Generative Large Language Models Empowering Psychometrics: Advantages, Challenges, and Applications

Authors: Tian Xuetao, Zhou Wenjie, Luo Fang, Qiao Zhihong, Feng Yi, Feng Yi

Date: 2025-10-22T00:00:00+00:00

Abstract

Generative Large Language Models (LLMs) are artificial intelligence models pre-trained on large-scale corpora that present unprecedented opportunities and challenges to the field of psychometrics. This paper synthesizes the evolution of interdisciplinary research between artificial intelligence and psychology, summarizes the significant advantages of LLMs in empowering psychometrics, identifies the critical challenges of LLMs in psychological applications, and proposes research directions for psychometric studies based on LLMs. Specifically, LLMs can generate coherent natural language text based on context, with the potential to transform traditional test interaction methods; LLMs demonstrate breakthrough capabilities in processing ultra-long texts and multimodal data, and their powerful content understanding abilities can comprehensively acquire and analyze psychological information of test takers; LLMs facilitate real-time analysis and personalized feedback, promoting a shift from outcome evaluation to process evaluation. Despite challenges in the practical application of LLMs, including stability, creativity, and scalability, they demonstrate broad application prospects and research value in areas such as situational judgment test generation, collaborative problem-solving ability assessment, intelligent diagnosis and treatment of mental health, and test item quality analysis.

Full Text

Empowering Psychometrics with Generative Large Language Models: Advantages, Challenges, and Applications

XUETAO TIAN[1], WENJIE ZHOU[2], FANG LUO[1], ZHIHONG QIAO[1], YI FENG[3]

[1]Faculty of Psychology, Beijing Normal University, Beijing, 100875, China
[2]Berkeley School of Education, University of California, Berkeley, 94720, USA
[3]Mental Health Center, Central University of Finance and Economics, Beijing, 100081, China

Abstract

Generative Large Language Models (LLMs) represent a class of artificial intelligence models pre-trained on massive corpora, bringing unprecedented opportunities and challenges to psychometrics. This paper synthesizes the developmental trajectory of interdisciplinary research between artificial intelligence and psychology to summarize the significant advantages of LLMs in empowering psychometrics, identify critical challenges in their psychological applications, and propose future research directions for LLM-based psychometric studies. Specifically, LLMs can generate coherent natural language text based on context, with the potential to transform traditional test interaction paradigms. Their breakthrough capabilities in processing ultra-long texts and multimodal data enable comprehensive capture and analysis of participants' psychological information. LLMs also facilitate real-time analysis and personalized feedback, promoting a shift from outcome-based to process-oriented evaluation. Despite practical challenges related to stability, creativity, and scalability, LLMs demonstrate substantial promise and research value in domains such as Situational Judgment Test generation, collaborative problem-solving assessment, intelligent mental health diagnostics, and test item quality analysis.

Keywords: Generative Large Language Models, Psychometrics, Artificial Intelligence, Automated Assessment, Interactive Testing

Psychometrics, as a foundational domain of psychological research, is dedicated to developing and refining tools and methods for assessing individual psychological traits. With societal progress and technological development, psychometrics faces new challenges: how to comprehensively improve the speed, precision, and ecological validity of psychological trait assessment (Luo et al., 2021). To meet the new demands of modern psychometrics, artificial intelligence technology has emerged as a transformative force. For instance, researchers have introduced automated item generation techniques based on machine learning and natural language processing into test development workflows to improve efficiency (Gotz et al., 2023; Hommel et al., 2022; Laverghetta Jr & Licato, 2023). Machine learning methods have also been integrated into measurement models such as Item Response Theory (IRT) and Cognitive Diagnostic Models (CDM) to enhance the precision of individual trait identification (Bergner et al., 2012; Martínez-Plumed et al., 2016; Pliakos et al., 2019; Wang et al., 2023). However, the most significant limitation of existing AI-integrated assessment methods lies in their dependence on large amounts of high-quality annotated data. Whether for guiding test development or training scoring models, data acquisition requires substantial time and human resources (Ersozlu et al., 2024). Moreover, constrained by data volume, these models typically exhibit poor generalization,

performing well on one test but poorly on another, and failing to adapt to new task requirements (Janiesch et al., 2021).

The rapid development and iteration of Generative Large Language Models (Generative LLMs, commonly abbreviated as LLMs) have created new opportunities and transformations for psychometrics. LLMs are AI technologies pretrained on large-scale corpora that can capture complex contextual semantic information and support fine-tuning optimization for specific scenarios (Zhao et al., 2023). Relevant core concepts and definitions are shown in Table 1 . Conducting psychometric research based on LLMs facilitates comprehensive intelligence from data acquisition and analysis to result feedback, making psychological assessment more efficient and precise. LLMs can also generate highly flexible and diverse natural language, providing rich possibilities for psychometrics development. However, whether LLMs can replace test developers, administrators, scorers, feedback providers, or even test-takers has sparked extensive exploration and debate in the psychometrics field (Buongiorno et al., 2024; Chiu et al., 2024; Goretzko & Bühner, 2022; Ke et al., 2025; Lee & Yeo, 2022; Pellert et al., 2024). This paper examines the developmental history of psychometrics and its intersection with AI technology to clarify the significant advantages of generative LLMs in empowering psychometric research and applications regarding interaction methods, content understanding, and scoring techniques, while also noting that LLMs still face technical challenges in stability, creativity, and scalability. Furthermore, through simulating trait-relevant behavioral contexts, constructing standardized agent interactions, implementing dynamic emotion recognition dialogues, and simulating expert/examinee roles, the integration of LLMs and psychometrics demonstrates promising applications in Situational Judgment Test generation, collaborative problem-solving assessment, intelligent mental health diagnostics, and test item quality analysis, serving as key breakthrough directions for LLM-empowered psychometrics.

2.1 Generative LLMs Transform Psychological Test Interaction Methods

As technology has evolved, the interaction forms of psychological testing have undergone tremendous transformation, from paper-and-pencil tests to computerized assessments, and now to individualized conversational interactions with computers through intelligent technology. This section briefly introduces the developmental history and characteristics of test formats, exploring how LLM-based interaction methods advance psychometric paradigms.

Paper-and-Pencil Psychological Tests

Paper-and-pencil tests have a long history as the earliest form of psychological assessment, with interaction mediated through paper and writing instruments. While psychology widely recognizes that Binet and Simon constructed the first modern intelligence test in 1905 (Boake, 2002; Matarazzo, 1992), records show that paper-based ability tests were widely used in China over 1,000 years ago (Zhang & Luo, 2020). Paper-and-pencil tests typically follow a fixed format,

with examinees answering predetermined questions in sequence. Item formats include objective types such as multiple-choice, true/false, matching, and fill-in-the-blank, or open-ended subjective questions requiring written responses (Berry, 2008).

Computerized Psychological Tests

With the popularization of computers and networks, computerized testing has enabled more flexible interaction methods. Beyond conventional items from paper-and-pencil tests, computerized assessments allow for more complex and realistic scenario-based items. For example, PISA (Programme for International Student Assessment) added problem-solving tests in 2012 using application problems aligned with real-life scenarios, providing new possibilities for assessing higher-order abilities, and in 2015 employed fixed human-computer interaction items to measure collaborative problem-solving skills (OECD, 2013; 2017). Gamified assessment and game-based assessment represent important developments in computerized testing. Combining gaming' s engaging qualities can effectively reduce test anxiety (DeRosier & Thomas, 2019; Mavridis & Tsiatsos, 2017), with process data and scoring data from games used to assess personality traits and cognitive abilities, showing substantial potential (Haizel et al., 2021; Kim et al., 2016; Sun et al., 2018). Computerized Adaptive Testing (CAT) combines computer technology with IRT to select items matching examinees' ability levels based on their responses in real-time, thereby reducing unnecessary items, shortening test duration, improving efficiency, and enabling cross-item and cross-time equating. CAT is currently widely used in major international assessments such as GRE, GMAT, and Duolingo.

LLM-Empowered Psychological Tests

LLM-based human-computer interaction will profoundly impact psychometric applications. In previous computerized tests, computers primarily received human information through directive interaction modes—for instance, in game-based assessments, test-takers conveyed decisions via mouse and keyboard. While effective, this limited operational space meant the captured data could not comprehensively reflect test-takers' psychological traits and decision-making processes. With LLMs, test interaction becomes more natural and flexible. A significant advantage of LLMs is their ability to engage in natural language interaction, dramatically expanding computers' capacity to capture psychological information. Through conversation, LLMs can obtain real-time linguistic features including tone, semantics, and sentence structure (Wu et al., 2024; Xu et al., 2023). This not only helps identify test-takers' emotional states, cognitive load, and motivation levels but also captures more nuanced psychological changes through multi-turn dialogues. For example, in stress-related contexts, LLMs can assess test-takers' linguistic performance across different stress levels through continuous conversation, enabling more accurate measurement of coping strategies and psychological resilience.

Furthermore, through robotic agents, LLMs can simulate diverse scenarios and dynamically adapt to different testing needs, significantly enhancing test im-

plementation flexibility. LLM-based dialogues and robotic agents can assume various social roles differing in age, profession, and education level, engaging test-takers in in-depth conversations to actively elicit psychological responses, thereby obtaining richer psychological information and achieving more precise and personalized assessment (Hu, 2024; Kharitonova et al., 2024; Yang et al., 2024). This drives psychometrics'transformation from traditional testing models toward intelligent and ecological approaches.

2.2 Generative LLMs Break Through Test Content Understanding Capabilities

As psychometrics continues to develop, researchers face not only structured data processing but also the complexity of unstructured data such as interview records, counseling dialogues, and audio-visual materials. These unstructured data typically contain rich semantic information and emotional expression, crucial for deeply analyzing individuals' psychological states, behavioral patterns, and emotional changes. How to efficiently understand and process such unstructured data has become a key technical challenge. This section briefly reviews the development of AI content understanding capabilities and their applications in psychometrics, elaborating on LLMs'breakthroughs in ultra-long text processing and multimodal data understanding.

Breakthroughs in LLMs' Long-Text Understanding Capabilities
Text data is one of the most common and easily collected data types in psychological research. In psychometrics, text analysis has become an important research method, with information mined from textual data used to analyze linguistic characteristics of different mental health states (Eichstaedt et al., 2018; Jose et al., 2022), support personality prediction (Majumder et al., 2017; Rahman et al., 2019; Ren et al., 2021), and conduct social-emotional analysis (Antypas et al., 2023; Vosoughi et al., 2018). Understanding individual differences through text analysis is built upon text representation, whose technological development has undergone four stages: word-based, topic-based, word vector, and pre-trained language models (see Figure 1

).

Traditional text representation techniques such as Bag of Words (BoW) and TF-IDF were among the earliest applied to text data processing. These models represent text as unordered word collections for basic statistical analysis and feature extraction. While widely applied, they cannot capture contextual relationships between words, making it difficult to understand polysemy, synonymy, and context-dependent text, with obvious limitations in processing long texts and complex semantics (Asudani et al., 2023; Ludwig et al., 2021; Zhang et al., 2010). To overcome these limitations, distributed models emerged, representing words as high-dimensional vectors to capture semantic relationships between words, achieving significant progress in text processing. Latent Semantic Analysis (LSA) represents distributed models through singular value

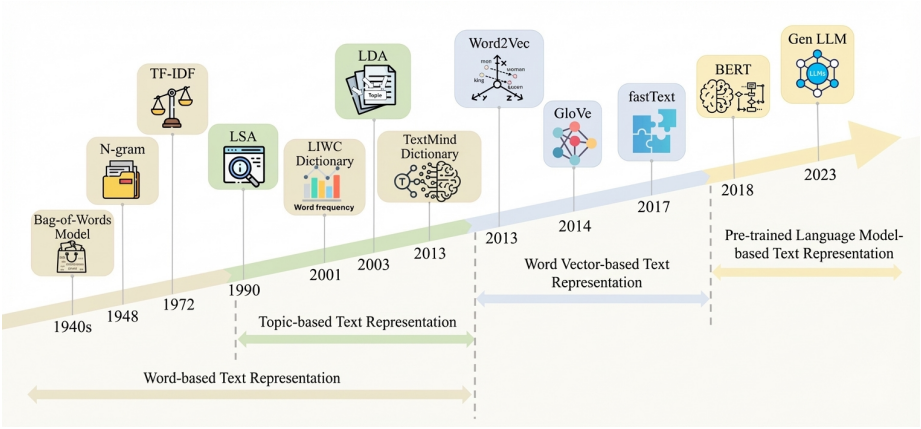

Figure 1: Figure 1

decomposition of term-frequency matrices, mapping high-dimensional text data to low-dimensional latent semantic spaces to reveal underlying semantic relationships between texts and words (Deerwester et al., 1990). Subsequently, to address issues of ignoring word order, syntactic relationships, logic, or morphology (Landauer et al., 1998), models like Word2Vec and GloVe were proposed, substantially enhancing text semantic understanding and enabling more precise analysis of latent semantic differences in open-ended responses (Dipietro et al., 2008; Jatnika et al., 2019; Ma & Zhang, 2015; Pennington et al., 2014; Uymaz & Metin, 2022), thereby improving psychometric accuracy (Foltz et al., 2023; Sonabend W et al., 2020). However, as data scenarios became more complex, these methods still struggled to meet demands for ultra-long text processing, particularly when deep contextual understanding and integration of multiple data sources were required.

The introduction of Pre-trained Language Models (PLMs) marked a new phase in natural language processing. Building upon distributed models, PLMs can dynamically adjust word representations based on context, enabling capture of complex semantic relationships. ELMo (Embeddings from Language Models) represents early PLMs, using bidirectional Long Short-Term Memory (LSTM) networks to capture dynamic word changes in context, making each word' s representation dependent on both itself and its context (Peters et al., 2018). This dynamic embedding approach significantly improved model performance across various NLP tasks. Considering LSTM' s sequential processing inefficiency and difficulty in parallelization for long texts, BERT (Bidirectional Encoder Representations from Transformers) achieved deep contextual understanding through Transformer architecture and bidirectional encoding mechanisms (Vaswani et al., 2017). Unlike traditional models, BERT employed Masked Language Model and Next Sentence Prediction tasks during pre-training, demonstrating excellence across multiple NLP tasks (Vaswani et al., 2017). The GPT (Generative

Pre-trained Transformer) model focuses on generating coherent subsequent text from given input. Unlike BERT, GPT adopts a unidirectional Transformer architecture, generating text through sequence prediction tasks (Radford et al., 2018). While its unidirectional structure presents limitations in handling bidirectional contextual dependencies, its text generation capabilities make it perform exceptionally well in many generative tasks. Context-aware models represented by ELMo, BERT, and GPT-1 learn fundamental language abilities through large-scale unlabeled text data during pre-training, then adapt to specific tasks using labeled data during fine-tuning to better solve downstream NLP tasks like cloze or summarization. Since then, numerous PLMs have been developed, but prior to LLMs, they were constrained by input text length limitations, making them difficult to apply in ultra-long contextual scenarios.

The emergence of LLMs such as GPT-3/4 and LLaMA provides new solutions for text analysis in psychological research, especially for long-text data analysis (Acheampong et al., 2021). Unlike early PLMs, these LLMs can solve new NLP tasks without relying on fine-tuning with downstream task data, instead using "In-Context Learning (ICL)"to leverage few-shot data during interaction (Brown et al., 2020). LLMs'leap in long-text understanding capability primarily benefits from synergistic advances in algorithms, encoding, and hardware. At the algorithmic level, researchers have developed more efficient attention mechanisms like FlashAttention (Dao et al., 2022) to reduce computational complexity, enabling processing of longer texts. At the encoding level, technologies such as Relative Position Embedding (RoPE, Lazos et al., 2005) and Attention with Linear Biases (ALiBi, Press et al., 2022) enhance models' extrapolation capabilities for unseen long texts. At the hardware level, developments in large-memory GPUs, tensor parallel computing, and memory optimization strategies provide the physical foundation for running large models with longer context windows. These technological advances collectively drive exponential growth in LLM context length—for example, OpenAI' s GPT models increased context length from 2,048 tokens in GPT-3 and 4,096 tokens in GPT-3.5 to 8k/32k tokens in GPT-4, and up to 128k tokens in GPT-4o, enabling models to process book-length texts of hundreds of thousands of words in a single pass.

LLMs demonstrate powerful capabilities in ultra-long text understanding. First, through extensive pre-training corpora and complex network structures, LLMs can maintain semantic consistency across long sequences and capture subtle semantic changes in multi-turn dialogues or lengthy narratives, enabling better contextual coherence. Second, through ultra-large-scale text pre-training, LLMs accumulate rich world knowledge. Compared to traditional machine learning models, these LLMs not only understand diverse contexts but also demonstrate deep world knowledge understanding during task execution. According to the CompassRank report, GPT-4o achieved a total score of 467 in science and 531 in humanities on the 2024 Chinese National College Entrance Examination (New Curriculum Standard I), exceeding Guangdong Province's undergraduate admission threshold, with 111.5 in Chinese and 141.5 in English, indicating that the latest generation of LLMs possesses substantial knowledge reserves. LLMs also

exhibit general task-solving capabilities, showing abilities far beyond traditional models even without optimization for specific downstream tasks. This generality enables LLMs to gradually replace many task-specific solutions in NLP, such as machine translation, text classification, and text retrieval. LLMs have also created new research paradigms for psychology, demonstrating advantages in identifying mental health conditions, personality, depression, moral values, and political orientation (Acheampong et al., 2021; Bai et al., 2025; Brady et al., 2021; Roy et al., 2022; Xu et al., 2024; Yang et al., 2023). Additionally, LLMs show strong reasoning capabilities when facing complex tasks, solving reasoning problems involving complex knowledge relationships and mathematical reasoning tasks (Ahn et al., 2024; Huang & Chang, 2023). LLMs' long-text understanding capabilities enable them to receive test-takers' genuine intentions expressed in natural language and execute complex task instructions when applied to psychometric scenarios (Rathje et al., 2024), providing technical support for ecologically valid measurement tasks.

Breakthroughs in LLMs' Multimodal Data Understanding Capabilities

As technology evolves, psychometrics has gradually expanded to collect and analyze various complex and diverse multimodal data, including textual data from interview records and open-ended questions, process data such as movement paths, click behaviors, and response times from virtual reality or gamified tests, physiological data like heart rate, skin conductance, EEG, and respiration from biosensors and wearable devices, vocal features including intonation, speech rate, volume, pauses, and emotional expression, and visual information such as facial expressions, body posture, and eye-tracking from video recordings. In psychometrics, multimodal data can provide richer individual information, particularly valuable when assessing complex psychological traits (Lin et al., 2020; Obrenovic & Starcevic, 2004; Palumbo et al., 2020; K. Sharma & Giannakos, 2020). However, effectively fusing and analyzing these different modalities has remained a challenge. With widespread application of speech analysis and computer vision technologies, psychometrics' ability to understand multimodal data has substantially improved. For example, speech analysis technology can identify and analyze pronunciation, tone, and voice quality, finding broad application in language tests like TOEFL iBT and Mandarin proficiency tests (Huawei & Aryadoust, 2023; Palanivinayagam et al., 2023). Computer vision technology also plays important roles in personality analysis based on audio-video data, automated interview scoring, and movement assessment (Debnath et al., 2022; Haizel et al., 2021; Silva et al., 2021; Zhang et al., 2024). These technologies not only enhance test intelligence but also support measurement diversification and result accuracy. However, current multimodal data understanding remains largely at the feature extraction level, struggling to achieve human-computer multimodal data interaction.

Generative LLMs provide new solutions for multimodal data understanding and fusion. Previous multimodal analysis research generally employed feature fusion methods, such as simultaneously extracting non-verbal information like facial

expressions, body movements, and vocal intonation to complement textual data and form psychological profiles (Wang et al., 2024). However, this approach struggles to capture cross-modal information relationships. LLMs achieve deep fusion and collaborative reasoning of cross-modal information by combining multimodal-specific encoders with shared Transformer layers, mapping text, images, audio, and other inputs into a unified semantic vector space (embedding space) (Wang et al., 2025). For example, models like GPT-4 and PaLM2 can process textual and visual information simultaneously in conversational form, demonstrating excellent performance across multimodal tasks and extracting meaningful psychological trait features (Wu et al., 2024; Xu et al., 2023). As psychometrics continues to develop, researchers will face more challenges in processing unstructured data containing rich semantic information and emotional expression crucial for understanding individuals' psychological states and behavioral patterns. LLMs' breakthroughs in multimodal data understanding provide new pathways for addressing these challenges.

2.3 Generative LLMs Broaden Psychometric Scoring Methods

Psychometric scoring and evaluation methods have undergone significant transformation. First, the evaluation 主体 has gradually shifted from human raters to computer-automated scoring, improving efficiency and consistency while reducing misjudgment and unfairness. More importantly, evaluation content has moved from traditional outcome-orientation toward process-orientation. Process-oriented evaluation enables real-time assessment and item adjustment at the test level while collecting and analyzing process data for more comprehensive measurement of individual psychological states. Temporally, process-oriented evaluation combines information from multiple observations to establish dynamic profiles of individual psychological development, better facilitating personal growth. This section focuses on typical application scenarios of psychometrics in educational assessment, examining LLMs' key roles in current psychometric evaluation through the lens of educational test scoring technology development.

From Human Subjective Scoring to Automated Scoring
Educational test scoring has transformed dramatically from human-dependent to automated approaches. In early paper-and-pencil tests, scoring relied primarily on manual review by teachers or evaluators following preset standards for both objective and subjective items. However, this approach suffered from significant limitations, including rater bias, consistency issues, and heavy workload (O. L. Liu et al., 2014). To address these problems, computer technology was gradually introduced into psychometric scoring.

Initial automated scoring focused on objective items, using Optical Character Recognition (OCR) to rapidly identify and score large volumes of multiple-choice and true/false questions (Alomran & Chai, 2018; McKenna, 2019; Memon et al., 2020). While greatly improving efficiency, this remained limited to objec-

tive assessment. With advances in Natural Language Processing (NLP) and machine learning, automated scoring expanded to subjective items. Early text similarity methods using BoW or TF-IDF features calculated similarity between responses and standard answers to achieve basic automated scoring functionality (Ramnarain-Seetohul et al., 2022; Wang, 2022), but performed poorly with complex semantics (Dai et al., 2024). With supervised learning algorithms like Convolutional Neural Networks, Recurrent Neural Networks, and sequence regression, automated scoring technology further developed. Trained on large annotated datasets, these models could score more complex subjective items, significantly improving accuracy and consistency (Chen & Zhou, 2019; Liang et al., 2018). However, these models depend on large high-quality training data and remain limited in handling diverse item types and data (Devine et al., 2023).

LLMs demonstrate enormous potential in subjective item scoring due to their ability to process complex natural language texts and their exceptional semantic understanding capabilities. Models like GPT-3/4 and BERT, through pretraining on massive corpora, can precisely discern subtle semantic differences in answers to achieve more accurate test scoring (Fernandez et al., 2022; Lee et al., 2024; Ludwig et al., 2021; Shin et al., 2024; Takano & Ichikawa, 2022; Yancey et al., 2023; Zhu et al., 2022). Additionally, in LLM-empowered educational testing, computers can not only passively receive information but also autonomously adjust assessment processes based on test-taker feedback. Through dialogue, LLMs can analyze responses in real-time and dynamically adjust question difficulty, content, or context (Chiu et al., 2024; van Velthoven et al., 2018; Zhang et al., 2024). This adaptive assessment process promises more effective capture of test-takers' latent traits. For example, in cognitive ability testing, LLMs can adjust item complexity based on immediate performance, ensuring results authentically reflect ability levels (Hu, 2024). This adaptive testing not only improves efficiency but also reduces fatigue and anxiety, yielding more reliable data. LLM-based tests can also promote student progress through automated feedback, generating detailed explanations that help examinees understand weaknesses and provide improvement suggestions, thereby enhancing learning outcomes (Alomran & Chai, 2018; Gabbay & Cohen, 2024; Shaik et al., 2022). Moreover, LLM-based feedback systems show potential in improving positive emotions, empathy, and creativity (Dong et al., 2024; Meyer et al., 2024; Sharma et al., 2023; Stamper et al., 2024).

From Outcome Evaluation to Process Evaluation

Traditional educational tests typically focus on outcome evaluation, assessing examinees' final answers. Whether multiple-choice, true/false, or open-ended subjective questions, raters evaluate responses against preset correct answers or scoring standards (Berry, 2008; Landauer et al., 1998). This outcome-oriented approach often overlooks thinking processes, emotional reactions, and behavioral patterns exhibited during test-taking (Schulte-Mecklenbeck et al., 2011). While outcome evaluation can score final performance, it cannot comprehensively reflect psychological states and cognitive processes, particularly limiting when assessing complex psychological traits like motivation, emotional response,

and problem-solving strategies. With computer technology 普及 and psychometric model advances, process-oriented evaluation has become an important assessment component. Process evaluation focuses not only on final answers but also incorporates various data generated during testing, such as response times, mouse clicks, and eye-tracking trajectories. These process data provide detailed information about test-taking behavior, including thinking paths, strategy selection, and cognitive load (Cho et al., 2020; Liao & Jiao, 2023; Ma & Guo, 2019; Man et al., 2022). By analyzing process data, researchers can more comprehensively evaluate psychological states and behavioral patterns. For example, in problem-solving tests, process evaluation analyzes not only whether correct answers were given but also examines steps, response times, and operation sequences during problem-solving to infer strategies and cognitive complexity (Chen, 2020; Chen et al., 2019; He et al., 2021).

With Multi-modal LLMs development, LLMs can simultaneously process and understand various data forms including text, images, and audio. This cross-modal processing capability enables more comprehensive and accurate assessment of individual psychological states (Dong et al., 2024). For example, LLMs can comprehensively evaluate emotional states and mental health by analyzing examinees' written responses, vocal intonation, facial expressions, and other process-oriented multimodal information (Li et al., 2023; Wu et al., 2024; Zhang et al., 2024).

In summary, LLMs have greatly advanced psychometric scoring methods, with LLM and MLLM technologies driving tests from traditional "measurement-evaluation" toward "measurement-evaluation-feedback-development."

3 Challenges of Generative LLMs in Empowering Psychometrics

While generative LLMs offer tremendous potential in transforming interaction methods, content understanding, and evaluation approaches in psychometrics, creating numerous innovations, we must recognize that current LLM technology still has limitations. This section discusses challenges in stability, creativity, scalability, ethics, data security, and cost, summarizing current research approaches to address these issues for better discussion of LLMs' potential in psychometrics.

3.1 LLM Stability Issues

Generative LLMs demonstrate enormous application potential, yet stability issues pose major obstacles to widespread adoption. These include output instability, occasional context loss, factual or commonsense errors, and cultural-linguistic biases. First, LLM outputs are often sensitive to minor input variations, leading to inconsistent results. Huang et al. (2024) compared consistency retention rates of 26 LLMs before and after receiving perturbed inputs, finding that GPT-4' s accuracy decreased by 10-21% after perturbation, while PaLM2'

s accuracy dropped by 8-25%, demonstrating that subtle input differences can cause performance degradation. Similarly, Zhao et al. (2024a) found that when input content was paraphrased while preserving semantics, LLM output consistency rates were only 57-80%. This inconsistency proves particularly unreliable for psychometric applications requiring high standardization. For example, in automated scoring tasks, LLMs might assign different scores to identical or similar answers, challenging assessment fairness and reliability. Research shows that simply changing answer order in evaluation templates can alter or distort GPT-4' s ranking of answers, with GPT-4 showing scoring biases toward answers in specific positions, longer answers, or answers similar to its own generated content (P. Wang et al., 2023; Zheng et al., 2023). To address output consistency issues, Zhao et al. (2024b) proposed Supervised Fine-Tuning (SAT) and Consistency Alignment Training (CAT). In the SAT stage, models generate multiple paraphrases of original instructions, pairing these rewritten instructions with original training data to create augmented samples, improving models' generalization to diverse instruction expressions while precisely understanding core semantics. In the CAT stage, multiple generated responses are scored to optimize models for producing consistent, desirable responses, thereby improving both diversity and consistency. Wang et al. (2023) proposed Balanced Position Calibration (BPC) and Multiple Evidence Calibration (MEC), with BPC calculating average scores through multiple answer order changes and MEC integrating multiple evaluation results for the same answer set to obtain more stable and accurate final scores.

Context loss represents another common stability issue in long conversations or complex tasks. As conversations progress, models may forget previous contextual information, resulting in incoherent and illogical outputs due to insufficient long-term memory capabilities, posing challenges for maintaining consistent memory across extended periods or sessions. This problem is particularly prominent in human-computer psychological counseling, where forgetting previous context affects individual psychological assessment accuracy and presents significant challenges for repeated assessments and dynamically changing psychometric tasks. When relevant information appears in the middle of input context, many LLMs show substantially reduced accuracy because attention mechanisms during pre-training and fine-tuning tend to focus more on beginning and ending content while neglecting middle sections. Specifically, as textual distance increases, attention decay makes it difficult for models to attend to distant middle portions, causing information loss or misinterpretation (Liu et al., 2023). To address this, researchers have proposed methods like Position-Agnostic Multi-step QA (PAM QA) through multi-step question decomposition training, enabling models to search for and extract relevant content across different positions to balance attention distribution. This approach significantly improves model performance on long-text inputs, particularly in multi-document question-answering tasks (He et al., 2024).

Additionally, factual errors represent a critical issue. Although LLMs can generate seemingly authentic answers, these may be inaccurate or completely wrong—

a phenomenon known as LLM Hallucination (Huang et al., 2025). Since models cannot verify generated content' s authenticity, this represents a major deficiency for psychometric tasks requiring factual accuracy. LLM outputs depend on large-scale natural language data, typically producing "average" viewpoints common on the internet or in popular books rather than necessarily conforming to psychological theories or evidence-based standards. Moreover, LLMs may inherit and amplify existing biases from training data, with outputs reflecting cultural and social biases that make them potentially unreliable for generating mental health-supportive language (Demszky et al., 2023). Current research attempts to improve LLM output validity and reduce cultural-linguistic biases and hallucinations through more diverse training data, fact-checking and self-reflection mechanisms, external database connections, and optimized prompting (Demszky et al., 2023, 2024; Huang et al., 2024; Ji et al., 2023; Rawte et al., 2023; Stade et al., 2024; Wei et al., 2024).

Measurement invariance must be considered when applying LLMs. Given rapid LLM technology iteration, measurement tools built on closed-source models (like GPT or Gemini series) may experience performance and output drift due to underlying models' "silent updates" (Chen et al., 2023; Ma et al., 2024). This uncertainty severely undermines measurement stability and may alter longitudinal research results. For closed-source models, regular API monitoring mechanisms should be established using typical response datasets to verify output consistency. For open-source models like DeepSeek and Qwen, using fixed-version APIs and conducting fixed-version local deployments are important pathways to ensuring measurement invariance.

3.2 LLM Creativity Issues

LLM applications in psychometrics also face creativity challenges. Although LLMs show great potential in generating linguistic content, with the latest generation performing comparably to or even exceeding human test-takers on creativity tests in some dimensions (Bellemare-Pepin et al., 2024; Guzik et al., 2023), they still exhibit several problems. First, LLMs lack genuine originality, generating content through recombination of existing patterns from training data rather than novel conceptualization. While LLMs' creativity test performance averages higher than human means, their distribution variance is much smaller, making it difficult for them to achieve extremely high scores (Hubert et al., 2024; Bellemare-Pepin et al., 2024), suggesting LLMs currently struggle with breakthrough innovation. In psychometrics, this may result in generated assessment tools or items lacking novelty, unable to break existing measurement paradigms. Due to dependence on learned structures and common patterns, LLMs may exhibit templated generation that remains at the textual surface concept level when creating creative content (like situational simulation items or open-ended questions in psychometrics). This limitation becomes particularly evident in psychometric tasks requiring diversity and complexity to capture individual differences. Second, LLMs cannot create new constructs. Psychological

research often requires proposing new psychological constructs to assess not-yet-fully-understood traits. However, LLMs can only generate content based on existing linguistic patterns without transcending training data. Whether LLMs possess genuine creativity remains an unresolved research question (Zhao et al., 2024b), limiting their application in developing new psychometric tools. To address these issues, researchers propose using multiple LLM agents playing different roles in discussions to simulate human collective creativity formation, enhancing LLM innovation capabilities (Lu et al., 2024), or employing associative thinking strategies to improve LLMs' ability to integrate different concepts (Mehrotra et al., 2024), which may become new approaches for LLM-empowered psychometrics.

3.3 LLM Scalability Issues

LLM scalability also represents an important limiting factor for practical use. Despite demonstrating strong capabilities in text generation and natural language processing, LLMs face challenges in scaling psychometric applications. First, LLMs have limitations in adapting to new constructs. Psychometric constructs continuously evolve with new psychological theories and empirical research, while LLMs typically rely on existing datasets for training, making them inadequate for handling data outside sample distributions. LLMs lack flexibility in effectively integrating new psychometric constructs, limiting their application in dynamically developing psychology fields. Although some studies have validated LLMs' reliability and validity in scale development based on new constructs (Hoffmann et al., 2024; Ouédraogo et al., 2024), researchers require strong critical thinking and professional expertise to provide appropriate instructions at each scale development step. Additionally, current research lacks comprehensive evaluation of LLM applications like automatic item generation and LLM-simulated test-takers across different domains, with their psychometric properties and content reliability not yet extensively validated (Circi et al., 2023; Zhu et al., 2022).

LLM scalability issues also manifest in multimodal data integration. Although LLMs can analyze and output multimodal data, they are often limited to shallow understanding, unable to fully capture complex relationships between different data types. In other words, current multimodal LLM applications in psychometrics are insufficient to fully exploit multimodal data potential, with such complex relationships still requiring manual pre-alignment. Moreover, LLMs face challenges in cross-cultural and linguistic adaptability. Psychometric tools often require cross-cultural application, demanding that LLMs handle languages and psychological constructs across different cultural backgrounds. However, since LLM training data may be biased toward specific cultural or linguistic contexts, their scalability and adaptability in cross-cultural situations are clearly inadequate. This limitation may cause psychometric tools to exhibit bias in different cultural environments, reducing their validity and universality (Du et al., 2023; Huang et al., 2024; Wang et al., 2023). Furthermore, psychometrics encompasses

multiple domains such as cognitive assessment, emotional measurement, and social evaluation. LLMs may not maintain consistent performance and accuracy when scaling to these different domains, with limited generalization capabilities (Ahn et al., 2024), requiring researchers to fine-tune models according to specific domain needs. Additionally, current LLMs have suboptimal mathematical computation and reasoning abilities, potentially exhibiting logical confusion and inconsistency when assisting in generating items measuring higher-order cognitive functions. Research suggests that causal knowledge graph technology and chain-of-thought techniques can help LLMs implement more logical task execution processes and discover potential associations between concepts (Tong et al., 2024). OpenAI' s latest GPT-o1 model makes preliminary explorations in this area, using chain-of-thought technology to substantially improve logical reasoning and computational abilities, but its high computational requirements make it difficult to 普及. Overall, LLM scalability issues present multiple challenges for psychometric applications. Achieving truly widespread LLM application in psychometrics still requires greater progress in model flexibility, multimodal data understanding, multi-domain scalability, and cultural adaptability.

3.4 LLM Ethics, Data Security, and Cost Issues

Applying LLMs to psychometrics requires careful consideration of ethics, security, and cost issues. First, LLMs' "knowledge" derives from pre-training corpora that inevitably contain societal biases regarding culture, gender, race, etc. (Chen et al., 2024; Dai, Xu, et al., 2024; Taubenfeld et al., 2024). Uncritical application of these models in psychological assessment may systematically disadvantage specific groups. For instance, models trained primarily on Western cultural corpora may make inaccurate or incorrect judgments when assessing mental states of individuals from non-Western cultural backgrounds (Rao et al., 2025; Sakai et al., 2025). Therefore, bias detection and calibration before application, along with ensuring training data diversity and representativeness, are crucial.

Second, data privacy and security concerns are paramount. Psychometric assessment, especially mental health diagnosis, involves highly sensitive personal data. Data leakage risks during transmission and storage cannot be ignored when using cloud-based LLM APIs (Ke et al., 2025; Lawrence et al., 2024). Additionally, data used for fine-tuning models may be extracted or misused. Ensuring full-process data anonymization and exploring locally deployable open-source models are key pathways to protecting participant privacy.

Finally, high costs represent a practical bottleneck and challenge for widespread LLM adoption in psychometrics. Whether calling top-tier closed-source model APIs for large-scale data processing or fine-tuning and deploying open-source models, substantial computational resources and financial investment are required. Current open-source models based on more efficient architectures like Mixture-of-Experts, such as DeepSeek v3 and r1, GPT-oss 20b and 120b, and the Qwen 2.5 series, achieve performance comparable to large closed-source mod-

els at lower training and deployment costs, offering new possibilities for reducing LLM costs in large-scale psychometric applications.

4 Application Prospects of Generative LLMs in Empowering Psychometrics

Building upon discussions of LLMs' significant advantages and technical challenges, this section proposes key potential applications for LLM-empowered psychometrics, including Situational Judgment Test generation, collaborative problem-solving assessment, intelligent mental health diagnostics, and test item quality analysis, providing directions for future psychometric research and applications.

4.1 Situational Judgment Tests: LLM-Based Test Generation

Generative LLMs, through fusing massive textual data, develop deep understanding of psychological traits across populations and can generate questions aligned with specific personality theoretical frameworks. For example, when designing Big Five personality tests, LLMs can create new items based on existing dimensions (openness, conscientiousness, extraversion, agreeableness, neuroticism), ensuring items accurately reflect test-takers' personality traits while maintaining good reliability and validity (Gotz et al., 2023). Beyond self-report personality tests, Situational Judgment Test (SJT) development better leverages LLMs' content generation capabilities. SJTs are commonly used selection tools effective for predicting job performance, assessing decision-making ability and behavioral tendencies by presenting examinees with simulated work-related scenarios and requiring them to select optimal response strategies (Burrus et al., 2012). Current personnel selection SJTs face problems of item exposure preventing reuse, and SJT development particularly depends on domain experts and rigorous development processes. LLMs, through learning from massive textual data, acquire abilities to "simulate" or "role-play" different personality trait-related behavioral performances, a capability validated in previous research (Hewitt et al., 2024; Ke et al., 2025; Jiao et al., 2025). This ability to simulate individuals' different behavioral performances across situations due to internal trait differences can assist in generating reliable situational judgment items, with validated reliability and validity comparable to or exceeding human-developed items in cognitive and personality test generation (Laverghetta Jr. & Licato, 2023; Li et al., 2025).

In practice, SJT development generally involves multiple steps: scenario selection and item development, typical behavioral response option development, and response option scoring design (McDaniel et al., 2007). In scenario selection and item development, psychologists or domain experts typically determine scenarios based on experience and literature review, then write items—a process requiring substantial manual effort including reading materials, group discussions, and repeated revisions. Generative LLMs' intrinsic knowledge supports

generating diverse, realistic work-related simulation scenarios, with experts serving as evaluators of generated content to reduce workload and improve efficiency. In typical behavioral response option development and scoring, each scenario may have multiple options requiring careful design and writing by test developers to reflect different behavioral tendencies, with scoring heavily dependent on domain experts' understanding and judgment of scenarios. Generative LLMs, relying on their understanding of trait differences, can generate large numbers of response options meeting specified scoring conditions through diversified content generation. Combined with expert experience, selecting less distinguishable option combinations across different score levels can efficiently complete preliminary SJT development. Notably, LLM-generated tests still require data collection through administration to complete test quality validation.

4.2 Ability Testing: LLM-Based Collaborative Problem-Solving

Ability assessment focuses on evaluating individuals' cognitive abilities and professional skills, typically using multiple-choice questions, situational simulations, and computer-interactive tests (Luo et al., 2021). While these methods have some validity for assessing individual abilities, they have clear limitations in simulating real-world collaborative task scenarios. Although AI technology has increasingly enhanced interactive functions in ability tests, test-takers' interaction partners remain machine agents with fixed logic. While this benefits standardization, it fails to activate individuals' communication and collaboration abilities.

LLMs can effectively change this situation by playing researcher-defined roles and engaging in free dialogue with test-takers (Jandaghi et al., 2023). Specifically, researchers have designed prompt frameworks enabling LLMs to simulate human brainstorming processes and participate in creative problem-solving (Chang & Li, 2024). As LLM problem-solving capabilities continue improving, LLM-based agents can serve not merely as human tools but as standardized "partners" in problem-solving processes, activating and externalizing test-takers' collaboration-related abilities (such as communication, leadership, and collaborative problem-solving) for better measurement. Compared to human-human collaboration assessment paradigms where human partners' behavioral variability introduces additional measurement error (Biswas et al., 2010; Stadler et al., 2020), LLM agents can exhibit more stable and controllable behavioral patterns through pre-defined task strategies, reducing systematic error from assessment partners. Simultaneously, compared to fixed-logic human-computer collaboration paradigms, this approach better reflects real collaboration scenarios and has greater ecological validity.

Implementing assessment in LLM-based collaborative problem-solving requires consideration of two aspects: designing specific assessment task scenarios and achieving automated ability evaluation. To ensure validity, task scenarios obviously cannot be too simple to be completed easily by individuals or LLMs alone. Task design must consider openness and ambiguity, allowing multiple solution

paths to comprehensively examine decision-making ability and adaptability under uncertain conditions, fully mobilizing test-takers' collaboration willingness and requiring repeated evaluation, selection, or revision of LLM outputs. Test-takers can deeply immerse themselves in problem-solving through free interaction with LLMs, producing behavior responses consistent with their ability traits. Automated ability evaluation requires comprehensive consideration of associations between test-taker-LLM data interaction forms and multidimensional abilities activated during collaborative problem-solving. Data interaction may involve multiple modalities including text, images, and audio, while abilities should encompass both problem-solving and communication-collaboration dimensions. Through comprehensive analysis of these multimodal data, a comprehensive ability profile can be constructed, breaking through previous ability assessment limitations and providing more accurate, detailed results for practical applications like personnel selection.

4.3 Mental Health Testing: LLM-Based Intelligent Diagnostics

Mental health testing represents an important psychometric application direction for assessing depression, anxiety, stress, and other mental health conditions. Traditional methods like self-report questionnaires suffer from strong subjectivity, and in today' s era of rapid information flow, intentional faking with abnormal motivation creates significant difficulties for mental health diagnosis. While clinical interviews and behavioral observation have some validity, they are time-consuming and costly, making large-scale screening difficult (Jiang et al., 2022). Generative LLMs provide new opportunities for mental health testing by processing complex text inputs, generating targeted questions and feedback, and identifying implicit emotional states.

LLMs can advance mental health testing toward intelligent diagnostics through interview-based test formats and continuous expression-based psychological state assessment. Currently, mental health testing in China is primarily used for group screening, with at-risk individuals identified through screening further evaluated by professional counselors (Fang et al., 2018). To ensure at-risk individuals are not missed, this process often requires participation of numerous professional counselors, whose numbers typically cannot meet follow-up demands after large-scale screening. In this context, LLMs can serve as auxiliary tools by simulating counselor roles and engaging test-takers in natural dialogue to guide expression of inner feelings (Chen et al., 2023). For example, LLMs can be trained to employ specific counseling techniques, such as using cognitive restructuring or generating highly empathetic responses to help clients overcome communication barriers caused by shame or distrust (Xiao et al., 2024), thereby creating a safe expression environment. This automated interview approach not only reduces professional workload but also collects large amounts of valuable real-time data without interrupting dialogue flow, with regular online follow-ups potentially enabling timely detection of psychological changes. In psychological state assessment, test-takers no longer

provide uniform feedback and scoring as in traditional self-report scales but expose their authentic states and underlying causes through personalized Q&A processes. Analyzing individuals' continuous expressions in dialogue with LLMs yields more precise and detailed mental health diagnostic results while providing personalized suggestions and support for both test-takers and interveners. Existing LLM-based mental health assessment systems have been developed by training models to learn psychiatric interview frameworks and diagnostic standards, mimicking clinical doctors' psychiatric interview assessments and demonstrating high consistency with psychiatrists' evaluations (Bi et al., 2025).

4.4 Item Evaluation: LLM-Based Test Item Quality Analysis

Item evaluation is a crucial component of psychometrics aimed at ensuring test tool validity and reliability. Traditional item evaluation methods typically rely on expert review and statistical analysis, which are time-consuming and potentially subjective (Zhao et al., 2013). LLMs have been used for role-playing in many applications with good results (Lu et al., 2024; Shen et al., 2024). If LLMs can simulate domain experts or test-takers at different ability levels, they can enable LLM-based item quality analysis, improving evaluation efficiency and accuracy.

LLMs simulating domain experts can automatically assess item content quality, including language clarity, logical consistency, and difficulty level. By using multiple general or domain-trained LLMs, inter-rater consistency can be calculated like traditional expert review, selecting high-scoring items for application and reducing expert dependency. LLMs can also role-play test-takers from different cultural backgrounds, ages, and ability levels. Theoretically, by generating "virtual" answers matching corresponding representative populations and combining methods like Item Response Theory, item difficulty and discrimination parameters can be preliminarily assessed, providing a fast, low-cost pre-testing method for item quality analysis (Liu et al., 2025; Lu & Wang, 2024). However, for simulated test-takers, several issues exist: culturally, mainstream LLM training corpus bias makes it difficult to grasp deep values and expression habits of specific cultures, with generated answers potentially showing "averaged" or "Westernized" tendencies. Regarding age, models inadequately model language styles and cognitive characteristics of children and adolescents, struggling to authentically reproduce their response patterns. For ability levels, LLMs tend to generate logically coherent "standard answers" while potentially inadequately simulating knowledge gaps and error patterns typical of low-ability test-takers. Therefore, using LLM-simulated test-takers as item analysis tools remains exploratory, with results that cannot replace real human data but can serve as beneficial supplements to expert review and early item analysis tools. LLM-empowered item quality analysis helps ensure items are suitable for broad test-taker populations, improving test tool fairness and representativeness.

This paper explores opportunities and challenges that generative LLMs bring

to psychometrics, emphasizing their significant advantages in transforming test interaction methods, enhancing multimodal data processing capabilities, and broadening scoring approaches (see Figure 2

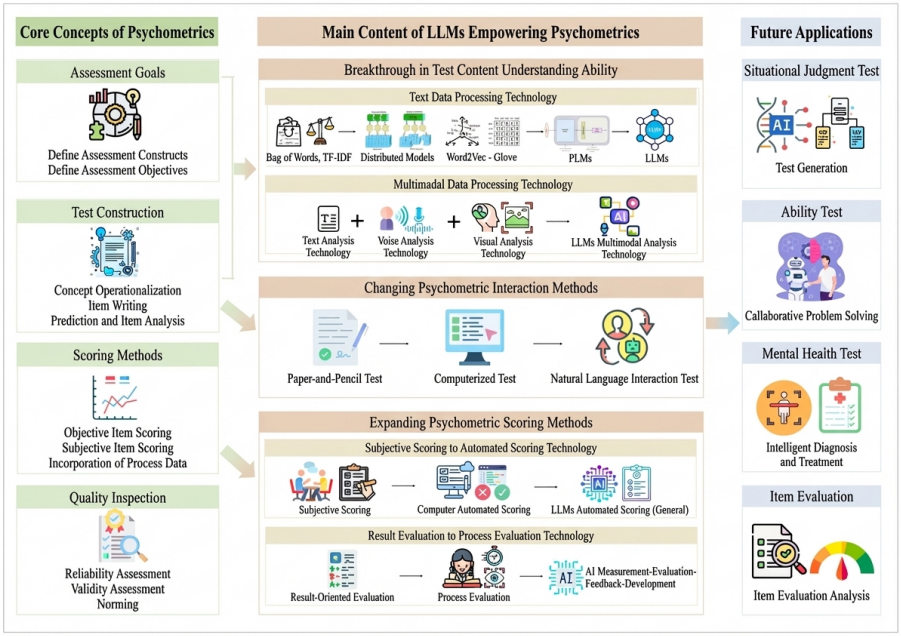

Figure 2: Figure 2

). Despite facing technical challenges such as stability, creativity, and generalizability, LLMs demonstrate broad prospects in Situational Judgment Test generation, collaborative problem-solving assessment, intelligent mental health diagnostics, and item quality analysis. With continuous technical optimization, such as improving model performance through instruction fine-tuning and consistency alignment training, LLMs will continue driving psychometrics toward greater intelligence, personalization, and efficiency.

Figure 2 Main content framework of LLMs empowering psychometrics

References

Fang, X., Yuan, X., Hu, W., Deng, L., & Lin, X. (2018). Development of the Chinese college student mental health screening Scale. *Psychology and Behavior Research*, 16(1), 111–118.

Jiang, L., Tian, X., Ren, P., & Luo, F. (2022). New mental health assessment with artificial intelligence assistance. *Advances in Psychological Science*, 30(01), 157–167.

Jiao, L., Li, C., Chen, Z., Xu, H., & Xu, Y. (2025). When AI "has" personality: The influence of good and evil personality roles on large language models' moral judgment. *Acta Psychologica Sinica*, 57(6), 929-949.

Luo, F., Tian, X., Tu, Z., & Jiang, L. (2021). New trends in educational evaluation: A review of intelligent assessment research. *Modern Distance Education Research*, 33(05), 1-12.

Ke, L., Li, Z., Liao, J., Tong, S., & Peng, K. (2025). Validity of large language models simulating regional psychological structures: An empirical test of personality and well-being. *Psychological Science*, 48(4), 907-916.

Zhao, S., Shi, Y., & Zhu, D. (2013). Application of item response theory in item quality evaluation for large-scale selective examinations. *Journal of Educational Studies*, 9(01), 58-65.

Sun, X., Li, J., & Fu, Z. (2018). Using game log-files to predict students' reasoning ability and mathematics achievement: Application of machine learning. *Acta Psychologica Sinica*, 50(7), 761-770.

Zhang, H., & Luo, F. (2020). Development of psychological and educational measurement in China. *Journal of Educational Measurement and Evaluation*, 1(1), Article 8.

Note: The reference list continues with the English-language citations provided in the original text, which are preserved exactly as given.