

GENERATIVE AI

in Test Development & Psychometrics

*An Analytical Review of Its Potential
and Limitations*

MEASURE
LEARNING

Aurelie Lecocq, Ph.D., MBA

*Director, Business Strategy and Growth,
Measure Learning*

About the Author



Dr. Aurelie Lecocq serves as the Director of Business Strategy and Growth at Meazure Learning, where she leads strategic initiatives to ensure the psychometric quality and defensibility of high-stakes licensure and certification programs. With extensive experience in psychometric consulting and bilingual proficiency in English and French, Dr. Lecocq works closely with regulatory bodies, professional associations, and subject matter experts to design and maintain valid, reliable, and fair assessment processes across healthcare and other professional sectors. Her expertise

spans competency framework development, test blueprinting, item and test analysis, standard setting, and statistical audits.

In her leadership role, Dr. Lecocq oversees a team of senior psychometricians, driving the continuous improvement of psychometric services and fostering innovation in assessment practices. She partners with clients to implement cutting-edge psychometric solutions that align with regulatory requirements and industry standards. Dr. Lecocq's goal is to deliver tailored, data-driven solutions that address each client's unique challenges while advancing the quality and impact of their assessment programs.

Dr. Lecocq holds a Ph.D. in Education from Université de Bourgogne (France) and has completed two postdoctoral fellowships in Psychoeducation in Canada. She also earned two master's degrees—a Master of Monitoring and Evaluation of Educational Systems from IUT Denis Diderot (France) and an MBA from Carleton University (Canada). Her international experience includes leadership roles in France, Vietnam, and Australia, which have shaped her approach to driving business growth and excellence in diverse cultural contexts.

Prior to joining Meazure Learning, Dr. Lecocq worked as an independent consultant, offering services in program evaluation, statistical analysis, survey design, and public policy analysis to a wide range of public and private sector clients in France and Canada. Her contributions have consistently supported clients in achieving their strategic objectives through evidence-based assessment and evaluation practices.



LET'S CONNECT ON LINKEDIN

A woman with curly hair is smiling and looking down at a desk. On the desk, there is a dark mug, a pen, and some papers. The background is slightly blurred, showing some indoor plants.

Table of Contents

- 2** Introduction
- 3** Exploring Generative AI Models
- 4** Performance of GenAI as a Test-Taker in High-Stakes Credentialing
- 7** Use of Generative AI Models in Assessment Development and Psychometrics
- 18** Legal and Ethical Considerations of Using Generative AI in Psychometrics
- 19** Principles for Responsible Use of GenAI in Test Development
- 20** Moving Forward With Generative AI
- 21** Citations

Introduction

The potential use of generative AI (GenAI) models in psychometrics is extensive and could be linked to activities such as job task analysis (JTA), competency development, item bank management, item writing, exam form assembly, standard setting, statistical analysis, reporting, and exam form translation. GenAI has the potential to make these processes more efficient and scalable, but it's crucial to consider the potential challenges and ethical implications.

There is no manual yet for the proper use of GenAI in psychometrics, and accreditation bodies have yet to determine the scope of its usage for certification. With so many possibilities, GenAI can feel intimidating and overwhelming for various stakeholders, including subject matter experts (SMEs), psychometricians, and industry leaders.

This report aims to demystify the potential use of GenAI models in psychometrics. This includes discussing its promising uses, limitations, challenges, ethical questions, and practical considerations. The creation of this report involved a selective literature review of current trends, challenges, and opportunities related to the use of GenAI in psychometrics. Additionally, interviews were conducted with psychometric professionals to gather insights and perspectives on the integration of this technology in their work. The report:

- Synthesizes key findings and recommendations
- Provides an understanding of the role of GenAI models in psychometrics
- Identifies potential risks and benefits
- Proposes strategies for leveraging GenAI ethically and responsibly in testing practices

This analysis is valuable for several reasons. Firstly, it provides insights into the current usage of GenAI in test development and psychometric activities, helping assessment professionals stay informed about emerging trends and technologies. Secondly, it offers practical recommendations for integrating GenAI into existing testing practices, ensuring compliance with ethical and security standards. Finally, it helps our industry make informed decisions about the adoption of GenAI tools, balancing innovation with risk management.

Exploring Generative AI Models

Generative AI (GenAI) is a subset of artificial intelligence (AI) that's based on patterns and information learned from large datasets through a concept called **machine learning** (ML). These models are trained using unsupervised learning, an ML method where the models learn to generate new content (e.g., text, images, audio) without explicit examples or labels.

MACHINE LEARNING

is a form of AI that gives computer systems a way to analyze data, perform tasks, and “learn” without explicit instruction.

A key characteristic of GenAI models is the ability to produce content that is indistinguishable from content created by humans. This is achieved through neural networks, which are computational models inspired by the structure and function of the human brain. These networks consist of layers of interconnected nodes that process information and learn to create new content by adjusting the strength of the connections between nodes based on the input data.

GenAI models—such as ChatGPT, Bard, Gemini, or Bloom—are particularly effective at interpreting, summarizing, and generating human language. They can analyze text, images, and data to identify patterns, understand context, and generate coherent responses. This makes them well-suited for a wide range of applications, including chatbots, language translation, and content generation.

In recent years, GenAI models have made significant advancements, thanks in part to the availability of large datasets and improvements in computing power. These advancements have led to the development of models with impressive capabilities, such as generating realistic images, composing music, and even writing code.

GenAI models represent a powerful tool for generating new content and have the potential to revolutionize many fields, including psychometrics. Despite their capabilities, these models are not without limitations. They can sometimes produce outputs that are nonsensical, inaccurate, or inappropriate, especially when the input data is ambiguous or the model is not properly trained. Additionally, there are ethical considerations to take into account, such as the potential for bias in the training data or the misuse of AI-generated content.

Performance of GenAI as a Test-Taker in High-Stakes Credentialing

Evaluating ChatGPT's Ability to Pass a Multiple-Choice Exam

To understand how generative AI (GenAI) models can be applied in psychometrics, we must first consider how they would perform as test-takers. In simpler terms, if ChatGPT were to take a licensure exam with multiple-choice questions, would it pass?

GenAI models, particularly **large language models** (LLMs) like ChatGPT, have shown promising performance as test-takers in various medical licensure and certification exams that use multiple-choice questions. These exams included the:


- United States Medical Licensing Examination (USMLE)
- National Medical Licensing Examination in Japan
- Association of Social Work Boards (ASWB) Licensing Exams
- Dental Licensing Examinations
- German Medical Licensing Examination
- Registered Nurse License Exam in Taiwan

These models have demonstrated the ability to understand complex medical and social work-related text patterns as well as generate rationales aligned with safe and ethical practice.

In a 2023 study,¹ ChatGPT was evaluated for its test-taking performance on the USMLE. The results indicated that ChatGPT achieved a high accuracy rate in answering questions,² performing at or near the passing threshold of 60% accuracy.

LARGE LANGUAGE MODELS

are a form of machine learning trained on massive data sets in order to recognize, comprehend, analyze, and generate human language.



Similarly, in the National Medical Licensing Examination in Japan, ChatGPT 4.0 was found to reach the passing standard, providing accurate and relevant answers to exam questions. ChatGPT 4.0 was also able to pass the US and UK dental licensing examinations,³ the German medical licensing examinations⁴, and the Registered Nurse License Exam in Taiwan.⁵

The ASWB Licensing Exams present a unique challenge due to the nature of social work practice, which requires a deep understanding of human behavior and social systems. In a 2023 study,⁶ ChatGPT was used to answer ASWB practice questions, with results showing that the model could provide rationales for its answers that were aligned with social work principles and ethics. The study also found that ChatGPT's rationales were communicated in simple language likely to be comprehensible to a wide audience, indicating its potential use as a study aid for social work students.

The Evolution of ChatGPT as a Test-Taker

Some studies have compared the performance of ChatGPT versions 3.5 and 4, highlighting clear advancements in the latter. While ChatGPT 3.5 demonstrated valuable insights in its explanations, it also made notable errors, particularly in providing incorrect or inconsistent information. These mistakes, often due to misinterpretation or **AI hallucinations**, revealed gaps in its medical knowledge and reasoning.

In contrast, ChatGPT 4.0 showed significant improvements in both accuracy and reliability. Recent studies reported that ChatGPT 4.0 not only passed several medical licensure exams but also exhibited sound clinical reasoning and more consistent responses,⁷ suggesting its potential as a valuable tool in medical education and decision-making.⁸

AI HALLUCINATIONS

are incorrect, misleading, or false results generated by an LLM.

Recent studies reported that ChatGPT 4.0 not only passed several medical licensure exams but also exhibited sound clinical reasoning and more consistent responses, suggesting its potential as a valuable tool in medical education and decision-making.

Content and Quality of ChatGPT Decision-Making Rationales and Validity

In addition to assessing ChatGPT's ability to answer multiple-choice questions in licensure exams, many of the studies reviewed also investigate its capacity to offer rationales and reasons to support its decisions as well as evaluate the quality and validity of the rationales provided. A 2023 study focused on its ability to recognize social work-related text patterns and produce rationales aligned with professional practice standards.⁹ The researchers found that ChatGPT was able to provide several reasons to support its decisions, often in a way that was simple and understandable to a wide audience. These rationales demonstrated an understanding of professional practice standards and ethical guidelines, suggesting that ChatGPT has the potential to assist in decision-making processes in professional settings.

While its rationales were generally of high quality, there were instances where they were not as comprehensive or accurate as those provided by human experts. This highlights the need for further research and development to improve the quality and validity of AI-generated rationales in psychometrics.

Overall, ChatGPT was able to produce rationales that were aligned with professional practice standards. This suggests that the model has the potential to assist in decision-making processes in licensure exams and could be used by test-takers as a study tool. Additionally, the content and quality of decision-making rationales generated by ChatGPT suggest that GenAI models in general could enhance the field of psychometrics by providing another perspective on test items.

Use of Generative AI Models in Assessment Development and Psychometrics

We've noticed a literature gap regarding the practical use, effectiveness, accuracy, and validity of using GenAI models specifically in psychometric activities. As we explore potential applications of GenAI models in psychometrics, we'll be examining key activities such as job task analysis (JTA), competency development and blueprinting, item bank management, item writing, exam development, standard setting, statistical analysis, and reporting. Each of these steps in the assessment life cycle plays a crucial role in ensuring the validity, reliability, and fairness of assessments, making them ideal for GenAI integration to enhance efficiency and effectiveness.

Job Task Analysis

JTA is a systematic process that identifies and documents the tasks, knowledge, skills, and abilities required for competent and safe performance. GenAI models have the potential to play a significant role in enhancing the efficiency and effectiveness of JTA by automating various aspects of the process.

Key Areas Where GenAI Could Be Used in JTA

Identifying Role-Relevant Information: GenAI models could be used to analyze and summarize literature, standards, and educational programs related to the job or role being analyzed. These models could scan and process large volumes of text to identify relevant information such as job descriptions, educational requirements, and industry standards. This automated analysis could help in identifying key competencies and tasks associated with the job, providing a solid foundation for the JTA process.

Generating a List of Competencies: GenAI could generate a first draft of the list of competencies based on the analyzed data. The models could identify common themes and patterns in the data to suggest potential competencies required for the job. While these draft lists would still require human review and validation, they could serve as a starting point for further analysis, saving time and effort in the initial stages of JTA.

Potential Challenges of Using GenAI for JTA

Using GenAI models for JTA could offer benefits like increased efficiency, but it also comes with challenges, including:

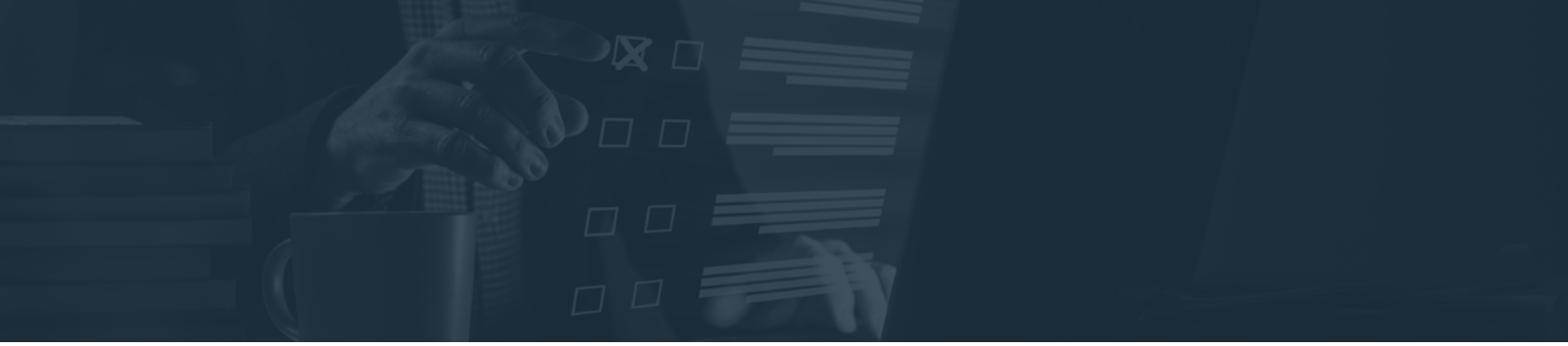
- GenAI may struggle to understand the nuanced context of job roles, leading to inaccuracies.
- Bias in training data can result in biased outcomes.
- Some job roles are complex, requiring human judgment that GenAI models may lack.
- GenAI may miss human insights and tacit knowledge.
- The accuracy of GenAI in JTA would depend on data quality.
- Subject matter experts (SMEs) and regulatory compliance may resist the use of GenAI.

Item Writing

Item writing is a fundamental aspect of assessment development, involving the creation of test items that accurately measure the knowledge, skills, and abilities of individuals. In the traditional process, items are typically written by SMEs based on their expertise in the field. This approach is time-consuming and costly, requiring significant input from both psychometricians and SMEs. GenAI, however, can potentially improve the efficiency of item writing and significantly reduce the time and cost associated with developing new items. In particular, it can assist in generating item drafts, reviewing existing questions, writing rationales, and finding references.

Key Areas Where GenAI Can Enhance Item Writing

Generating Item Drafts: As part of the initial stage, AI models can potentially analyze lists of competencies being tested alongside large datasets of existing items and use this information to generate new, original items. These drafted items can then be reviewed and edited by SMEs to ensure their accuracy and relevance. This automation can significantly speed up the item development process, allowing for the creation of a larger number of items in a shorter amount of time.



Reviewing Existing Items: GenAI can be used to review questions that have already been written by SMEs. Models can identify potential issues such as ambiguity, bias, or lack of clarity by analyzing the questions' content and structure. This automated review process can help ensure that the items meet the required standards for validity.

Writing Rationales for Items: Rationales explain why a particular answer is correct or incorrect, helping test-takers understand the reasoning behind the correct answer. This task is time-consuming for SMEs, but GenAI models can analyze the content of the items and write rationales that are clear, concise, and aligned with the intended audience.

Finding or Updating References and Supporting Materials: GenAI can streamline this process to help ensure that items are based on sound, evidence-based principles and are relevant to the intended content area. However, it is important to note that any copyrighted sources or information housed behind firewalls will not be accessible by most GenAI models.

Potential Challenges of Using GenAI in Item Writing

Psychometricians and test development professionals have valid concerns about the use of AI-generated items, including:

- AI-generated items may lack the creativity and nuanced understanding of human writers, potentially leading to technically accurate but shallow items.
- Bias in the training data can result in biased items that disadvantage certain groups.
- AI can assist in generating draft items, but human oversight is necessary to ensure quality and relevance.
- Implementing GenAI for item writing requires significant investment in technology and resources.
- Ethical questions arise regarding the delegation of decision-making tasks to AI and the need to protect security and privacy.
- Providing information or creating content through GenAI could potentially be accessible to the public if the appropriate guardrails are not put in place.

Item Bank Management and Maintenance

The organization, maintenance, and analysis of item banks is a critical yet time-consuming process for psychometricians. It requires careful manual oversight to track item performance, avoid duplication, and update content, but the integration of GenAI could streamline these tasks in several ways.

Key Areas Where GenAI Could Help With Item Banks

Assigning Tags and Keywords: GenAI models could analyze the content of item banks and automatically assign tags or keywords to items based on their content. This could help in quickly identifying and retrieving items that are relevant to specific topics or concepts, making it easier to construct exams that are aligned with the intended learning outcomes.

Identifying Duplicates: Analyzing the content of the item bank, GenAI models could identify duplicate or similar items. This would help ensure the quality and diversity of the item bank by removing redundant or outdated items.

Exposing Item Bank Gaps: GenAI could highlight areas where new items need to be developed to ensure adequate coverage of the content domain. GenAI could also assist psychometricians with the development of item writing targets.

Analyzing Item Performance Data: GenAI models could analyze the performance of items in previous exams to identify items that may be too easy or too difficult, items that may be biased toward certain groups, or items that may be functioning poorly for other reasons. This information could help in improving the quality and fairness of exams constructed from the item bank.

Potential Challenges of Using GenAI for Item Bank Management and Maintenance

Despite these benefits, there are also challenges and limitations associated with the use of GenAI in item bank management. One challenge is ensuring the accuracy and reliability of GenAI models in categorizing items, identifying duplicates, and analyzing item performance, especially since their limitations in interpreting the nuances of human language and context can impact their accuracy.



Exam Form Assembly

Exam form assembly is a complex process that involves selecting questions from an item bank, ensuring that they align with the blueprint parameters, and analyzing the content to flag enemies or similarities in content. GenAI can potentially play a role in streamlining and improving this process.

Key Areas Where GenAI Could Assist in Form Assembly

Selecting Items: An important aspect of exam development is selecting questions that meet the requirements of the exam blueprint, which outlines the content areas and cognitive levels that the exam should cover. While we don't envision GenAI models to automatically select questions from the item bank to match the blueprint parameters, they could be used as a secondary tool to analyze the exam form and verify a proper fit with the blueprint parameter. This could add an additional layer to support the exam's validity.

Diversification: GenAI models could help ensure the selected questions are diverse and cover a wide range of topics and difficulty levels. They could analyze the content of the selected questions to identify any enemies or similarities in content. This could help exam developers make sure the exam is fair and balanced.

Analyzing Content: GenAI could evaluate the readability of the questions to determine whether they're suitable for the target population. The complexity of the questions could also be analyzed to ensure they're neither too easy nor too difficult.



Potential Challenges of Using GenAI for Exam Form Assembly

Overreliance on automation in this stage of test development can pose numerous challenges:

- Balancing test forms for factors like difficulty and content coverage can be complex and may require extensive fine-tuning and validation.
- Human oversight and quality control are essential for ensuring fairness, validity, and reliability. Certification and licensure exams must undergo a rigorous process to establish content validity. This ensures the exam adequately covers all relevant topics and tasks within the domain, aligning with the knowledge, skills, and abilities required for the profession. The validation process often involves input from SMEs to confirm accuracy and reliability in assessing candidates' competencies, and GenAI does not possess such nuanced skills yet.
- Biases in both the data used to train GenAI models, as well as the models themselves, can result in biased test forms and limit the models' ability to detect bias in items.
- Security and privacy concerns are also paramount, as GenAI systems must adhere to strict standards that help protect sensitive test content.
- The cost and resources required to implement GenAI for test form assembly can be significant.
- Lack of model transparency and ethical considerations regarding the delegation of decision-making tasks to GenAI are additional challenges that need to be addressed.

For these reasons, we would still recommend that test form assembly be done by psychometricians. However, GenAI could be used to perform an analysis of the content to flag potential **enemies** or problematic items.

Statistical Analysis

Analysis involves interpreting statistical data related to the performance of exam questions and the exam as a whole. GenAI can automate some aspects of this process by flagging questions that may need further review.

ENEMY ITEMS

are any two (or more) test questions that should not appear in the same exam form.

Key Areas Where GenAI Could Aid Statistical Analysis

Interpreting Item Statistics: An important part of analysis involves the interpretation of item statistics, such as item difficulty and discrimination. Item difficulty refers to how easy or hard an item is for examinees, while discrimination refers to the extent to which an item can differentiate between high- and low-performing examinees. GenAI could potentially analyze these statistics to identify items that may be too easy or too difficult, or items that are not effectively discriminating between examinees. By flagging these items, GenAI could help exam developers and SMEs make informed decisions about which items to retain, revise, or remove from the exam.

Analyzing Overall Performance: GenAI could also analyze the overall performance of the exam, including the reliability and validity of the exam scores. Reliability refers to the consistency of the exam scores, while validity refers to the extent to which the exam measures what it is intended to measure.

Identifying Other Issues: GenAI could flag questions that may be problematic for other reasons, such as potential bias or cultural insensitivity. It could analyze the question content to identify any potential issues and flag them for further review by human experts.

Potential Challenges of Using GenAI for Statistical Analysis

GenAI's limitations in statistical analysis or interpretation of items and exams include:

- It may interpret statistical results differently from human experts, leading to potential biases in the analysis. Some statistical concepts and analyses are complex and may require human expertise for accurate interpretation. In those cases, GenAI may struggle with contextual understanding, especially in complex statistical concepts.
- GenAI models are usually based on pre-defined rules and patterns. This may limit their ability to adapt to new or unexpected statistical scenarios, potentially leading to misinterpretations of results.

While GenAI could automate certain aspects of statistical analysis, human oversight and expertise are still essential for ensuring the accuracy, validity, and reliability of statistical interpretations. Because of this, we recommend that psychometricians continue analyzing and interpreting exam-related statistics, perhaps using GenAI to flag potentially problematic items.

Translation and Test Adaptation

Translation and test adaptation are crucial aspects of exam development. GenAI already plays a significant role in facilitating the translation and adaptation process, making it more efficient and accurate.



Key Areas Where GenAI Could Enhance Test Translation and Adaptation

Translating Exam Language: One way in which GenAI can be used for translation is by providing automated translation services. AI-powered translation tools, such as DeepL Translator or IBM Watson Language Translator, can quickly translate exam questions and content into multiple languages, allowing exams to be administered to a more diverse group of candidates. These tools can also help ensure consistency in translation across different languages, reducing the risk of errors or misinterpretations.

Translating Cultural Nuances: GenAI could also be used to adapt exams to different cultural contexts or educational systems. For example, models could analyze exam content and identify cultural references or language nuances that may not be appropriate for certain cultural or linguistic groups. By flagging these issues, GenAI could help exam developers modify the content to be more culturally sensitive and relevant to diverse populations.

Adapting Existing Items: GenAI can help modify exam questions to suit different linguistic or cultural contexts by analyzing their structure and content before suggesting modifications. This could help ensure the exam questions are fair and accessible to all candidates, regardless of their background or language proficiency.

Potential Challenges of Using GenAI for Translation and Adaptation

Accuracy in this area extends beyond words to include cultural nuances and context—elements that are challenging for existing GenAI models. Here are a few specific challenges that may arise from its use:

- AI-powered translation tools—while significantly improved—may still produce inaccurate or misleading translations, particularly for complex content.
- GenAI may struggle to capture the nuances of language—including idiomatic expressions, cultural references, and regional variations—leading to translations that are not culturally appropriate or easily understood by the target audience.
- GenAI may produce translations that are contextually incorrect or confusing because it doesn't fully grasp the context of exam questions or content.
- AI-powered translation tools may not support all languages or region-specific dialects (e.g., Canadian French or Colombian Spanish), which can be a limitation for exams that need to be translated into less common languages or those with unique regional linguistic differences.
- Using GenAI for exam translation may raise issues regarding the protection of sensitive exam materials.
- Despite advancements in GenAI, human review by a professional translator and bilingual SMEs is still necessary to ensure the quality and accuracy of translated exam content.

Standard Setting

Standard setting is the step in which the passing score or cut score for an exam is established. GenAI could play a role in standard setting by helping generate a draft of the minimally competent candidate (MMC) based on a review of educational requirements, program curricula, and lists of competencies.

Key Areas Where GenAI Could Assist in Standard Setting

Identifying Core Knowledge, Skills, Abilities, or Competencies: GenAI could analyze educational requirements, program curricula, lists of competencies, and professional standards to identify the essential knowledge, skills, abilities, or competencies needed for a given profession.

Generating Draft Profiles and Performance Descriptors: Based on the core competencies uncovered, GenAI models could then generate a draft profile of the MMC and create performance-level descriptors—which define the minimum level of competence required to practice in the field—alongside a description of a non-competent and highly competent candidate. This draft profile could then serve as a reference for SMEs during standard setting activities.

Analyzing Competency Lists or JTAs: GenAI could review lists of competencies or job task analyses (JTAs) to identify the key competencies that are relevant to the profession. This analysis could help exam developers and SMEs understand the skills and knowledge that are necessary for successful performance in the field. This information could then be used to inform the standard setting process and ensure that the passing score reflects the skills and knowledge required for competent practice.

Potential Challenges of Using GenAI for Standard Setting

Concerns about using GenAI models for standard setting activities include:

- GenAI models can lack transparency and explainability, making it difficult to understand how standards are set and whether the results can be trusted. This lack of transparency may lead to questions about the validity and fairness of the standards set by these models.
- They may introduce biases into the standard setting process, particularly if the training data used to develop them is biased. This could result in standards that are not representative of the population being assessed or that unfairly advantage or disadvantage certain groups.
- The standard setting process is complex, requiring human judgment and expertise to make nuanced decisions. GenAI may struggle to replicate this human judgment, leading to standards that are not appropriate or meaningful.

Reporting

Reporting involves summarizing the methodology, activities, objectives, main results, and technical details of an assessment. Psychometricians often spend considerable time creating these reports because they are essential for customers and are often used during their accreditation process.

Key Areas Where GenAI Could Be Used for Reporting

GenAI could make some aspects of the reporting process more efficient and less labor-intensive for psychometricians. Here are two such aspects:

Generating Summary Reports: GenAI could be used to analyze assessment data and automatically generate summary reports based on a pre-defined template. These reports could include information about the test development process, item analysis, test equating, and standard setting. GenAI could be used to generate tables, charts, and graphs to illustrate key findings, making it easier for stakeholders to interpret the results.

Preparing Statistical Reports: GenAI could help prepare statistical reports by analyzing data and identifying trends or patterns that may be of interest to stakeholders. It could also help pinpoint outliers or anomalies in the data, which can be important for ensuring the validity of the assessment results.

Potential Challenges of Using GenAI for Reporting

By leveraging GenAI technologies, psychometricians could streamline the reporting process and focus their efforts on more strategic tasks, such as interpreting results and improving assessment quality.

However, we do have several concerns:

- There is always a risk of errors in the data or in the models used to analyze the data, which could lead to incorrect or misleading reports.
- AI-generated reports may lack the human touch and contextual understanding that psychometricians can provide, potentially leading to reports that are less insightful or actionable.
- AI models are trained on data, and if this data is biased, the reports generated by them may also be biased. This could result in reports that reflect and perpetuate existing biases and stereotypes, leading to unfair or discriminatory outcomes.
- Due to concerns about transparency and accountability, it may be difficult to understand how GenAI arrived at its conclusions. This could raise questions about the validity and reliability of the reports, particularly in high-stakes situations.

Legal and Ethical Considerations of Using Generative AI in Psychometrics

The integration of generative AI (GenAI) into psychometric activities introduces complex legal and ethical considerations that must be carefully navigated. A primary consideration is the potential infringement of copyright laws. GenAI models can be trained on vast amounts of text, including copyrighted material. This raises questions about the originality and ownership of the generated content. Organizations that are contemplating the use of GenAI in their activities must ensure that all content produced complies with copyright laws and does not violate intellectual property rights.

ISSUES THAT COULD IMPACT ASSESSMENT INTEGRITY, VALIDITY, RELIABILITY, AND EQUITY:

- Infringement of copyright laws
- Introduction or amplification of bias
- Deviation from relevant competency measurements
- Dissemination of inaccurate or false information

Another key consideration is its impact on the validity and fairness of assessments. GenAI models are not immune to bias and can inadvertently perpetuate or amplify existing biases present in the data they are trained on. This raises concerns about the fairness and equity of assessments generated using AI. Psychometricians and subject matter experts (SMEs) must carefully review and validate AI-generated content to ensure it's free from bias and accurately reflects the knowledge, skills, and abilities being measured.

Additionally, there is a risk that GenAI may produce false data or generate inaccurate resources. Organizations using GenAI must be vigilant in verifying the input and output of the systems they use to prevent the dissemination of incorrect information that could undermine assessment validity and reliability.

While GenAI can streamline certain aspects of assessment development, as previously explored, human oversight and verification are essential. Psychometricians and SMEs play a critical role in reviewing and validating content to ensure its accuracy, relevance, and alignment with professional standards. This human oversight is crucial for maintaining the integrity and validity of assessments.

Principles for Responsible Use of GenAI in Test Development

Ensuring adherence to psychometric standards, test integrity, security, and fairness is paramount when integrating generative AI (GenAI) models into psychometrics. They must be implemented responsibly to maintain the validity and reliability of assessments. Here are some key considerations:

Psychometric Standards: GenAI applications in psychometrics should align with established psychometric standards and guidelines, such as those set by the American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), National Commission for Certifying Agencies (NCCA), or ISO/IEC 17024. These standards ensure that assessments are valid, reliable, and fair. They can also be used as input for GenAI models so they're trained on data that aligns with psychometric standards. For example, training data can include items that have been validated by psychometricians and subject matter experts, ensuring that AI-powered tools learn to generate items that meet the required standards.

Security: GenAI systems used in psychometrics should adhere to strict security protocols to protect sensitive test content and data. This includes encryption of data, secure storage practices, and access controls to prevent unauthorized use or disclosure.

Fairness: GenAI models should be designed and validated to ensure fairness across diverse populations. This includes monitoring for biases in GenAI models and adjusting models as necessary to mitigate these biases.

Validation: All GenAI-driven processes should undergo rigorous validation so they produce results that are comparable to traditional psychometric methods. This may involve conducting validation studies to demonstrate the validity and reliability of AI-generated items or scores.

Transparency: GenAI systems used in psychometrics should be transparent, with clear documentation of how decisions are made. This transparency helps ensure accountability and allows stakeholders to understand and trust the results produced.

Integrating GenAI models into psychometrics requires a careful balance between leveraging the benefits of the technology and maintaining the standards of the field. GenAI can be a valuable tool for enhancing the validity and reliability of psychometric assessments if it adheres to psychometric standards, test integrity, security, and fairness.

Moving Forward With Generative AI

The use of generative AI (GenAI) in psychometrics presents both challenges and opportunities for the field. As AI technologies continue to evolve, those in professional credentialing and assessment must stay informed about emerging trends and developments to effectively leverage its potential benefits.

A Call to Action for Test Development Professionals

Continued Investigation: Ongoing research and development are needed to explore the full potential of GenAI in psychometrics. This includes investigating new models and techniques that can enhance the efficiency, accuracy, and fairness of psychometric assessments.

Community Collaboration: There must be collaboration and partnership between psychometricians, AI experts, and other stakeholders to advance the use of GenAI. By working together, these groups can identify new opportunities for GenAI integration, address ethical and legal considerations, and develop best practices for its implementation in psychometric assessment development.

Steadfast Allegiance to Standards: Those involved in assessment must continue to uphold professional standards and guidelines to ensure that GenAI is used ethically and responsibly in assessment development. This includes transparency in GenAI decision-making processes, validation of AI-generated content, and mitigation of bias in GenAI models.

In conclusion, the integration of GenAI in psychometrics has the potential to transform the field by offering new capabilities and efficiencies. However, moving forward with this integration requires a thoughtful and collaborative approach to address the challenges and opportunities presented by this technology. By embracing GenAI responsibly, psychometricians can enhance assessment quality, validity, and fairness, ultimately benefiting both candidates and the broader field of psychometrics.

Citations

1. Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2), e0000198.
2. Yanagita, Y., Yokokawa, D., Uchida, S., Tawara, J., & Ikusaka, M. (2023). Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. *JMIR Formative Research*, 7, e48023.
3. Chau, R. C. W., Thu, K. M., Yu, O. Y., Hsung, R. T. C., Lo, E. C. M., & Lam, W. Y. H. (2024). Performance of generative artificial intelligence in dental licensing examinations. *International Dental Journal*.
4. Meyer, A., Riese, J., & Streichert, T. (2024). Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. *JMIR Medical Education*, 10, e50965.
5. Huang, H. (2023, October). Performance of ChatGPT on registered nurse license exam in Taiwan: a descriptive study. In *Healthcare* (Vol. 11, No. 21, p. 2855). MDPI.
6. Victor, B. G., Kubiak, S., Angell, B., & Perron, B. E. (2023). Time to move beyond the ASWB licensing exams: Can generative artificial intelligence offer a way forward for social work? *Research on Social Work Practice*, 33(5), 511-517.
7. Yanagita, Y., Yokokawa, D., Uchida, S., Tawara, J., & Ikusaka, M. (2023). Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. *JMIR Formative Research*, 7, e48023.
8. Meyer, A., Riese, J., & Streichert, T. (2024). Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. *JMIR Medical Education*, 10, e50965.
9. Victor, B. G., Kubiak, S., Angell, B., & Perron, B. E. (2023). Time to move beyond the ASWB licensing exams: Can generative artificial intelligence offer a way forward for social work? *Research on Social Work Practice*, 33(5), 511-517.