DOI: 10.1002/tl.20631

ORIGINAL ARTICLE

WILEY

Enhancing qualitative research in higher education assessment through generative AI integration: A path toward meaningful insights and a cautionary tale

Ruth C. Slotnick | **Joanna Z Boeing**

Office of Assessment, Bridgewater State University, Bridgewater, Massachusetts, USA

Correspondence
Ruth C. Slotnick, Bridgewater State University, Bridgewater, MA, USA.
Email: rslotnick@bridgew.edu

Abstract

This study explores the use of generative AI, specifically Google's Bard and OpenAI's ChatGPT, to enhance qualitative research within higher education assessment, focusing on institutional assessment practitioners. Using a dataset focused on diversity, equity, and inclusion (DEI) from annual faculty assessment reports, we tested traditional analytical methods and compared them to AI-assisted techniques, with a particular emphasis on AI's capacity to improve qualitative analysis. By exploring AI's benefits and limitations in qualitative assessment, we not only advocate for the thoughtful integration of AI technologies but also underscore the critical importance of human expertise in maintaining the depth and integrity of qualitative inquiry. We present a step-by-step practical guide for the assessment practitioner to integrate AI into the qualitative research process, highlighting AI's potential to deepen insights while upholding research integrity and emphasizing the necessity of human oversight.

INTRODUCTION

The integration of generative AI (GenAI) in higher education presents exciting opportunities, especially for institutional assessment practitioners. These professionals, instrumental in using program-level data to drive improvements, could gain from AI's ability to efficiently manage large datasets and expedite quantitative and qualitative analysis. In this article we focus on qualitative analysis, acknowledging the unique challenges that come with incorporating AI into qualitative assessment practices. Such challenges require a thoughtful and systematic approach to ensure the preservation of traditional qualitative inquiry's integrity and depth. Addressing this need, our study examined the application of advanced AI tools—namely, Google's Bard (renamed Gemini in early 2024;

 © 2024 Wiley Periodicals LLC.

https://gemini.google.com) and OpenAI's ChatGPT-3.5 (https://chat.openai.com)—to qualitative analysis through a methodological lens that combines reflective thematic analysis with AI-driven insights.

Broad discussions about the potential of AI in the field of higher education assessment are occurring. This study embarked on a more granular exploration, systematically testing AI's effectiveness in analyzing a subset of qualitative data drawn from annual faculty assessment reports at a medium-sized comprehensive public university in the northeastern United States, where we work in an office of assessment. After presenting the results, we outline the policy and practice implications of integrating AI into qualitative research. Through a focused examination, we delineate the benefits and limitations of AI in higher education assessment, moving beyond general considerations to offer grounded insights. This approach underscores the need for such research, as discussions about what AI can do often outweigh actual testing of the tools. The article concludes with a step-by-step guide for conducting qualitative research with AI, offering strategic steps for thoughtful AI integration into assessment practices. This guide ensures that the possible benefits and pitfalls of AI are considered, informing future practice and policy development in the assessment field.

TERMINOLOGY, STUDY CONTEXT, AND AI TOOLS

For this article, "assessment practitioners" refers to higher education professionals, specifically institutional staff members who collaborate with departmental faculty or divisional units to evaluate student learning outcomes. This role is vital for aligning educational outcomes with student achievements and institutional goals. Additionally, we refer to diversity, equity, and inclusion (DEI) broadly to encompass a range of equity dimensions, including but not limited to antiracism, decoloniality, diversity, equity, inclusion, and racial justice, which aligns the definitions and understandings adopted by our institution.

Our study used the DEI-focused questions from the annual assessment reports for analysis because the open-ended items allowed for free responses and testing of the AI tools' knowledge base of DEI-related resources. We tested ChatGPT-3.5 (September 11, 2023, update, free version) and Bard (September 27, 2023, version). These AI tools were chosen for their accessibility, widespread use, and relevance for data analysis tasks. Although we evaluated the capability of other platforms, such as Microsoft Bing's AI (now referred to as Copilot; https://copilot.microsoft.com), they were too limited or unsuitable for the qualitative analysis required for this study. It is important to note that the sophistication and capabilities of AI tools continue to rapidly evolve.

LITERATURE REVIEW

This literature review focuses on the use of qualitative research by the assessment practitioner, the opportunities and challenges that AI integration pose for campus practitioners, the effectiveness of AI in qualitative data analysis based on existing research, and the ethical considerations vital for successful human–AI collaborations in qualitative analysis.

Qualitative research and the role of assessment practitioner

The evaluation of academic programs is critical for ensuring teaching and learning quality in higher education. Many assessment scholars have addressed the value of qualitative research for the assessment professional. Assessment practitioners, acting as methodologists (Jankowski & Slotnick, 2015), are instrumental in gathering and analyzing data to evaluate program outcomes, identify improvement areas, and guide strategic decision-making, ultimately enhancing student success. Qualitative research, offering insights into the lived experiences of faculty, students, and stakeholders, is a key tool for assessing the quality of student learning. Despite its value, Suskie (2009) noted that qualitative methods remain "underused and underappreciated in many assessment circles" but can add a "human dimension" relevant to course and program levels (pp. 32–33). Echoing this sentiment about humanizing assessment, Maki (as cited in Weave Education, 2023) called for a shift away from numeric-centric reporting toward exploration of diverse pathways for understanding students' learning experiences. Levin (2023) reinforced this point, stating that "we simply can't 'numbers our way' into understanding students' unique challenges, especially those from under-resourced communities" (para. 3). Qualitative research is important for capturing nuanced narratives of student growth and development, which are often missed by quantitative approaches. The challenge often lies in the time-intensive nature of qualitative analysis, especially for assessment professionals who frequently work in isolation (Nicholas & Slotnick, 2018), which highlights the potential role of AI as a supportive tool.

Opportunities and challenges for AI in higher education assessment

The rapid evolution of AI has increased its potential to aid in the technical aspects of assessment work. Jankowski (as cited in Janio, 2023) and Henning (2023) outlined AI's utility in streamlining tasks such as creating assessment reports, providing feedback, summarizing accreditation progress, assisting with various analytic tasks such as statistical and thematic analysis, and creating rubrics. Shea and Walton (2023) demonstrated AI's applications in assessment at a state-level conference, highlighting its speed, multi-paragraph input processing, and long memory as key benefits for assessment practitioners. More recently, Walton et al. (2024) showcased a campus chatbot, capable of providing administrative and strategic plans upon request, and assisting with syllabus checking, in-class project development, and student feedback. Singer-Freeman (2023) similarly focused on use of AI in classroom-based assessment strategies, particularly on how to use findings to implement immediate pedagogical improvements, increase equity, and foster innovation. However, she also noted challenges affecting AI's dependability such as response inconsistency and inability to retrieve data when internet access is limited.

Insights from Jankowski and Henning (2024) regarding the application of GenAI in student affairs assessment are pertinent to this article. They explored the use of chatbots for tutoring, retention efforts, policy inquiries, facilities management, and predictive analytics. Highlighting the user-friendly nature and broad applications of tools such as ChatGPT, the authors also cautioned users to be mindful of academic integrity issues with AI-generated content. They emphasized the need to verify outputs from AI tools, due to potential biases in training data. Despite these challenges, Jankowski and Henning suggested that AI could assist student affairs professionals with, for example, drafting and revising learning outcome statements and job descriptions.

As AI continues to reshape the landscape of academic and student affairs assessment, its implications for qualitative research practices demand closer scrutiny. Amid the efficiency gains and broad applicability of AI in enhancing educational assessment, the intersection of AI with qualitative analysis introduces unique challenges.

Balancing AI advancements with human expertise in qualitative research

AI offers a promising opportunity to streamline qualitative analysis for assessment practitioners. For example, researchers have found notable efficiency gains when using AI to replace manual coding (Gao et al., 2023; Jiang et al., 2021; Zhang, Wu, Xie, Kim, et al., 2023). AI can also assist with thematic analysis (Zhang, Wu, Xie, Lyu, et al., 2023; Zhao et al., 2021). However, its overreliance on efficiency and speed might erode coding creativity, nuance, and diversity, possibly compromising the depth of insights provided by traditional qualitative analysis (Feuston & Brubaker, 2021; Gao et al., 2023; Jiang et al., 2021). As in non-AI-assisted research, the voice of minoritized populations is also at risk of being overlooked. AI struggles to capture the subtleties of human experiences and may miss "contextualized insights" (Jiang et al., 2021, p. 20) and edge cases with a small *n* within the datasets (Feuston & Brubaker, 2021). Novice qualitative researchers might even inadvertently prioritize AI-generated outputs without critically evaluating their context or considering the unique insights gained from traditional qualitative analysis (Gao et al., 2023). Other known drawbacks of using AI are the time and skill required for prompt engineering, which is necessary for ensuring high-quality AI-generated responses, and the time required to review responses for thematic analysis (Zhang, Wu, Xie, Lyu, et al., 2023).

Additionally, researchers have found that AI introduces bias, repetition, and generic responses, raising questions about the validity of its output (Feuston & Brubaker, 2021; Zhao et al., 2021). Moreover, each platform contains developer bias (Baidoo-Anu & Ansah, 2023). As a recent article in *Harvard Business Review* stated, "Bias can creep into algorithms in several ways. AI systems learn to make decisions based on training data, which can include biased human decisions or reflect historical or social inequities, even if sensitive variables such as gender, race, or sexual orientation are removed" (Manyika et al., 2019, para. 4). Even more alarming, "new research suggests human users may unconsciously absorb these automated biases" (Leffer, 2023, para. 2). Therefore, although AI tools offer promising advancements in efficiency, their integration into qualitative research should be carefully mediated to ensure they complement, rather than replace, human expertise. As Feuston and Brubaker (2021) asserted, AI should "augment rather than automate human qualitative analytic practices" (p. 5). One way to mediate their use is employing Lubars and Tan's (2019) delegation framework, which is often cited in task delegation with AI. This framework acknowledges critical factors in AI delegation, including motivation for a task, difficulty of the task, risk perception, and trust in AI.

Ethical considerations and the human–AI partnership

Assessment scholars' call for finding the equilibrium between human interpretation and the collaborative role of AI in qualitative research holds significance for assessment practitioners who handle sensitive institutional data, making the ethical considerations of high importance. This ethical awareness is emphasized in Montenegro and Henning's (2022) examination of the relationship between research paradigms and methodologies adopted by assessment practitioners in campus-based research for improving academic programs and student learning. Acknowledging the inherent bias embedded within the assessment practices of the higher education system, the authors advocated for assessment practitioners to adopt an equity lens. They emphasized the crucial role of assessment as a tool not only for cultivating equitable educational experiences but also for promoting equality through fair assessment practices to actively address and dismantle barriers to student success. Furthermore, they stressed that this commitment entails an ongoing process of questioning, reflecting on, and improving practices, policies, and perspectives, ultimately positioning the removal of barriers to student success as an imperative act of social justice. In addition to ethical considerations of bias and barriers, the ethics of the use of AI tools must also be addressed. Assessment practitioners must ensure that data are protected and deidentified, and that institutional policies are adhered to prior to uploading any data to AI cloud services.

METHODOLOGY

This study, conducted over a 2-week period, used a dataset from the university's annual assessment system. The dataset was primarily composed of qualitative responses, but also included quantitative data presented in tables and charts. The analysis focused on responses to questions about the integration of DEI within program-level assessment efforts. To ensure privacy, all identifying information was removed from the dataset.

In the initial phase, we decided to hand code the data without the use of AI, in order to ensure the accuracy and reliability of the data coding process. The analysis employed reflective thematic analysis, a methodical and iterative approach that identifies patterns and themes in qualitative data, especially effective for survey analysis. Following Braun and Clarke's (2006) six steps, reflective thematic analysis prioritizes discussion, validation, and intercoder agreement, ensuring the main tenets of rigorous qualitative analysis are heeded (Saldaña, 2013). Additionally, we used the ATLAS Looking at Data protocol, developed by the Center for Leadership & Educational Equity (n.d.). Designed for efficient data examination, this protocol is particularly suited for analyzing under time constraints.

In the subsequent phase, we used Bard and ChatGPT-3.5 only for theming, using several rounds of specific prompts (see Appendix) to evaluate AI's proficiency in interpreting the quantitative and qualitative elements of the dataset. A key aspect of the study's methodology was constant-comparative analysis. This iterative technique, which alternates between human-only and AI-assisted analysis, is instrumental in validating and triangulating findings throughout the research process (Glaser & Strauss, 1967). By meticulously comparing insights derived independently from human and AI analyses, we ensured a rigorous validation process. This approach produced reliable and valid results within a limited timeframe, underscoring the study's commitment to methodological rigor despite constraints. It offers versatility in analyzing quantitative and qualitative data across various settings, including educational and professional contexts.

DATASET

To evaluate the effectiveness of AI in qualitative data analysis and DEI-focused assessment questions, we extracted a subset of data from annual assessment reports, covering 93 graduate and undergraduate programs across four colleges during the 2022–2023 academic year. The reports, submitted by departmental faculty via Qualtrics (https://www.qualtrics.com), addressed two questions about the integration of DEI within program-level assessment efforts (see Figure 1). The data were scrubbed of any identifying details such as program name, faculty name, and program-specific information. This decision ensured privacy but might have removed context that would have been valuable for the analysis. The DEI-related dataset allowed for a focused analysis but limited the scope of the study.

The first question asked faculty respondents to indicate what DEI practices their program uses out of a list of 13 possible options, which included "other" and allowed respondents to fill in details as well as "none at this time." Respondents could select any number of options. The second question asked faculty respondents to provide any related

Which of the following practices related to anti-racism, decoloniality, diversity, equity, inclusion, and racial justice does the degree program include or utilize?

- Included in the learning outcomes.
- Included in the assessment plan.
- Specifically addressed in departmental policies
- Department faculty committee
- Curricular revisions
- Professional development
- Providing resources to faculty and staff within the department
- Collection of disaggregated demographic data
- Analysis of assessment results by disaggregated data
- Diverse student advisory groups
- Developed anti-bias leadership competencies.
- Other - please describe: ______________
- None at this time.

Use the space below to add any comments regarding practices related to anti-racism, decoloniality, diversity, equity, inclusion, and racial justice that the degree program includes or utilizes.

FIGURE 1 List of DEI practices and open-ended response question.

comments. Individuals from nearly all programs responded, with one faculty member representing each program. However, this person may not have known about all DEI-related practices used in their program. The resulting dataset included three elements: a multi-colored horizontal bar chart showing selected DEI practices categorized by college, a table presenting selected DEI practices categorized by college, and 30 open-ended responses. The data were extracted from Qualtrics, de-identified, and managed across Microsoft Teams, Word, and Excel for further analysis and review.

DATA ANALYSIS AND RESULTS

The data analysis process included three rounds: independent analysis, collaborative analysis, and AI integration. The initial two rounds focused on coding the data, and the third round involved multistep phases of independent analysis, collaborative discussion, and AI testing. Although the process is presented as linear, in practice it was iterative.

First round: Familiarizing and coding—Human only, independent

To initiate the data analysis process, each researcher worked alone to become familiar with the initial dataset. One researcher summarized the data, examined the qualitative comments, highlighted, bulleted answer choices, and clustered responses into 13 categories, while maintaining a researcher-reflective journal to track insights, hunches, and any other nuances during the testing process. The other researcher bulleted a themed list of answer choices by college; categorized the overall themes into faculty, student, program, and other; and maintained researcher notes for coding of themes. Our primary objective was to derive a comprehensive understanding of the data, particularly focusing on responses to the first question about DEI practices used in the respondent's program. During this independent analysis phase, each researcher employed a unique approach to examine the data, highlighting areas of note and capturing key observations. These individual efforts laid the foundation for the subsequent rounds of collaborative analysis and AI integration.

Second round: Refining the coding scheme—Human-to-human, collaborative

To refine the coding process and ensure consistency across the dataset (bar chart, table, and open-ended responses), we embarked on a collaborative analysis phase guided by a modified version of the ATLAS Looking at Data protocol (Center for Leadership & Educational Equity, n.d.). This structured approach facilitated a thorough comparison of our individual analyses, allowing identification of areas of agreement and possible discrepancies. Through examination of our assigned codes, we engaged in an intensive agreement process, meticulously documenting the rationale behind each coding decision in reflective journals and researcher notes. This rigorous approach ensured that the final coding scheme emerged from a consensus-driven process, fostering consistency and reliability across the dataset. The insights captured during this collaborative phase proved invaluable in shaping the subsequent round of analysis.

Third round: Theme extraction and AI-assisted theme generation

The third round of analysis focused on theme extraction and the evaluation of AI-assisted theme generation. The third round included three distinct phases: (a) independent theme identification, in which we independently identified themes within the dataset, meticulously examining each component to uncover overarching patterns and recurring concepts; (b) collaborative agreement on themes, in which we transitioned from independent to collective consensus, convening to reconcile our independently identified themes and refine and finalize the thematic framework, ensuring consistency and alignment in their interpretations; and (c) human–AI collaborative theme generation, in which we assessed the capabilities of AI-assisted theme generation tools, engaging Bard and ChatGPT with a series of prompts to extract themes from each component of the dataset.

Testing AI analysis of quantitative data (bar chart)

The initial prompt challenged the AI models to analyze the quantitative data in the bar chart. Bard initially struggled to interpret the chart but with additional description in the prompt was eventually able to recognize the colleges and practices. It generated a table that made no sense, identifying only 5 of the 13 possible practices and making errors in some of the frequencies. ChatGPT could not analyze the chart because the free version could only interpret the text as of the date of the study.

Testing AI analysis of quantitative data (table)

Following the pattern used for the bar chart, the next step involved testing the AI tools' capabilities in analyzing the quantitative data in the table. The data table, structured with columns representing the four colleges and their respective reported DEI-related practices, provided frequency and percentage data for each practice across all colleges. To effectively guide the AI tools' analysis, we crafted a series of prompts designed to elicit specific insights, such as the most prevalent DEI-related practices for each college and overall trends across all colleges. Bard and ChatGPT were then tasked with analyzing the table based on these prompts and revealing the quantitative patterns within the dataset.

We carefully evaluated the AI tools' responses against our own analysis, assessing for accuracy, completeness, and possible misinterpretations. Bard made some errors and misinterpretations, including, in the third prompt, erroneously identifying the least common DEI practice, which it had reported correctly in the second prompt. ChatGPT converted the table to a single stream of text but was able to interpret the data as a table and describe the data in a clear format with percentages and counts for each of the college responses. However, it provided no interpretation when initially prompted, other than identifying the highest percentage reported. This rigorous evaluation process led to the refinement of the prompts and iterative reanalysis, ensuring that the AI tools produced reliable and consistent insights aligned with our findings.

Testing AI analysis of open-ended responses

The final step in this round was evaluating the capabilities of Bard and ChatGPT in analyzing the open-ended responses, overall and by college (Table 1). We developed a series of prompts to guide the AI tools in extracting themes and insights from the qualitative data (see Appendix). The initial prompt challenged the AI models to analyze the data and identify major themes within the 30 open-ended responses, which comprised 1619 words. When asked to provide references, the AI generated references that were real but outdated for the field.

DISCUSSION AND FUTURE CONSIDERATIONS

Based on the results of our analysis, we discuss the use of manual coding versus AI-assisted coding, the usefulness of AI in theming, and the ability of AI to source assessment-related literature.

Manual versus AI-assisted coding

This study did not test the strength of AI in qualitative data coding. Instead, an intensive, manual, deep reading approach was applied to analyze the data. This approach, paired with extensive longitudinal knowledge of academic departmental progress in assessment, institutional emphasis on DEI initiatives, and broader campus initiatives, enabled a comprehensive analysis. Insights were revealed that a more AI-automated process might have overlooked. The decision against using AI for coding was driven by an intention to thoroughly engage with the data. As Saldaña (2021) stated, "Coding is not a precise science; it is primarily an interpretive act" (p. 7).

TABLE 1 Results of human-to-human synchronous collaborative with AI assistance on open-ended responses.

Prompts	Prompt action	Result: Bard	Result: ChatGPT
Initial prompt	Analyze the data and list major themes for the 30 open-ended responses	Initially, Bard could not accommodate the text, so we converted the Word file to JPG. Bard identified four major themes: commitment to antiracism, decoloniality, DEI, and racial justice; student success; professional development; and community engagement.	ChatGPT offered 13 themes with comments and a summary paragraph.
Second prompt	Analyze and theme the data by college	Bard was unable to process the data due to misreading combined text as blank pages. It provided steps in how to analyze data.	ChatGPT offered reasonable themes.
Third prompt	Analyze themes in relation to the literature on DEI work in higher education	Bard provided an analysis of the themes.	ChatGPT provided an analysis of the themes and cited references.
Fourth prompt	Back the analysis with references to current higher education assessment literature	Bard provided a list of 14 references ranging from 2002 to 2020.	ChatGPT stated that it does not have access to up-to-date references but suggested general areas in which to find literature related to higher education assessment and DEI, with examples for each category.
Fifth prompt	Provide top five recommendations that a university or college office of assessment can use to activate or strengthen DEI work in a degree program	Bard provided five recommendations.	ChatGPT stated it was providing five recommendations but provided 10.
Sixth prompt	Provide sources for the recommendations in the fifth prompt	Bard offered four sources along with its own knowledge of DEI.	ChatGPT said its recommendations were based on widely accepted principles and strategies for promoting DEI in higher education and included six sources with examples, acknowledging its last knowledge update occurred in September 2021.
Seventh prompt	Provide more specific recommendations for building a sustainable equity practice in degree programs	Bard responded with five additional recommendations.	ChatGPT provided 14 additional specific recommendations.

A handful of researchers have tested the use of AI for qualitative data coding (e.g., Gao et al., 2023; Jiang et al., 2021, Lubars & Tan, 2019; Zhang, Wu, Xie, Kim, et al., 2023; Zhao et al., 2021), underscoring its potential. However, its limitations must be acknowledged. AI lacks the nuanced understanding that researchers develop through long-term campus knowledge and experience in interpreting data reports and identifying patterns. Consequently, assessment practitioners should proceed with caution when employing AI as a coding tool and ensure they scrutinize and validate any AI output. Overreliance on AI's frequency counts, as noted by Feuston and Brubaker (2021), can lead to overlooking crucial "edge cases" and "one-off examples" (pp. 19–21) embedded within the data. To effectively harness AI's capabilities while maintaining control over the analysis, Lubar and Tan's (2019) delegation framework provides a valuable guideline. This underscores the importance of a balanced approach when considering the use of AI for coding.

AI's utility in theming

We evaluated AI tools for theming using the free versions of Bard and ChatGPT, restricting access to advanced features available in ChatGPT's premium version. This restriction could have impacted the tool's performance and constrained the depth of analyses possible. Limitations also emerged in how the tools handled various data types. For instance, the free version of ChatGPT processed text but not images. Bard had text input limitations due to token limits, and ChatGPT faced constraints related to its knowledge base, last updated, at the time of the study, in September 2021.

The effectiveness of AI hinges on the art of prompt engineering, also known as "prompt art," which involves carefully crafting prompts to guide AI models toward generating informative and relevant outputs (Wang & Jin, 2023). This process often relies on prompt chaining, where a series of prompts are sequentially fed to the AI model to refine and enhance its responses. The size of prompts is constrained by token limits, which dictate the maximum number of characters that can be incorporated into a chatbot request. As Wang and Jin (2023) stated, "A well-crafted prompt helps the model generate more accurate and relevant responses" (p. 5). Achieving rich and accurate AI responses often requires a layered and time-consuming process of data precleaning, output formatting, and careful iteration of prompts.

We observed AI's value in facilitating gap analysis (Feuston & Brubaker, 2021). Additionally, the AI tools demonstrated rapid data analysis, ranging from 3 to 15 s. However, a risk of oversimplification or misinterpretation exists, particularly with complex qualitative data, highlighting the necessity for careful interpretation of AI-generated insights. Bard misinterpreted bar chart elements, possibly due to issues related to formatting and color-coding. Furthermore, Bard displayed errors in analyzing table data, even after having previously provided correct interpretations. When prompted for dataset themes, ChatGPT provided more of a summary than a thematic analysis. However, the ability to copy output from both platforms was highly useful.

The process of data theming is intensive and time-consuming (Saldaña, 2021), making tools that expedite it potentially beneficial. However, researchers must address possible errors introduced through additional prompts, emphasizing the need to verify output accuracy. Understanding the text limitations that may lead to data reductionism is crucial and aligns with qualitative data analysis principles (Saldaña, 2021). Wang and Jin (2023) cautioned against the autoregressive nature of AI, highlighting the tendency to accumulate errors and imperfections from previous iterations, particularly in generating lengthy texts (p. 8). The substantial time investment required for prompt iteration is also a caution (Zhang, Wu, Xie, Kim, et al., 2023). To ensure credible AI-generated output, assessing the trustworthiness and rationale of prompt choices is vital. Zhang, Wu, Xie, Kim, et al. (2023) cautioned about the lack of transparency and traceability in AI's data output. Additionally, researchers must evaluate the cost implications of using paid versions of AI tools such as ChatGPT and other AI platforms, in consideration of their potential for more accurate and comprehensive analyses.

AI's ability to source assessment-related literature

We anticipated limitations in AI's ability to access relevant literature, a prediction based on our familiarity with AI platforms and separate literature reviews. ChatGPT primarily sourced older materials, whereas Bard exhibited better performance due to its Google search functionality. The need emerged to define DEI-related terms, such as DEI, social justice, racial justice, and equity. In addition, AI platforms lacked knowledge of campus initiatives, such as drivers for DEI work, racial justice task forces, national and state calls for DEI-focused assessment work, and institutional and specialized accreditation standards.

These limitations (see Figure 2 for an example) highlight the central role of human researchers in contextualizing literature vis-à-vis the data. The findings emphasize the need to provide specific contextual information regarding campus, regional, and national initiatives for AI to generate accurate recommendations. We initially doubted the tools' ability to cite relevant literature, but found that the citations were relevant, though outdated. We were also surprised to discover that the tools presented additional insights absent from our own analyses.

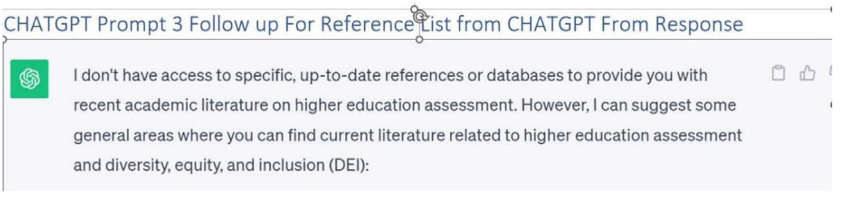

FIGURE 2 Follow-up prompt asking ChatGPT to clarify reference list for DEI sources.

IMPLICATIONS FOR POLICY AND PRACTICE

The results of this study have implications for policies and practices. The integration of AI into academic assessment practices marks a critical juncture for higher education institutions, necessitating a comprehensive approach that includes not only assessment office staff but also upper-level administration, such as provosts and presidents, in understanding and guiding its use. The policies and practices that follow are targeted to assessment professionals, yet they can also provide valuable insights for other campus offices and programs.

Policy implications for campus practice

Include assessment office staff in campus AI policy groups due to their campus-wide leadership and unique perspective across colleges, engagement with student performance–related data and faculty perception data, and management of sensitive campus information. Establish explicit campus guidelines and protocols to ensure transparency in AI-assisted analyses. These guidelines should foster accountability in results interpretation, maintain integrity in blinding processes through adherence to institutional review board protocols, and ensure ethical conduct.

Practice implications for assessment professionals

Address ethical considerations related to AI integration into assessment processes. Specifically, explore situations where human judgment and expertise hold paramount importance. Finding the right equilibrium between AI-generated insights and human oversight, as Lubars and Tan (2019) emphasized, is crucial. Develop effective training modules for assessment professionals to adeptly use AI in their practice. In addition, consider establishing frameworks to assess competencies in AI usage. Investigate the application of platforms such as Bard, ChatGPT, or other GenAI tools in the accreditation processes, annual assessment reporting, departmental assessment work, and general education assessment with an equity lens. Additionally, explore ways to leverage these tools to support innovations that foster inclusive learning environments and evidence-based decision-making.

AI IN THE WORK OF THE ASSESSMENT PRACTITIONER: A TOOL TO GET STARTED

AI presents an opportunity to expedite certain aspects of assessment work. The 10-step guide integrating AI into assessment (Table 2), developed as a result of this study, presents an intersection of roles articulated by Jankowski and Slotnick (2015), such as methodologist, facilitator/guide, narrator/translator, visionary/believer, and political navigator. Jankowski (as cited in Bheda et al., 2022) has further asserted the importance of the equity position that assessment practitioners should adopt, adding equity champion, ally, and activist to the original framework (Jankowski, 2022), which this study captures under the label of social justice activist. The guide is also informed by Lubars and Tan's (2019) delegation framework and Braun and Clarke's (2006) reflexive thematic analysis steps. Assessment professionals are well-positioned in their roles to lead the integration of AI and assessment practices, paving the way for transformative change (Bheda et al., 2022). AI has the potential to revolutionize assessment practices, offering a promising avenue for future research and practice.

STUDY LIMITATIONS

Although the study findings offer valuable insights for assessment practitioners contemplating AI integration in qualitative analysis, it is imperative to note the limitations. Though our decision to hand code and theme the data without the use of AI in the initial phase ensured the accuracy and reliability of the data coding process, it might have introduced human bias into the analysis. The small sample size constrains the generalizability of the two AI tools' effectiveness in diverse assessment-related research contexts. To enhance robustness, a larger and more varied sample is recommended. The study did not use statistical methods for reliability and validity but did employ methodological strategies to ensure trustworthiness (Noble & Smith, 2015). The specificity of prompts for AI-generated

TABLE 2 AI-assisted data analysis: 10-step guide for assessment practitioner roles.

Step	Action	Rationale	Assessment practitioner role
1	Review AI capabilities and platforms	Determine strengths and weaknesses of AI tools for project analysis.	Assessment/method expert Political navigator Visionary/believer
2	Prepare materials for AI analysis ensuring confidentiality	Collect and clean data. Protect privacy by removing personal identifiers. Ensure ethical handling of data within academic standards (institutional review board).	Assessment/method expert Political navigator
3	Initial AI testing	Assess AI's insight generation. Align AI capabilities with academic assessment needs.	Assessment/method expert Political navigator
4	Refine and test prompts for further AI analysis	Create prompts guiding AI analysis for academic program insights. Refine AI prompts for accuracy and relevance. Enhance datasets as needed.	Assessment/method expert Political navigator
5	Test AI analysis with different platforms	Consider output variation from different AI platforms. Choose platform(s) with the most reliability and accuracy.	Assessment/method expert
6	Troubleshoot AI implementation	Address technical challenges and adjust accordingly for more reliable results.	Assessment/method expert
7	Evaluate AI-assisted analysis	Ensure that results are accurate and relevant. Identify researcher and AI biases.	Assessment/method expert Narrator/translator Political navigator
8	Review results and prepare for dissemination	Develop materials outlining the strengths and weaknesses of AI-assisted analysis.	Narrator/translator Political navigator Visionary/believer
9	Advanced knowledge and practical applications	Share results with stakeholders. Leverage AI to enhance academic success and improve programs through data-driven insights. Consider additional field testing.	Facilitator/guide Method expert Narrator/translator Visionary/believer
10	Consider the impact of AI-assisted analysis	Examine and reflect on equity and social justice implications of AI insights. Use AI with fairness and inclusivity-mindedness.	Social justice activist Political navigator

analysis could also affect the quality of results, and the study's brief duration may limit the depth of analysis. Lastly, the exclusive focus on Bard and ChatGPT may overlook the value of other AI tools for qualitative analysis. The decision to use AI for theming but not for coding in the subsequent phase is another limitation that was discussed in the discussion section of the paper. Exploring a broader range of AI tools would offer a more comprehensive evaluation for specific use cases, especially with the aforementioned rapid advances in AI.

CONCLUSION

This study examined the use of AI tools in qualitative and quantitative analysis for a component of campus-based assessment. The results demonstrate that assessment practitioners can strategically employ AI to glean more comprehensive, actionable insights about student learning. However, successful integration necessitates careful control so that AI enhances rather than replaces human expertise. When judiciously implemented, AI can increase efficiency, uphold ethical standards, and promote inclusivity. While embracing technology's potential, assessment practitioners must also critique and hold it accountable, as Penn (2022) called for. Thoughtful scholars should lead AI integration, advancing assessment scholarship on human–AI collaboration.

With careful facilitation, AI can become a valuable tool for practitioners to promote more effective, equitable evaluation practices in service of student learning and success. However, it remains precisely that—a tool to augment human skills, not a replacement for studied analysis. Keeping the human firmly in control while leveraging AI's strengths is essential for harvesting the benefits on terms that prioritize human judgement and ethics. This study contributes to the growing field of AI in education, offering valuable insights and directions for future research and practice. As AI continues to evolve and permeate various aspects of higher education, the exploration of its potential, while remaining mindful of its limitations and ethical implications, becomes increasingly important.

REFERENCES

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of Artificial Intelligence*, *7*(1), 52–62. https://doi.org/10.2139/ssrn.4337484

Bheda, D., Jankowski, N. A., & Felten, P. (2022). An invitation to a beginning rather than the end. In G. W. Henning, G. R. Baker, N. A. Jankowski, A. E. Lundquist, & E. Montenegro (Eds.), *Reframing assessment to center equity* (pp. 326–340). Stylus Publishing.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(1), 77–101. https://doi.org/10.1191/1478088706qp063oa

Center for Leadership & Educational Equity. (n.d.). *ATLAS looking at data*. https://www.schoolreforminitiative.org/download/atlas-looking-at-data/

Feuston, J. L., & Brubaker, J. R. (2021). Putting tools in their place: The role of time and perspective in human–AI collaboration for qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, *5*($CSCW_2$), Article 469. https://doi.org/10.1145/3479856

Gao, J., Choo, K. T. W., Cao, J., Lee, R. K. W., & Perrault, S. (2023). CoAIcoder: Examining the effectiveness of AI-assisted human-to-human collaboration in qualitative analysis. *ACM Transactions on Computer-Human Interaction*, *31*(1), 1–38. https://doi.org/10.1145/3617362

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory. Strategies for qualitative research*. Aldine.

Henning, G. (2023, November 12–14). *Integrating artificial intelligence into assessment practice* [Conference session]. 2023 NASPA Region I Conference, Portland, ME, United States.

Janio, J. (2023, October 8). *Assessment and AI: A shifting terrain* [Video]. YouTube. https://www.youtube.com/watch?%3DRrlwHLPh1PM&t%3D6s

Jankowski, N. A. (2022, July 17–20). *Equity and excellence: Supporting students in teaching and learning* [Presentation]. Opening plenary for SACSCOC Summer Institute, Lake Buena Vista, FL, United States.

Jankowski, N. A., & Henning, G. W. (2024, February 4). *AI and assessment in student affairs*. Student Affairs Assessment Leaders. https://studentaffairsassessment.org/entries/announcements/ai-and-assessment-in-student-affairs

Jankowski, N. A., & Slotnick, R. C. (2015). The five essential roles of assessment practitioners. *Journal of Assessment and Institutional Effectiveness*, *5*(1), 78–100. https://doi.org/10.5325/jasseinsteffe.5.1.0078

Jiang, J. A., Wade, K., Fiesler, C., & Brubaker, J. R. (2021). Supporting serendipity: Opportunities and challenges for human–AI collaboration in qualitative analysis. *Proceedings of the ACM on Human–Computer Interaction*, *5*($CSCW_1$), Article 94. https://doi.org/10.1145/3449168

Leffer, L. (2023, October 26). Humans absorb bias from AI—and keep it after they stop using the algorithm. Scientific American. https://www.scientificamerican.com/article/humans-absorb-bias-from-ai-and-keep-it-after-they-stop-using-the-algorithm/

Levin, J. (2023, November 1). *In the age of numbers, don't lose sight of the qualitative data*. University Business. https://universitybusiness.com/in-the-age-of-numbers-dont-lose-sight-of-qualitative-data/

Lubars, B., & Tan, C. (2019). Ask not what AI can do, but what AI should do: Towards a framework of task delegability. *arXiv*. https://arxiv.org/abs/1902.03245

Manyika, J., Silberg, J., & Presten, B. (2019, October 25). What do we do about the biases in AI? *Harvard Business Review*, *0*, 00–00. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Montenegro, E., & Henning, G. W. (2022). Why the intersection of assessment and equity? In G. W. Henning, G. R. Baker, N. A. Jankowski, A. E. Lundquist, & E. Montenegro (Eds.), *Reframing assessment to center equity: Theories, models, and practice* (pp. 18–35). Stylus Publishing.
Nicholas, M. C., & Slotnick, R. C. (2018, April). *A portrait of the assessment professional in the United States: Results from a national survey*. National Institute for Learning Outcomes Assessment. http://learningoutcomesassessment.org/occasionalpaperthirtyfour.html
Noble, H., & Smith, J. (2015). Issues of validity and reliability in qualitative research. *Evidence-Based Nursing*, *18*(2), 34–35. https://doi.org/10.1136/eb-2015-102054
Penn, J. D. (2022). Advancing assessment of learning in higher education as a discipline: Benefits, tensions, and next steps. *Research & Practice in Assessment*, *17*(2), 5–18. https://www.rpajournal.com/advancing-assessment-of-learning-in-higher-education-as-a-discipline-benefits-tensions-and-next-steps/
Saldaña, J. M. (2013). *The coding manual for qualitative researchers* (2nd ed.). Sage Publications.
Saldaña, J. M. (2021). *The coding manual for qualitative researchers* (4th ed.). Sage Publications.
Shea, P., & Walton, D. (2023, April 28). *AI in education: Implications for assessment and assignment design* [Conference presentation]. 11th Annual AMCOA Conference, Westborough, MA, United States. https://www.youtube.com/watch?v=d7Xo36E2DSY
Singer-Freeman, K. (2023, November 5). *Grand challenges in assessment in the age of AI* [Webinar]. The California Outcomes & Assessment Coordinator Hub. https://coaches.institute/2023/11/05/friday-slo-talk-coaches/
Suskie, L. (2009). *Assessing student learning: A common sense guide*. John Wiley & Sons.
Walton, D., Slotnick, R., & Shea, P. (2024, April 26). Take a walk on the wild side: Empowering assessment practices with generative AI [Conference presentation]. Massasschusetts Department of Higher Education. https://www.mass.edu/strategic/documents/AMCOA%20Conference/2024/Generative_AI_Assessment_Workshop-AMCOA2024-Slotnick-Walton-Shea.pdf
Wang, S., & Jin, P. (2023). *A brief summary of prompting in using GPT models*. Qeios. https://doi.org/10.32388/imzi2q
Weave Education. (2023, November 2). *Webinar: Humanizing assessment reporting with Dr. Peggy Maki | Weave professional development* [Video]. YouTube. https://www.youtube.com/watch?v=AMZgi4xFRUA
Zhang, H., Wu, C., Xie, J., Kim, C., & Carroll, J. M. (2023). QualiGPT: GPT as an easy-to-use tool for qualitative coding. *arXiv*. https://arxiv.org/abs/2310.07061
Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2023). Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. *arXiv*. https://arxiv.org/abs/2309.10771
Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. *arXiv.org*. https://arxiv.org/abs/2102.09690

How to cite this article: Slotnick, R. C., & Boeing, J. Z. (2024). Enhancing qualitative research in higher education assessment through generative AI integration: A path toward meaningful insights and a cautionary tale. *New Directions for Teaching and Learning*, 1–17. https://doi.org/10.1002/tl.20631

APPENDIX

Analysis prompts tested in AI

This index provides a structured guide to the prompts used in our analysis. It covers key question formulation, detailed data analysis (both quantitative and qualitative), and follow-up inquiries related to diversity, equity, and inclusion (DEI) in higher education. Although not exhaustive, this summary offers an overview of the main prompts employed to explore DEI practices within academic programs.

- Bar chart and table analysis prompts:
  - Prompt for bar chart: “Analyze this image and describe the data without interpreting it for the following chart.”
  - Follow-up prompt: “Analyze this image, describe the data, and interpret it for the following chart.”

  - Follow-up prompt: "Analyze this image and describe the data without interpreting the data for the following chart showing counts of degree programs by college indicating practices related to DEI."
  - Follow-up prompt: "Which of the following practices related to anti-racism, decoloniality, diversity, equity, inclusion, and racial justice does the degree program include or utilize?"
  - Follow-up prompt: "Analyze this image and describe the data without interpreting the data for the following chart showing counts of degree programs by college indicating practices related to DEI. The *n*'s for each college represent the total number of programs reporting in each college."
  - Comprehensive analysis prompt: "Analyze this image, describe the data, interpret it, and analyze the data for themes based on the question prompt: Which practices related to anti-racism, decoloniality, diversity, equity, inclusion, and racial justice does the degree program include or utilize? Note: Percentages in the table are out of the total respondents for each college and the total as indicated in the column header."
- Open-ended response analysis prompts:
  - Initial analysis prompt: "Analyze the following open-ended response data and list major themes."
  - Detailed analysis prompt: "Analyze the following open-ended response data and list major themes by college."
  - Literature connection prompt: "Follow-up prompt: Analyze these themes in relation to literature on DEI work in higher education. Do you have any current higher education assessment literature references to back your analysis?"
- Recommendations to strengthen DEI institutionally and with programs prompts:
  - Recommendations request: "Provide five top recommendations that a university or college office of assessment can use to activate or strengthen DEI work in a degree program."
  - Source inquiry: "Follow-up prompt: What are your sources for these recommendations?"
  - Specificity request: "Follow-up prompt: Be more specific about recommendations for building a sustainable equity practice in an academic degree program."
  - Descriptive analysis: "Describe what you see and put into context."

AUTHOR BIOGRAPHIES

Ruth C. Slotnick is Director of Assessment at Bridgewater State University.

Joanna Z Boeing is Assistant Director of Assessment at Bridgewater State University.