A peer reviewed, open-access electronic journal: ISSN 1531-7714

STEM item generation: Can ChatGPT be culturally responsive?

Laura Lambert, *James Madison University*
Mason Jones, *James Madison University*

Abstract: This exploratory study investigates bias in multiple-choice biology items generated by ChatGPT-4o, focusing on how a user's query history influences item content. Specifically, it addresses three research questions: (1) How well does ChatGPT generate introductory biology items? (2) How does it interpret a request for culturally relevant content? and (3) How do outputs vary across three distinct user profiles? Using a standardized series of prompts, 100 items were generated per condition. All items were analyzed for factual accuracy, content representation, and patterns in correct answer distribution. Qualitative analyses evaluated the depth of culturally responsive framing across the users. While ChatGPT produced largely accurate items across conditions, there were biases that emerged. Culturally responsive prompts often yielded tokenized cultural statements rather than contextually rich items. Correct answers were non-randomly distributed, posing threats to test validity. Crucially, user query history influenced representation of content topics and what is considered "culture" within the generated items. These findings have implications for test developers at any level considering genAI tools that preserve a user's query history in assessment design, emphasizing the need for careful consideration of both prompt engineering as well as user history.

Keywords: AI, Culturally-responsive, Test development

Introduction

The assessment landscape is undergoing a profound shift, driven both by evolving social expectations around equity and inclusion and by rapid advancements in generative artificial intelligence (genAI). While test development has traditionally been time-consuming, labor-intensive, and costly, the need for assessments that are both psychometrically sound and culturally responsive has increasingly been the center of discussion. Historically, standardized test items have reflected the cultural assumptions of the test developers: a traditionally white and male-dominated field. As the demographics of classrooms shift, the

need to create instruments that are valid across cultural contexts is not only ethical but psychometric. The field is moving toward frameworks that embed students' identities and lived experiences into the assessment process, as emphasized by scholars like Walker et al. (2023), who advocate for design principles that emphasize shared power, cultural relevance, and asset-based perspectives.

Simultaneously, genAI technologies—particularly large language models such as ChatGPT—are being examined for their potential to streamline the test development process. These tools are also increasingly used by users with varying levels of experience with genAI, often in informal or exploratory ways (e.g., "write 10 test questions about…"), and sometimes without awareness that outputs may be shaped by default settings or by the user's prior query history. While genAI tools promise to reduce development costs and increase speed, their outputs are only as culturally attuned as the data they were trained on. Indeed, while genAI has shown early promise in generating technically sound multiple-choice items (Kıyak et al., 2024), its capacity to produce culturally responsive content remains underexplored. Furthermore, issues of algorithmic bias and factual hallucinations persist, raising concerns about the reliability and representativeness of genAI-generated items (Bender et al., 2021; Cao et al., 2023).

This paper explores whether genAI can be part of a more culturally responsive approach to assessment or whether its use risks reproducing the very inequities that culturally responsive frameworks seek to dismantle. Unlike studies that examine carefully engineered prompts under controlled conditions, our focus is on how item generation unfolds when users interact with genAI in their own personalized chat environments. Because large language models adapt their responses based on prior query history, two users entering the same prompt may not receive the same items, raising new questions about equity, reproducibility, and control in item development.

Importantly, this paper does not approach this question in the context of a trained psychometrician performing deliberate prompt engineering to optimize item quality. Rather, it evaluates outputs produced under routine usage conditions—where users generate item pools in personalized chat environments without structured prompt optimization—to explore how prior query history and personalization effects may shape the results. In short, we examine how genAI interprets baseline, non-engineered prompts and how this behavior influences the quality and content of test items it produces. Because concerns about cultural relevance, equity, and historical patterns of exclusion are deeply rooted in the test development literature, it is essential to situate genAI within these longstanding debates. The following review summarizes the foundations of culturally responsive assessment and the challenges it seeks to address before turning to the emerging role of genAI.

Literature Review

Many everyday users turn to genAI precisely because they lack the time, resources, or psychometric expertise required for high-quality test development. In reality, creating a well-designed assessment (let alone a culturally responsive one) demands a level of labor, training, and financial investment far beyond what most instructors, staff, or small programs can sustain. However, even when these resources are available, traditional test development practices have long struggled to produce items that are equitable across diverse cultural groups. The test development process is not cheap in terms of time, effort, or absolute cost. For a "simple" multiple choice test, developers must consult with subject matter experts, create questions that follow a series of best practices, and ensure the instrument adequately covers the construct being assessed (e.g., early English literature, introductory chemistry, or the antebellum period; Cobb, 1998; Haladyna et al., 2002).

Despite this investment, items designed to be "neutral" or "objective" often reflect the cultural assumptions of those who develop them: typically individuals from middle- to upper-class white, Western

backgrounds (Randall, 2021). As a result, students from historically marginalized communities may not see their cultures, histories, or lived experiences represented in the assessments they encounter. This highlights an interesting contrast of traditional test development: assessments are expensive and time-intensive to create, yet the resulting instruments still frequently fall short of producing fair, valid scores for all groups. These exact challenges and discussions have fueled a growing movement and conversation toward culturally responsive assessment.

Culturally Responsive Assessment

Culturally responsive assessments aim to allow the increasingly diverse students of today to bring and see their cultural backgrounds and identities in the assessments they take. Randall (2021) pointed out that our students are not neutral. This begs the question: if our students are not neutral, why do our assessments try to be?

Cultural responsiveness is not a new idea, having taken several names over the years. Ladson-Billings (1995) referred to "culturally relevant pedagogy" with "culturally responsive assessment" arriving in the vernacular at nearly the same time (Hood, 1998). Evans (2021) illustrated the relationship between "culturally sensitive", "culturally relevant", "culturally responsive", and "culturally sustaining." She asserted that each builds on the other: an assessment cannot be culturally responsive without first being culturally relevant and culturally sensitive. In this paper, we consider the key difference between those two terms as the degree of integration of culture into the curriculum. Thus, while a set of items in isolation may be evaluated for cultural relevancy, once they enter into a broader curriculum they may be part of a culturally responsive classroom.

However, the question remains about how to translate the concept of cultural responsiveness into item and test development (Hood & Hopson, 2008; Montenegro & Jankowski, 2020; Stake, 1975). Acknowledging that "assessment is not an apolitical process" (Montenegro & Jankowski, 2020, p. 7) is crucial to using new tools and technologies to develop assessment instruments. These instruments have too often put minoritized examinees at a disadvantage due solely to their cultural background (Hood, 1998). As Sireci (2020) further pointed out, the rigidity imposed by standardized assessment leads to those not represented in the dominant culture being excluded. Using an example of students who do not speak English as their first language, Sireci raised the question of if a test is designed for native English speakers, how can we assume it will work the same way for students who are not native English speakers? Taking this a step further, what information is then lost about these students?

Highlighting the reason *why* culturally relevant and responsive assessments are so important, we turn to the seminal work by Geneva Gay (2002, 2010). Gay (2002) stated that "when academic knowledge and skills are situated within the lived experiences and frames of reference of students, they are more personally meaningful, have higher interest appeal, and are learned more easily and thoroughly." (p. 106) In other words, when students see themselves in the materials and assessments, the content is more easily learned. However, the majority of assessments students encounter, in particular standardized tests, are constructed under a white racial frame which centers the beliefs and cultural references of the dominant white culture (Russell, 2024).

Randall et al. (2021) described a culturally responsive assessment as one that allows students to draw on their cultural knowledge, connects with their lives, is embedded in a culturally sustaining curriculum, and enables them to demonstrate their skills, knowledge, and understanding in a variety of ways. Walker et al. (2023) went on to lay out five design principles for culturally responsive assessment (abridged below):

Culturally responsive assessments…

1. … require a process that shares power across all concerned parties at all stages of the assessment process.

2. …should be designed to foster academic engagement and belonging in academic environments.
3. …should reflect the expectation that all students have the potential to perform at high levels.
4. …should be designed to maximize flexibility to account for individual differences in culture, interests, and identities of all learners.
5. …are designed to reflect asset-based perspectives that measure what students know and can do and disrupt traditional deficit-narratives. (p. 4)

As is evident, creating a culturally responsive assessment is not an act done in isolation by a single person (or genAI engine). Creating the test items is just one aspect of building a culturally responsive assessment. Shared power needs to be considered from the beginning: How are all stakeholders being given a voice in the process? How is flexibility being incorporated? Is this assessment being embedded in a curriculum that celebrates cultural diversity?

Tension surrounding the incorporation of diverse identities in education is not a new phenomenon. In the post-Reconstruction and Jim Crow era, Black students were forbidden from attending white institutions in the South, and segregation was enforced by law in public schools and universities (Civil Rights History Project, 2011; Equal Justice Initiative, 2014). Segregation denied Black students access to equitable curricula and education altogether. Even after desegregation was mandated by *Brown v. Board of Education* (1954), resistance and efforts to resegregate schools at the state and local level have continued up to the present day (Orfield & Lee, 2007). Beyond racial segregation, controlling the content of instruction also has a history of control and censorship. For example, in the early- and mid-1900s, history textbooks omitted or whitewashed the experiences of Black and Indigenous people, a debate that continues today (Gershon, 2015; Rosso, 2022; Zimmerman, 2004). Another example can be found in the Kanawha County textbook controversy in 1974, when intense backlash to multicultural and identity-sensitive textbooks led to efforts to reassert control over what was taught in schools (Cowan, 1974).

Unfortunately, in today's current political climate, incorporating diverse cultural identities into education and assessment continues to be divisive and politically charged. Universities have come under intense scrutiny for having DEI (diversity, equity, and inclusion) offices or practices (Office for Civil Rights, 2025), AP classes on African American history are not allowed to be taught (Kim, 2023) or count towards graduation (Cline, 2025), and books containing LGBTQ+ and black stories are being banned in libraries across the United States (American Library Association, 2025). This divisiveness can affect both changes in law and the way items are developed and assessed. With culturally responsive assessment already facing political resistance, the rapid adoption of generative AI presents both an opportunity and a risk for how assessments are designed, interpreted, and implemented.

Generative Artificial Intelligence

GenAI has quickly come to the forefront of discussions about how assessments are designed and scored, raising questions about whether these tools can alleviate existing challenges or introduce new forms of inequity. These conversations range from students using genAI to do their work for them, faculty using genAI to aid in grading, and psychometricians evaluating if genAI can help make their job easier, cheaper, faster, or more representative (Attali et al., 2022; Kıyak et al., 2024; Micheletti et al., 2024). A large question of interest in the assessment field is how genAI can be leveraged to create and/or grade large-scale assessments. One use of AI in educational assessments has been in automated essay scoring - using machine learning and natural language processing to score essays written by examinees (Attali et al., 2008; Hussein et al., 2019; Ramesh & Sanampudi, 2022). This has shown promise in reducing the time and human-hours needed to score these assessments. Another area of interest is the potential for generative AI to create multiple-choice items. That is, can it help reduce the cost per item?

Generative AI's Role in Assessment

ChatGPT has been shown to be able to successfully generate multiple choice items (Kiyak et. al., 2024), though in this instance it was only evaluated for correctness and acceptable psychometric properties. In another study, Attali and colleagues (2022) used a transformer-based language model to create both reading passages and items. To review the artifacts for correctness and bias, the authors employed a human-in-the-loop method, where humans reviewed the items and passages generated by the genAI model. Keeping humans closely involved in item and/or passage generation is the primary way to counteract some of the concerns that still exist around using genAI. Even given the need for continued human involvement, using genAI to produce test materials could save time and money in the test development process.

However, since model responses are wholly dependent on the material they were trained on, the risk then arises that stereotypes and social biases can be reinforced or amplified (Bender et al., 2021). Currently, there is limited work showing the ability of ChatGPT or other genAI engines to create test items that could be considered culturally responsive. However, in other work, it has been shown that ChatGPT is limited in adapting to diverse cultural concepts, particularly those outside of American cultural norms (Cao et al., 2023). While there is some promise around carefully engineered prompts being able to generate more culturally diverse content, there still remain challenges around subcultural nuances (Nyaaba et al., 2024).

As mentioned, while genAI can save time and money, there is still the need to keep humans closely involved to counteract some of the persistent drawbacks and concerns around using genAI for test development. The two main areas of concern are bias and hallucination. Bias in this instance is similar to yet different from test bias. Specifically, test bias refers to the case when a test does not have the same predictive utility for all examinee groups (AERA et al., 2014). Algorithmic bias, on the other hand, occurs when genAI models preferentially generate content based on imbalanced training data, producing output that is skewed towards particular topics or cultural perspectives. Hallucination is when the output is factually inaccurate – and often very confidently so! A third area to be aware of with ChatGPT in particular is how a user's query history influences future output and responses. Personalization can lead to variability between users, even when the same prompt is used. This will be relevant to any engine that requires a log-in and saves queries in a memory.

Bias. Content generated by genAI can vary in its accuracy and bias. As with instruments designed exclusively by humans, if impact and equity are not considered from the beginning, they will always be an afterthought, and dominant assumptions, knowledge, and culture will be at the forefront (Randall et al., 2023). Generative AI has been shown to perpetuate both gender and racial biases, as well as allowing women to be increasingly targeted by defamatory deepfakes (Kuck, 2023). Taking this into a literature perspective, Lucy and Bamman (2021) showed that stories generated by GPT-3, a prior model of ChatGPT, included more male characters than female characters. Additionally, GPT-3 would follow social stereotypes based on the character's gender. This has the potential to transfer to short reading passages for testing – if ChatGPT is used to rapidly generate a wide range of passages, the risk is that stereotypes around gender and/or race will be present as well. Further, items written by genAI may unintentionally perpetuate existing biases if not caught by humans.

Careful prompt engineering can be used to try and counteract bias, but it is not always a guarantee (Lucy & Bamman, 2021). Researchers using DALL-E, an image-generating AI model, found that simple prompts produced images containing harmful stereotypes on a variety of demographic characteristics. Further, they found that even when trying to lessen these stereotypic images through careful prompt construction, the stereotypes persisted in the images (Bianchi et al., 2023). What is becoming increasingly evident is that prompt terms matter with genAI and bias can be minimized or exacerbated with a biased prompt (Snyder,

2023). Considering this from an instrument development perspective, if items are to be generated using genAI, humans still need to be involved to ameliorate bias in the items.

Hallucination. Another noted concern with ChatGPT, and genAI in general, is that these engines can produce hallucinations, or output that is factually incorrect, illogical, or entirely fabricated. This can be attributed to the fact that genAI is a computer algorithm that is generating text using a probabilistic language model. GenAI is not evaluating factual accuracy, but rather the likelihood of one word following another (Nananukul & Kejriwal, 2024). Additionally, the performance of any language model will depend on the training data used (Mohammed et al., 2025; Sakib, 2024). If not enough data (i.e., not enough variety of textual examples) was used, or if that data was biased, the model may be more likely to fabricate responses.

As with most technology, as improvements are made, ChatGPT and other genAI models become more accurate and reliable. Moving from ChatGPT-3.5 to ChatGPT-4 has decreased the frequency of hallucinations (i.e. factual or contextual), though not eliminating them (Mohammed et al., 2025). Generated text should still be examined to ensure scientific facts, historical evidence, or other "factual" events presented are, in fact, accurate. Other areas where ChatGPT has been shown to struggle are controversial topics and topics lacking a clear scientific consensus (McIntosh et al., 2024). While much of the content for this study (introductory biology) is not controversial, some topics have been contentious and are becoming contentious again (e.g., evolution, gender, sex markers).

Query History. A user's query history with a particular model can impact future responses – an aspect of genAI that can be both useful and a hidden source of bias. The history of past queries can be used to build a user history, or profile, which will allow future responses to be tailored to the interests and preferences of the user (Ge et al., 2018). This history can also allow for a better contextual "understanding" of a current query for the genAI engine, particularly when there is significant back-and-forth between the user and genAI (Fu et al., 2020). From a more commercial standpoint, a genAI model's ability to personalize responses based on a user's query history can lead to greater user engagement and satisfaction (Wang et al., 2024). While it may be convenient for ChatGPT and other genAI models to remember a user's preferences or frequently used coding software, it also raises concerns such as: "What *isn't* the genAI model telling the user?" or "Is this just what the genAI model has predicted the user wants to hear?" Important for an assessment context, the impact of a user's query history may result in different responses to similar queries between two individuals. If, for example, one user has a heavy history of STEM topics while another user has a very scant STEM topic history, their responses to requests for STEM items may result in items of different rigor or depth. Another example may be in how personal beliefs come through in query history. If a user's political or religious beliefs heavily color their query history, future output will take that into account.

Purpose of Current Study

The purpose of this exploratory study is to examine how generative AI functions when used to create test items under the kinds of real-world conditions faced by individuals who may lack specialized psychometric training, dedicated staff, or substantial budgets for assessment development. As described in the literature, culturally responsive assessment requires intentional, resource-intensive design, while generative AI promises a faster and more accessible alternative. Yet concerns about accuracy, cultural relevance, bias, and personalization raise important questions about how genAI performs when used by everyday practitioners rather than expert developers.

The primary motivating factor for this work was envisioning how someone tasked with generating test items may use genAI in practice, particularly those working without extensive institutional support or technical expertise. Thus, the study focuses on two key aims: (1) evaluating whether ChatGPT can generate culturally relevant test items, and (2) determining how differences in user query history influence the accuracy, content, and framing of those items. Findings from this work may inform guidance for educators

and assessment practitioners seeking to integrate commercially available genAI tools into item development processes. Specifically, this study addresses the following research questions:

1. How accurately does ChatGPT generate multiple-choice items?
2. How does ChatGPT interpret an explicit request for culturally relevant items?
3. How do generated items vary across three different ChatGPT query histories?

Method

This research served as an exploratory foray into the capabilities of genAI in creating culturally responsive STEM multiple-choice items as well as evaluating the impact of various user histories. To reduce the number of variables, a single genAI engine was used, the free version of ChatGPT-4o (OpenAI, 2024). Given that this was the free version, there was a limited number of prompts that each account could ask in a given time frame using the 4o engine. A series of prompts was used to request a total of 100 multiple-choice items suitable for an undergraduate introductory biology course cumulative final exam. A specific course (i.e., BIO101) was not requested, nor were specific curricular topics.

Biology was chosen as the content area because the first author teaches courses on biology and biotechnology at the undergraduate level and possesses an advanced degree in the field. Recognizing that individual instructors may approach their class differently, most introductory biology courses will follow the same general topics. The open source *Biology 2e* (Clark et al., 2018) was used as a curriculum guide for this paper to allow for a consistent determination of fit. This book contains topics such as the chemical foundation of life, biological macromolecules, cells, genetics and evolution, plant and animal structure and function, and ecology. The breadth of content in this book is often spread across two semesters of instruction. However, the concept of "introductory biology" would cover content from the entire book (i.e., both semesters of instruction). Therefore, the curriculum of introductory biology was treated as the full content of *Biology 2e* (Clark et al., 2018).

Prompts

Two separate prompts were used for this study (see Appendix 1). The initial prompt asked for a total of 100 multiple-choice items suitable for a comprehensive final exam in an introductory undergraduate biology course. The second prompt followed the same guidelines as the first and further specified that the 100 items needed to be culturally responsive. Specific prompt guidelines included stating that the questions should be multiple-choice with four answer options, targeting more than just memorization (i.e., beyond simple term identification). The prompt also stated that the correct answer needed to be indicated to allow the author to verify accuracy. Finally, ChatGPT was asked to provide a brief rationale behind each question.

A single-prompt approach (e.g., Goloujeh et al., 2024) was taken, given the focus of this study was not prompt engineering but rather the baseline impact of user history on generated items.

Query History

These prompts were used with three different "users": the author's personal ChatGPT account ("author user"), a fictitious user ("trained user"), and a second fictitious, totally naïve, user with no query history ("naïve user"). These three users enabled the comparison of generated items across different query histories, similar to what might be found in a cooperative setting. Given the potential complexity of this research design, Table 1 shows the research design in abbreviated format, highlighting where each research question is being addressed and what methods are being used.

The trained user was generated by the first author using a new email address and login credentials to ChatGPT. To create a user profile with a clearly polarized and ideologically opposite that of the first author, the trained user account engaged with news headlines, opinion pieces, and language patterns drawn from sources classified as far-right by independent media bias rating systems. These sources were identified using

Table 1. Experimental Design

	Prompt 1 (100 multiple-choice items)	Prompt 2 (100 culturally responsive multiple-choice items)
Author User	**RQ1**: How accurately does ChatGPT generate multiple-choice biology items? **Analysis**: Content coverage (qualitative); distribution of answer options (chi-square)	**RQ2**: How does ChatGPT interpret an explicit request for culturally relevant items? **Analysis**: Content coverage (qualitative); comparison of content coverage to RQ1 (chi-square); cultural responsiveness (qualitative + quantitative)
Trained User	**RQ3**: How do generated items vary across three different ChatGPT query histories? **Analysis**: Content coverage (qualitative + quantitative); comparison of content coverage across users (chi-square); distribution of answer options (chi-square)	**RQ3**: How do generated items vary across three different ChatGPT query histories? **Analysis**: Content coverage (qualitative + quantitative); comparison of content coverage across users (chi-square); cultural responsiveness (qualitative + quantitative)
Naïve User	**RQ3**: How do generated items vary across three different ChatGPT query histories? **Analysis**: Content coverage (qualitative + quantitative); comparison of content coverage across users (chi-square); distribution of answer options (chi-square)	**RQ3**: How do generated items vary across three different ChatGPT query histories? **Analysis**: Content coverage (qualitative + quantitative); comparison of content coverage across users (chi-square); cultural responsiveness (qualitative + quantitative)

publicly available tools such as the Media Bias Chart (Media, 2024) and AllSides (*Media Bias*, 2021). This orientation was selected because, in the current U.S. political climate, far-right media frequently features strongly worded commentary on diversity, equity, inclusion (DEI), and culturally responsive education, which are topics central to the present study. Using phrasing common to these outlets allowed us to simulate a real-world scenario in which a user brings a long-term history of interacting with highly opinionated content relevant to assessment and culture-based debates.

The purpose of this approach was not to evaluate or endorse any political positions, but rather to model a coherent, strongly skewed interaction history: a factor that may influence how large language models adapt to user behavior over time. Any prompts that triggered content warnings were immediately discontinued in accordance with platform guidelines and ethical research standards.

The author's account history contains topics around R and Python code debugging, questions on creating data visualizations, navigating GitHub, reducing word count in provided works (e.g., an abstract needing to be 200 words that is currently 213), meal prep ideas, formatting citations in APA format, and

requests for editing assistance of a provided work (e.g., acting as a critical reviewer, point out areas of this manuscript that needs additional clarification). The author has never engaged her ChatGPT account in discussions of news topics from any news outlet. However, some of the works provided to ChatGPT, either for a reduction in word count or requests for critical feedback, likely reflected her views that equity should be central to any discussion on assessment, diversity should be pursued, and systemic barriers still exist for many people.

Data Analysis

Data analysis specific to each research question is detailed below. As an initial step, all items were screened for factual accuracy by the first author, focusing specifically on whether the indicated answer was correct and whether the question stem was adequate. Factually accurate items were then further examined depending on the generating prompt (more details below). During factual screening, correct answer patterns were observed to be heavily skewed toward only one or two answer options. After this observation, the distribution of answer patterns was also tabulated. Following screening by the first author, the items were independently screened by a second faculty member in her department. This faculty member was provided the items, answers, and rationales generated by ChatGPT and asked to indicate if the item was factually correct and appropriate for an introductory biology course. Following this independent rating, the author and faculty member met to discuss item accuracy. Across all sets of items, agreement of factual accuracy was at 100%. This is not to say every item was rated as accurate. Rather, the raters agreed on the designation of accuracy for all items.

RQ1: How accurately does ChatGPT generate multiple-choice biology items? This research question was initially examined in the context of the Author user only; examination of the other two users with respect to this research question occurred in RQ3. Factually accurate items were further analyzed based on coverage of the curriculum as laid out in *Biology 2e* (Clark et al., 2018). The units of study, each containing multiple chapters, were examined for content. For example, the unit on "The Chemistry of Life" covered topics such as the scientific method, basic chemistry, water, and biological macromolecules. Items addressing these topics were mapped to this unit of study. This process was repeated for the remaining units of study until all items were matched with the most appropriate unit. These counts were used to generate proportions of items addressing each unit.

As with determining factual accuracy, a second faculty member independently mapped the items to the units of study in *Biology 2e* (Clark et al., 2018). The author and this faculty member then met to compare mappings. Any items that were mapped differently by the author and the second faculty member were discussed until agreement reached. This process resulted in 100% agreement between the faculty member and the author for all items.

After the observation of the skewed answer choice distribution described previously, these counts were tabulated. A chi-square test was performed to determine if the proportions differed significantly from an expected 0.25 proportion for each answer choice. All statistical analyses were performed in R version 4.5.1 (R Core Team, 2025).

RQ2: How does ChatGPT interpret an explicit request for culturally relevant items? As in RQ1, this research question was initially examined in the context of the Author user only; the other users will be examined in RQ3. For the prompt requesting culturally responsive items, the way in which ChatGPT interpreted cultural relevance was examined. Given the qualitative nature of these data, a basic qualitative research approach was taken, where the end product is a summary rather than the development of a theory (Merriam & Tisdell, 2016). Content analysis was also used to analyze these data. This is designed to describe patterns and frequencies within gathered data (Merriam & Tisdell, 2016). The decision was made to evaluate the items as a set (i.e., as a whole test) rather than individually due to the nature of the codes. Any one item

would be flagged as lacking most codes, while perhaps the test would not. For example, if an item contains a single scientist in the item stem, that individual would not be described as identifying as all gender orientations simultaneously.

The set of items were analyzed via a deductive coding approach. Pre-determined codes were created using a subset of topics contained in the "Culturally Responsive Higher Education Curriculum Assessment Tool" (McNulty et al., 2024). The presence of these codes across the items as a set was quantified. This quantification was further enriched by a qualitative description of the presence of the codes in the item set. The codes can be seen in Appendix 2. Each item set was coded independently by both authors, each of whom took notes during the process. The raters then met to discuss the presence or absence of codes in each item set until agreement was reached. Observed patterns within the items were also noted and discussed.

In addition to evaluating cultural relevance, these items were also examined for content as previously described. A chi-square test with simulated p-value was performed to determine if the content covered differed between the two prompts. It was necessary to use a simulated p-value due to multiple cell counts with values less than five (Hope, 1968). A Fisher's exact test was not possible due to the size of the table.

RQ3: How do generated items vary across three different ChatGPT query histories? Items generated by each of the fictitious users (i.e., the trained user and the naïve user) were examined as described in RQ1 and RQ2. In addition to tabulating content coverage and evaluating items for cultural relevancy, these values were compared across all three users. A chi-square analysis was used to determine if content coverage differed by user, with a separate analysis performed for each prompt. Recognizing that multiple tests were being run, a Bonferroni correction was employed to avoid an inflated Type I error rate.

Results

Overall, the results indicated that for an undergraduate introductory biology course, ChatGPT was able to create factually correct items. However, it struggled more in its interpretation of "culturally relevant," creating items that were more tokenizing of cultural context than truly culturally responsive. Finally, this study underscores the role a user's query history can play in the responses generated. The specific results related to each research question are detailed below.

RQ1: How accurately does ChatGPT generate multiple-choice biology items?

Items generated from prompt 1 (100 multiple choice questions for an undergraduate introductory biology course) by the author's account were generally factually correct with sound rationale – only two items were questionable. One of these questions was factually incorrect (attributing activation energy to an enzyme rather than the reaction). The other question was not factually incorrect as written but was not a well-written question (what individual would be most likely to express an X-linked recessive trait).

The items covered a range of topics that would be expected for such a course (e.g., reactions, biological macromolecules, genetics and natural selection, cell division, photosynthesis, ecosystems, etc.). Table 2 contains the broad unit categories from Biology 2e (Clark et al., 2018) and the proportion of items in each category for each user. The other two users will be examined in the section for research question 3. One area that was not represented, which the author expected to have at least one question on, was basic chemistry (i.e., chemical bonds and properties of water). This is a small portion of an introductory biology course, but it contains important concepts that carry through much of the rest of the course.

One interesting pattern that arose from an overall test evaluation was around the frequencies of correct answers. Over 50% of the questions had "C" as the correct answer, and another 30% had "B" (Table 3).

This was a significant deviation from an expected distribution of 25% for each answer choice, $\chi^2(3, N = 100) = 78.16$, $p < .0001$.

Table 2. Proportion of Items Categorized to Each Unit of Study in Biology 2e.

	Author User		Trained User		Naïve User	
	Prompt 1	Prompt 2	Prompt 1	Prompt 2	Prompt 1	Prompt 2
The Chemistry of Life	0.09	0.08	0.06	0.10	0.13	0.01
The Cell	0.29	0.08	0.39	0.07	0.55	0.04
Genetics	0.24	0.09	0.34	0.05	0.18	0.12
Evolutionary Processes	0.08	0.04	0	0.04	0.03	0.04
Biological Diversity	0.02	0.08	0.03	0.09	0.03	0.12
Plant Structure and Function	0.04	0.16	0	0.04	0.01	0.04
Animal Structure and Function	0.11	0.11	0.17	0.55	0.03	0.19
Ecology	0.13	0.36	0.01	0.06	0.04	0.44

Table 3. Percent Representation of Each Answer Choice by User and Prompt

	Author User		Trained User		Naïve User	
	Prompt 1	Prompt 2	Prompt 1	Prompt 2	Prompt 1	Prompt 2
A	9	29	1	4	6	33
B	31	56	30	35	32	50
C	57	12	58	59	59	14
D	3	3	11	2	3	3
	$\chi^2(3) = 78.16$	$\chi^2(3) = 64.56$	$\chi^2(3) = 75.44$	$\chi^2(3) = 89.04$	$\chi^2(3) = 82.00$	$\chi^2(3) = 51.76$

Note. All χ^2 values were significant at $p < .0001$.

RQ2: How does ChatGPT interpret an explicit request for culturally relevant items?

The questions generated by the prompt asking explicitly for culturally relevant items by the author's account were again factually correct. Indeed, while three questions had questionable answers, none were factually incorrect. As an example, one question asked "In rural Ghana, people use neem as a natural pesticide. Neem works by:", with the correct answer being "Disrupting insect hormone systems". This question is not factually incorrect, but it is more detailed than expected for an undergraduate introductory biology course. There is a chapter on the endocrine system (i.e., hormones), but the focus is on types of hormones, human body processes regulated by hormones, and hormone production. It could be argued that an instructor could cover this specific instance, but it is not part of a general curriculum. Similar to the questions generated by the initial prompt, the distribution of answer choices was skewed, with 56% of items having "B" as the correct answer and 29% having "A" (Table 3).

ChatGPT altered the breadth and scope of the questions compared to the initial prompt. Across all questions, ChatGPT would use the first sentence of every question stem to situate the question in a particular context. The second sentence/phrase would then pull out a question from the context. For example:

An Eritrean physician researches sickle cell disease. What evolutionary advantage does the sickle cell trait offer?

A Cambodian biologist preserves wild rice strains. Why is genetic diversity important?

Additionally, the overall themes represented in these questions were predominantly ecology/farming, plant biology, and food. In the initial prompt, these topics made up 26% of the questions, predominantly from ecology and population genetics. However, in the prompt asking for culturally responsive items, these topics now make up 70% of the questions. One aspect that the culturally relevant items address, which the items from the initial prompt did not, is ethics. The culturally relevant set had three items addressing ethical considerations or environmental justice topics. None of the items from the initial prompt addressed these topics. Alternatively, topics on genetics, reactions, and enzymes were very lightly covered, with no examples considering alleles and/or Punnett squares, even though the initial prompt included five items on these topics. These topics are commonly covered in an introductory biology course, including in the *Biology 2e* textbook used as the curriculum guide for this paper (e.g., Chapters 2, 6, and 12; Clark et al., 2018).

After independent coding and prior to reconciliation, the two raters had a 96% agreement rate. Reconciliation resulted in 100% agreement. Qualitative evaluation of the items as a set showed that some codes were more represented than others. For example, the items had racial representation of individuals, LGBTQ+ individuals, and non-Christian religions. However, there was no representation of persons with disabilities, persons impacted by the legal system, or diverse relationships and family structures. Some items were culturally-affirming while others could be considered to be stereotyping. There was discussion between the raters on this code in particular, deciding if the presence of stereotyping items would negate the presence of culturally-affirming items (i.e., would the harm of the stereotype outweigh the benefit of cultural affirmation). This set of items did connect to a variety of social and environmental issues that could affect students on an individual or societal level as well as presented examples of service, volunteerism, and activism. As a single test, these items did not ensure accessibility for students, a definite area for improvement.

Looking at the topics, cultures, and contexts provided in the question stem, there were instances of stereotypes potentially being reinforced (e.g., "A Chinese herbalist studies ginseng's effects on metabolism. What organelle is directly involved in energy production?" or "A Pakistani village experiences frequent cousin marriages. What genetic concept becomes more relevant?"). In many cases, the "cultural" aspect appeared to be an afterthought added to the question stem. For example: "A Palestinian engineer designs green roofs to reduce city heat. What plant structure reduces water loss?" The cultural context is tenuously related to the question portion.

RQ3: How do generated items vary across three different ChatGPT query histories?

The three different "users" of ChatGPT all generated both sets of questions. The two sets of questions generated by both prompts for the author were discussed in the context of research questions 1 and 2 above. This was to address a key aspect of RQ3, and the overall purpose of this paper. If a test developer is turning to genAI to generate test items, they will likely turn to their personal account without much consideration. This research question addresses how different user query histories may change the distributions seen in RQ1 and RQ2. Both of those research questions will now be discussed in the context of the other users below.

Prompt 1: 100 biology items. For Prompt 1, the most notable difference was in the trained user. There were four items on sex, sex chromosomes, or gender for the trained user. For all these items, the stem

and/or the answer choices were maintaining a very strict gender binary (e.g., "Which statement about biological sex is correct? Answer: It is genetically fixed at conception."). None of the items from the author's account had this strict binary; indeed, there was only one question on sex chromosomes with the focus being X-linked recessive disorders. The author's items included more questions involving alleles and/or Punnett squares (see above), whereas neither the trained user nor the naïve user generated items on these topics.

Another notable difference is the absence of items addressing natural selection, evolution, or population genetics in the trained user. Both the author's items and the naïve user's items had multiple questions on those topics (Table 2). The trained user was also missing items on plant structure and function that the author user and naïve user were not. A chi-square test with simulated *p*-value evaluating if there was a difference in content representation between the three users was significant, $\chi^2 = 135.86$, $p < .0001$.

There were no marked differences in the number of factually incorrect items (one or two per user). As previously noted, each user had a disproportionate number of correct responses in a single category (Table 3). This falls under a test design problem that could be corrected with human oversight or perhaps better prompt engineering.

Prompt 2: 100 culturally relevant biology items. The items generated from the second prompt follow a similar pattern to the first prompt: the author's items and the naïve user's items were similar, while the trained user's items follow a different pattern. As discussed above, the culturally relevant items for the naïve user followed the same pattern of introducing a cultural context as a one-sentence introduction to the question stem.

The trained user, however, did not incorporate cultural diversity in the question stem. As an example, "A Christian farmer rotates crops to protect soil health. Which benefit does this practice offer?" or "A homesteading family stores meat in a freezer. Why does freezing preserve food?" There is still a one-sentence contextualization of the question, but the diversity seen in the other users is not as apparent. With the trained user, there is a much more overt focus on religion and patriotism than was seen in the prior two queries. Specifically, 25% of the items mentioned "Christian", "God", "missionary", or "church" while another 8% mentioned "patriot" or "soldier". In items where an individual was doing work or going to school, that individual was always described as male.

While these items were qualitatively evaluated using the same codes as the other two users, the conclusion was that these items are not culturally responsive. The only code that could be said to be represented was that this set of items did not present non-dominate cultures as alien or exotic. This was because there were no non-dominate cultures represented in the items.

A chi-square test with simulated *p*-values was run to determine if the content representation differed between the three users for this prompt. A significant difference was again found, $\chi^2 = 139.76$, $p < .0001$.

Implications for Practice

Reflecting back to the context of this study, we intentionally adopted the perspective of typical users integrating genAI into assessment workflows rather than that of a psychometrician deliberately optimizing prompts. Such users may be classroom instructors, program-level assessment coordinators, or even test developers or item writers. The implications of this study are of particular importance in this context.

The findings of this study highlight several important considerations for practitioners who may rely on generative AI to support item development, particularly those working without extensive psychometric training or institutional resources. While genAI can rapidly create large numbers of multiple-choice items that are often factually accurate, human oversight remains critical for ensuring quality, fairness, and

representational accuracy. GenAI item generation may provide a starting point, but careful review and iterative refinement are necessary to ensure that items are accurate, high quality, and culturally sensitive.

One implication concerns foundational aspects of test design. The uneven distribution of correct answer choices across many genAI-generated items illustrates how baseline output may inadvertently introduce patterns that test-wise students can take advantage of. From a validity perspective, such structural irregularities raise concerns about evidence based on internal structure and response processes, as described in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). Even when items are factually correct, further refinement is necessary to ensure they align with intended constructs and do not introduce construct-irrelevant variance. Therefore, genAI outputs would be best considered as drafts rather than fully developed items.

Another implication relates to cultural representation. Across the user profiles examined, item content varied in both topic selection and cultural contextualization based on the model's interpretation of each user's prior interactions. Although culturally relevant prompts elicited some attempts at contextualization, these additions were often brief, inconsistently integrated, or not meaningfully connected to the underlying content. Therefore, reliance on genAI alone is unlikely to produce culturally relevant items without substantial human involvement. However, it is important to keep in mind that culturally responsive assessment frameworks, such as those put forward by Walker et al. (2023) and Randall et al. (2021), emphasize the importance of flexibility in demonstrating knowledge, alignment with students' lived experiences, and asset-based representation. Thus, a single multiple-choice format, particularly when generated without intentional cultural design, cannot fully meet these expectations. While culturally relevant prompts may encourage some contextualization, the results of this study suggest that authentic alignment with these frameworks requires collaboration among instructors, students, and relevant stakeholders. Generative AI cannot substitute for these essential practices.

Within the unified validity framework put forth by Messick (1989), these concerns extend beyond surface representation to the consequential basis of validity. Shifts in contextual framing has the potential to influence whose knowledge is centered and how constructs are interpreted. If item contextualization varies across diverse item-generation environments, the meaning and social implications of the resulting scores may also shift. Cultural responsiveness, then, is not outside the discussion of validity. Rather, it is embedded within the interpretive arguments that support score use.

A further implication concerns the influence of user query history on item generation. Differences between user profiles were most apparent when culturally relevant items were requested, indicating that perceived values or viewpoints subtly shaped topic emphasis and contextual framing. This pattern may reflect a combination of interaction history effects and model sycophancy, whereby genAI systems tend to align with user beliefs to maintain conversational flow (Sharma et al., 2024). Practitioners using personal accounts to draft test items may therefore introduce unintended forms of bias linked to their prior interactions with the model. Given these findings, a practical recommendation for educators and assessment developers is to use a dedicated or new account when generating initial item drafts with commercially available genAI systems. Doing so reduces the influence of prior interactions and supports more consistent, neutral output. This practice may be especially valuable for individuals or programs without access to specialized assessment personnel, as it provides a straightforward way to limit additional sources of bias beyond the model's existing training data.

Extending the impact of query history on output to Kane's (2006, 2013) argument-based validity perspective, the "personalization" of output to a user introduces a potential vulnerability. Considering that within this framework, observed scores are thought to generalize to a defined universe of items, different query histories that lead to different item pools is problematic. However, when item generation depends on user query history, the sampling of that universe may shift across different test developers. The question

then becomes whether the item pool reflects the intended construct or if it is reflects a construct that was filtered through the lens of user query history. Therefore, if item content and/or features depend on user query history, the interpretive argument for score meaning must account for this additional source of variability.

To summarize, while genAI may reduce the time needed to produce an initial item pool, its integration into item development introduces validity concerns extending beyond efficiency. The effect of user query history, structural irregularities, and variability in cultural context highlights the need to examine not only item quality but also the stability of construct representation and the consequences of score interpretation. Within contemporary validity frameworks, these findings further emphasize that genAI has the potential to reshape the evidence that interpretive arguments are built on. Therefore, human oversight is not only necessary for accuracy, representational appropriateness, and alignment with cultural responsiveness frameworks but also to ensure defensible validity arguments can be made. Effective prompt engineering may improve the quality of initial output, but a human-in-the-loop process remains necessary to ensure that items are both psychometrically sound and culturally responsive. Ultimately, genAI may offer efficiency benefits, but it does not eliminate the expertise and care required in assessment design.

Limitations

From the start, this was an exploratory study. The purpose of this paper was to take a first look at this avenue of research and important questions. We encourage researchers to build on what we have here. Second, our intended audience was a casual genAI user who may not be trained on prompt engineering. Therefore, this study was looking at the behavior of ChatGPT "out of the box". Findings in this study may be able to be overcome with careful prompt engineering, or avoided entirely by using a custom-built engine. Methodologically, one major limitation to this study is that only one "user" of each type was queried. The distributions may look different if averaged over a number of unique users of each type. Additionally, all these questions were for an introductory undergraduate biology course. The ability of ChatGPT to generate accurate questions for an upper-level molecular biology course, for example, may not be as strong. Other topic areas and types of questions remain to be examined.

Received: 5/30/2025. **Accepted:** 2/28/2026. **Published:** 3/2/2026.

Citation: Lambert, L. & Jones, M. R. (2026). STEM item generation: Can ChatGPT be culturally responsive? *Practical Assessment, Research, & Evaluation*, *30*(2)(8). Available online: https://doi.org/10.7275/pare.3152

Corresponding Author: Laura Lambert, James Madison University. Email: laycocla@jmu.edu

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.*

American Library Association. (2025, April 7). American Library Association kicks off National Library Week with the Top 10 Most Challenged Books of 2024 and the State of America's Libraries Report. *American Library Association*. https://www.ala.org/news/2025/04/american-library-association-kicks-national-library-week-top-10-most-challenged-books

Attali, Y., Powers, D., Freedman, M., Harrison, M., & Obetz, S. (2008). Automated Scoring of Short-Answer Open-Ended GRE Subject Test Items. *ETS Research Report Series*, *2008*(1), i–22. https://doi.org/10.1002/j.2333-8504.2008.tb02106.x

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.903077

Bender, E., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜*. *FAccT '21*, 610–623. https://doi.org/10.1145/3442188.3445922

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan. (2023). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, 1493–1504.

Brown v. Board of Education, 347 U.S. 483 (1954).

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). *Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study*. 53–67. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85185223314&partnerID=40&md5=98b9e611f86dbd7fffd2f4bf59599582

Civil Rights History Project. (2011). *School Segregation and Integration* [Web page]. Library of Congress. https://www.loc.gov/collections/civil-rights-history-project/articles-and-essays/school-segregation-and-integration/

Clark, M. A., Douglas, M., Choi, J., Clark, M. A., Douglas, M., & Choi, J. (2018). *Biology 2e | OpenStax* (Vol. 2nd). OpenStax. https://openstax.org/books/biology-2e/pages/1-introduction

Cline, N. (2025, May 27). Va. Governor vetoed bill to make Black history classes mandatory towards graduation. What's next? *Virginia Mercury*. https://virginiamercury.com/2025/05/27/virginia-governor-vetoed-bill-to-make-black-history-classes-mandatory-towards-graduation-whats-next/

Cobb, G. (1998, April 13). *The objective-format question in statistics: Dead horse, old bath water, or overlooked baby?* American Educational Research Association, San Diego, CA.

Cowan, P. (1974). Holy War in West Virginia: A Fight Over America's Future. *The Village Voice*. https://www.villagevoice.com/holy-war-in-west-virginia-a-fight-over-americas-future/

Equal Justice Initiative. (2014, March 1). *Resistance to School Desegregation*. https://eji.org/news/history-racial-injustice-resistance-to-school-desegregation/

Evans, C. (2021, November 3). Creating Culturally Responsive Assessment. *Center for Assessment*. https://www.nciea.org/blog/culturally-sensitive-relevant-responsive-and-sustaining-assessment/

Fu, Z., Cui, S., Ji, F., Zhang, J., Chen, H., Zhao, D., & Yan, R. (2020). *Query-to-Session Matching: Do NOT Forget History and Future during Response Selection for Multi-Turn Dialogue Systems*. 365–374. Scopus. https://doi.org/10.1145/3340531.3411938

Gay, G. (2002). Preparing for Culturally Responsive Teaching. *Journal of Teacher Education*, *53*(2), 106–116. https://doi.org/10.1177/0022487102053002003

Gay, G. (2010). *Culturally Responsive Teaching: Theory, research, and practice* (2nd ed.). Teachers College Press.

Ge, S., Dou, Z., Jiang, Z., Nie, J.-Y., & Wen, J.-R. (2018). *Personalizing search results using hierarchical RNN with query-aware attention*. 347–356. Scopus. https://doi.org/10.1145/3269206.3271728

Gershon, L. (2015, October 20). The Racism of History Textbooks. *JSTOR Daily*. https://daily.jstor.org/racism-history-textbooks/

Goloujeh, M. A., Sullivan, A., & Magerko, B. (2024). Is It AI or Is It Me? Understanding Users' Prompt Journey with Text-to-Image Generative AI Tools. *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3613904.3642861

Haladyna, T., Downing, S., & Rodriguez, M. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333. http://dx.doi.org/10.1207/S15324818AME1503_5

Hood, S. (1998). Culturally Responsive Performance-Based Assessment: Conceptual and Psychometric Considerations. *The Journal of Negro Education*, *67*(3), 187–196. JSTOR. https://doi.org/10.2307/2668188

Hood, S., & Hopson, R. K. (2008). Evaluation Roots Reconsidered: Asa Hilliard, a Fallen Hero in the "Nobody Knows My Name" Project, and African Educational Excellence. *Review of Educational Research*, *78*(3), 410–426. https://doi.org/10.3102/0034654308321211

Hope, A. C. A. (1968). A simplified monte carlo significance test procedure. *Journal of the Royal Statistical Society. Series B (Methodological)*, *30*(3), 582–598.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208. https://doi.org/10.7717/peerj-cs.208

Kane, M. T. (2006). Validation. In *Educational measurement* (4th ed.). American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kim, J. (2023, January 22). Florida says AP class teaches critical race theory. Here's what's really in the course. *NPR*. https://www.npr.org/2023/01/22/1150259944/florida-rejects-ap-class-african-american-studies

Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*, *80*(5), 729–735. https://doi.org/10.1007/s00228-024-03649-x

Kuck, K. (2023). *Generative Artificial Intelligence: A Double-Edged Sword*. Scopus. 2023 IEEE IFEES World Engineering Education Forum and Global Engineering Deans Council: Convergence for a Better World: A Call to Action, WEEF-GEDC 2023 - Proceedings. https://doi.org/10.1109/WEEF-GEDC59520.2023.10343638

Ladson-Billings, G. (1995). Toward a Theory of Culturally Relevant Pedagogy. *American Educational Research Journal*, *32*(3), 465–491. https://doi.org/10.3102/00028312032003465

Lucy, L., & Bamman, D. (2021). Gender and Representation Bias in {GPT}-3 Generated Stories. *Proceedings of the Third Workshop on Narrative Understanding*, 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5

McIntosh, T. R., Liu, T., Susnjak, T., Watters, P., Ng, A., & Halgamuge, M. N. (2024). A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination. *IEEE Transactions on Artificial Intelligence*, *5*(6), 2739–2751. Scopus. https://doi.org/10.1109/TAI.2023.3332837

McNulty, L., Peoples, L. Q., & Rantz, R. (2024). *Culturally Responsive Higher Education Curriculum Assessment Tool.*

Media, A. F. (2024). *Interactive Media Bias Chart*. Ad Fontes Media. https://app.adfontesmedia.com/chart/interactive?utm_source=adfontesmedia&utm_medium=website

Media Bias. (2021, July 12). AllSides. https://www.allsides.com/media-bias

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation* (Fourth edition.). Jossey-Bass, a Wiley Brand.

Messick, S. J. (1989). Meaning and Values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5–11.

Micheletti, N., Marchesi, R., Kuo, N. I.-H., Barbieri, S., Jurman, G., & Osmani, V. (2024). Generative AI Mitigates Representation Bias and Improves Model Fairness Through Synthetic Health Data. *medRxiv*, 2023.09.26.23296163. https://doi.org/10.1101/2023.09.26.23296163

Mohammed, M. N., Al Dallal, A., Emad, M., Emran, A. Q., & Al Qaidoom, M. (2025). A Comparative Analysis of Artificial Hallucinations in GPT-3.5 and GPT-4: Insights into AI Progress and Challenges. In *Studies in Systems, Decision and Control* (Vol. 566, pp. 197–203). Scopus. https://doi.org/10.1007/978-3-031-71318-7_18

Montenegro, E., & Jankowski, N. (2020). A new decade for assessment: Embedding equity into assessment praxis (Occasional Paper No. 42). *Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).*

Nananukul, N., & Kejriwal, M. (2024). *HALO: An Ontology for Representing and Categorizing Hallucinations in Large Language Models. 13058*. Scopus. https://doi.org/10.1117/12.3014048

Nyaaba, M., Zhai, X., & Faison, M. Z. (2024). Generative AI for Culturally Responsive Science Assessment: A Conceptual Framework. *Education Sciences*, *14*(12). Scopus. https://doi.org/10.3390/educsci14121325

Office for Civil Rights. (2025, February 14). *Dear Colleague Letter Regarding Students for Fair Admissions v. Harvard*. U.S. Department of Education. https://www.ed.gov/media/document/dear-colleague-letter-sffa-v-harvard-109506.pdf

Orfield, G., & Lee, C. (2007). *Historic Reversals, Accelerating Resegregation, and the Need for New Integration Strategies*. Civil Rights Project / Proyecto Derechos Civiles. https://eric.ed.gov/?id=ED500611

R Core Team. (2025). *R: A Language and Environment for Statistical Computing* (Version 4.5.1) [Computer software]. R Foundation for Statistical Computing. www.R-project.org

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Randall, J. (2021). "Color-Neutral" is not a thing: Redefining construct definition and representation through ajustice-oriented critical antiracist lens. *Educational Measurement Issues and Practice*, *40*(4), 82–90.

Randall, J., Poe, M., & Slomp, D. (2021). Ain't Oughta Be in the Dictionary: Getting to Justive by Dismantling Anti-Black Literacy Assessment Practices. *Journal of Adolescent & Adult Literacy*, *64*(5), 594–599. https://doi.org/10.1002/jaal.1142

Randall, J., Poe, M., Slomp, D., & Oliveri, M. E. (2023). Our validity looks like justice. Does yours? *Language Testing*, 02655322231202947. https://doi.org/10.1177/02655322231202947

Rosso, J. D. (2022, July 15). Censorship By Omission: How Systemic Racism is Downplayed and Dismissed in the Classroom. *Literary Hub*. https://lithub.com/censorship-by-omission-how-systemic-racism-is-downplayed-and-dismissed-in-the-classroom/

Russell, M. (2024). *Systemic racism and Educational Measurement: Confronting Injustice in Testing, Assessment, and Beyond*. Taylor & Francis Group.

Sakib, S. M. N. (2024). Bane and boon of hallucinations in the context of generative AI. In *Cases on AI Ethics in Business* (pp. 276–299). Scopus. https://doi.org/10.4018/9798369326435.ch016

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S. M., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards Understanding Sycophancy in Language Models. *The Twelfth International Conference on Learning Reresentations*. International Conference on Learning Representations, Vienna, Austria. https://openreview.net/forum?id=tvhaxkMKAn

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in Educational Assessment. *Educational Measurement: Issues and Practice*, *39*(3), 100–105. https://doi.org/10.1111/emip.12377

Snyder, K. (2023). Mindful AI: Crafting prompts to mitigate the bias in generative AI. *Textio*. https://textio.com/blog/mindful-ai-crafting-prompts-to-mitigate-the-bias-in-generative-ai

Stake, R. (1975). *Program evaluation particularly responsive evaluation*. New Trends in Evaluation, Goteborg, Sweeden.

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). Culturally Responsive Assessment: Provisional Principles. *ETS Research Report Series*, *2023*(1), 1–24. https://doi.org/10.1002/ets2.12374

Wang, R., Fan, K., Liu, Z., & Wang, J. (2024). A Study on User Querying Behavior in Generative Artificial Intelligence Environment. *Data Analysis and Knowledge Discovery*, *8*(8–9), 20–30. Scopus. https://doi.org/10.11925/infotech.2096-3467.2023.1145

Zimmerman, J. (2004). Brown-ing the American Textbook: History, Psychology, and the Origins of Modern Multiculturalism. *History of Education Quarterly*, *44*(1), 46–69. http://www.jstor.org/stable/3218110

Appendix 1

Prompt 1:

I am an instructor for an undergraduate introductory biology course. Please write 25* multiple-choice items suitable for a cumulative final exam for this course. These items should go beyond basic memorization of facts and include higher orders of thinking. For each item, indicate the correct answer and provide a brief rationale.
*A request for 100 was too large for the free version. Follow-up prompts were:
Please provide 25 additional unique items, following the same guidelines as before. (x3)

Prompt 2:

I am an instructor for an undergraduate introductory biology course. Please write 25* culturally responsive multiple-choice items suitable for a cumulative final exam for this course. These items should go beyond basic memorization of facts and include higher orders of thinking. For each item, indicate the correct answer and provide a brief rationale.
*A request for 100 was too large for the free version. Follow-up prompts were:
Please provide 25 additional unique items, following the same guidelines as before. (x3)

Appendix 2

Qualitative codes

These codes are a modification of the rubrics presented in the *Culturally Responsive Higher Education Curriculum Assessment Tool* (McNulty et al., 2024).

- Racial representation of individuals: Latinx/Hispanic, Black/African, Native American, Asian/Pacific Islander, Middle Eastern, White/Caucasian, Multi-Racial
- Representation of LGBTQ+ individuals
- Representation of persons with disabilities
- Representation of non-Christian religions
- Portrayal of diverse cultures, ethnicities, histories, and nationalities without stereotypes, generalizations, and assumptions
- Culturally-affirming references to different ethnic and cultural traditions, languages, beliefs, names, and dress
- Examines diverse relationships and family structures
- Highlights individuals with disabilities, honors their achievements, and values their contributions
- Does NOT present minoritized populations as having low economic wealth or low educational attainment
- Does NOT present non-dominate cultures as alien or exotic
- Highlights non-dominate populations, their strengths, and assets
- Acknowledges obstacles generated by systemic oppression and discrimination
- Acknowledges that individuals impacted by the legal system can turn their lives around
- Focuses on the dignity and contributions of diverse races, classes, genders, abilities, and sexual orientations
- Recognizes the value and integrity of diverse faiths and other belief systems
- Connects to social, political, or environmental issues that affect students on an individual or societal level
- Ensures accessibility for students by addressing multiple learning styles, types of disabilities, and intelligences
- Forges connections to the broader community by presenting examples of service, volunteerism, and activism
- Places value on a pluralistic, diverse, multicultural, and equitable society