

Identifying Items on Which Humans and Chatbots Diverge Using Differential Item Functioning

Licol Zeinfeld*
Ron Blonder

Alona Strugatski*
Shelley Rap

Ziva Bar-Dov
Giora Alexandron

Weizmann Institute of Science

*Equal contribution

{licol.zeinfeld, alona.faktor, ziva.bar-dov, ron.blonder, shelley.rap, giora.alexandron}@weizmann.ac.il

ABSTRACT

The rapid adoption of large language models (LLMs) in education raises profound challenges for assessment design. To adapt assessments to the presence of LLM-based tools, it is crucial to characterize the strengths and weaknesses of LLMs in a generalizable, valid and reliable manner. However, current LLM evaluations often rely on descriptive statistics derived from benchmarks, and little research applies theory-grounded measurement methods to characterize LLM capabilities relative to human learners in ways that directly support assessment design. Here, by combining educational data mining and psychometric theory, we introduce a statistically principled approach for identifying items on which humans and LLMs show systematic response differences, pinpointing where assessments may be most vulnerable to AI misuse, and which task dimensions make problems particularly easy or difficult for generative AI. The method is based on Differential Item Functioning (DIF) analysis – traditionally used to detect bias across demographic groups – together with negative control analysis and item-total correlation discrimination analysis. It is evaluated on responses from human learners and six leading chatbots (ChatGPT-4o & 5.2, Gemini 1.5 & 3 Pro, Claude 3.5 & 4.5 Sonnet) to two instruments: a high school chemistry diagnostic test and a university entrance exam. Subject-matter experts then analyzed DIF-flagged items to characterize task dimensions associated with chatbot over- or under-performance. Results show that DIF-informed analytics may provide a theory-grounded framework for understanding where LLM and human capabilities diverge, and highlight their value for improving the design of valid, reliable, and fair assessment in the AI era.

Keywords

Assessment, Generative AI, Differential Item Functioning

Licol Zeinfeld, Alona Strugatski, Ziva Bar-Dov, Ron Blonder, Shelley Rap, and Giora Alexandron. Identifying Items on Which Humans and Chatbots Diverge Using Differential Item Functioning. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyejeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 479–486. International Educational Data Mining Society (2026).

© 2026 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

<https://doi.org/10.5281/zenodo.21039815>

1. INTRODUCTION

The rapid adoption of generative AI (GenAI¹) tools in education has created both opportunities and risks. While these systems, particularly chatbots such as ChatGPT, can provide personalized explanations, feedback, and support for learners, their growing use also poses a profound threat to the validity of educational assessments [1, 4]. By making it exceptionally easy for students to generate fluent and convincing answers, GenAI lowers the barriers to academic dishonesty in ways that traditional safeguards struggle to address [1, 4, 10]. This risk is amplified in unproctored or remote environments, where monitoring is limited and misuse is harder to detect [27].

Importantly, the concern is no longer theoretical: recent studies document that students are already using GenAI for assignments and exams, directly undermining the validity and reliability of assessment results and raising urgent challenges for educational practice and policy [18, 23]. These changes point to a key observation: to ensure resistance to GenAI-related misconduct and to maintain validity and reliability in assessment contexts characterized by potentially heavy LLM use, it is necessary to understand where LLMs diverge from human learners and to characterize their capabilities and limitations across assessment tasks.

Current evaluations of LLM performance on assessment tasks are largely shaped by technical benchmarks (e.g., [3, 25, 14]). These benchmarks provide descriptive statistics that locate LLMs on various scales relative to human learners and provide evidence that LLMs are affected differently from humans by task features such as visual elements or sequential reasoning steps [24, 26]. Yet, little research has examined more generalizable ways of measuring LLM capabilities in relation to human learners. Assessment, by definition, is a proxy that seeks to generalize from sparse samples of student performance (e.g., across time or content). For human learners, the design and interpretation of such proxies have been refined through decades of assessment research and implementation. For GenAI agents, however, it remains far less clear what inferences assessments support.

Mature frameworks from educational measurement – such

¹In this paper, we use GenAI to refer to generative AI in the broad sense; LLMs to refer specifically to language-oriented models; and chatbots to refer to conversational agents powered by LLMs.

as, but not limited to, Item Response Theory (IRT) [5] – may offer principled ways to assess the capabilities of GenAI. Recent studies [21, 22, 13] have shown that psychometric modeling can be extended beyond its traditional applications to capture systematic differences between human learners and chatbots, mainly for identifying GenAI. Within educational data mining, a central paradigm is that item-level analysis can reveal underlying cognitive processes, providing a foundation for more valid assessment and adaptive learning design [2, 11].

Combining these approaches, our work moves the center of attention from *examinees* to *items*, and its goal is identifying items on which chatbots and humans exhibit differential behavior. Identifying such items serves two important purposes. First, to better understand the dimensions that make tasks easy or difficult for LLMs relative to human learners. Second, to apply these insights to the design of assessments that are better adapted – in terms of validity, reliability, and fairness – to scenarios in which students may work (legitimately or not) with GenAI.

Within psychometrics, there is a collection of methods known as Differential Item Functioning (DIF). DIF refers to situations in which an assessment item functions differently for subgroups of learners distinguished by a characteristic unrelated to the construct being measured (typically a demographic characteristic). In DIF terminology, the baseline group is referred to as the *reference* group, while the group of interest is referred to as the *focal* group. DIF methods go beyond simply comparing overall performance between the reference and focal groups by controlling for student ability, *thereby distinguishing true item bias from general group-level performance differences*.

DIF analysis is widely used in assessment design to flag biased or poorly constructed items, as such items compromise the validity and fairness of the assessment [17]. For example, a mathematics item requiring advanced reading comprehension may disadvantage learners with equivalent math ability but weaker language skills (e.g., second language learners). Methods to detect DIF include non-parametric methods, such as Mantel–Haenszel [16], or parametric ones, such as those based on IRT or logistic regression [15].

The rationale of the current work is to examine whether DIF methods can be useful to identify items on which human learners and chatbots differ. The observation is that chatbots can be referred to as the ‘focal’ group, while humans are the ‘reference’ (baseline) group. This is formulated through the following research questions (RQs):

RQ1: *Can DIF techniques identify items that function differently for human learners and chatbots?*

RQ2: *What key characteristics do subject-matter experts identify in items on which chatbots exhibit differential performance compared to human learners?*

To study these questions, we developed, in a stepwise manner, a method that combines DIF analysis with additional item-level criteria and identifies items on which chatbots and human learners show differential behavior. This method was

progressively refined and evaluated on human learners and GenAI responses to two assessment instruments taken from two very different contexts: a high school chemistry test administered as a formative assessment, and the quantitative section of a high-stakes psychometric entrance exam for higher education. The GenAI responses were generated using six chatbots (see the Methodology section for details). Following that, subject-matter experts conducted a qualitative analysis of the chemistry DIF items to identify key dimensions that may lead to the differential behavior.

The key contribution of this paper is proposing a theory-inspired and statistically robust method to identify items that exhibit differential behavior for chatbots relative to students, and demonstrating its effectiveness in providing item analytics to those who wish to incorporate GenAI-related considerations into assessment design.

2. PSYCHOMETRIC PRELIMINARIES

2.1 Item–Total Correlation and Rest Score

Item–total correlation (ITC) is defined as the correlation between the score on a single item and the *rest score* for that item, which is the aggregated performance across all the *other* items in the test (also named ‘corrected ITC’). It assesses the consistency of an item with the rest of the test, providing a measure of *item discrimination* – how well the item distinguishes between examinees with high versus low overall performance on the test. ITC analysis is used in test design to improve validity and reliability. Following common practice, ITC values of 0.20 or higher are considered acceptable [20]; we adopt this threshold in the present study.

2.2 Differential Item Functioning (DIF)

DIF methods test whether an item functions differently across groups of respondents conditioned on the respondents’ *abilities*. Here, ‘ability’ refers to a respondent’s overall proficiency on the instrument and is measured using the per-item rest score, defined as

$$R_{i,-j} = \sum_{k \neq j} X_{ik},$$

where $R_{i,-j}$ is the sum of the respondent i ’s correct responses across all items excluding the target item j . We use $R_{i,-j}$ as an observed ability proxy throughout the analyses. In DIF terminology, the *reference group* is the baseline for comparison (here, humans), and the *focal group* is the group tested for differences (here, chatbots). For analysis, we denote each response as $X_{ij} \in \{0, 1\}$ for respondent i on item j , where 1 indicates a correct answer and 0 an incorrect one. Conceptually, an item exhibits DIF if two respondents with the same ability, but from different groups, have systematically different probabilities of answering it correctly. Throughout, we adopt a consistent direction convention: **POS/NEG** (positive/negative) indicates a chatbot advantage/disadvantage (higher/lower conditional probability of answering correctly than humans at the same ability level).

2.2.1 Mantel–Haenszel DIF (MH-DIF). MH-DIF compares focal and reference groups within discrete strata of the ability proxy and aggregates these comparisons into a

common odds ratio (α_{MH}). Odds ratios less than one correspond to POS (chatbot advantage), while odds ratios greater than one correspond to NEG (chatbot disadvantage). Following ETS guidelines [29], effect sizes are categorized by the magnitude of $|\log(\alpha_{MH})|$: Category A (*negligible*) if $|\log(\alpha_{MH})| < \log(1.5)$, Category B (*moderate*) if $\log(1.5) \leq |\log(\alpha_{MH})| < \log(2)$, and Category C (*large*) if $|\log(\alpha_{MH})| \geq \log(2)$. MH-DIF is widely used for detecting *uniform DIF*, or differences that are stable across the ability range, and is particularly robust under stratification with limited sample sizes. For more details on MH-DIF, see [19].

2.2.2 Logistic Regression DIF (LR-DIF). LR-DIF provides a more flexible framework that models the probability of a correct response as a function of ability, group membership, and their interaction:

$$\text{logit } P(X_{ij} = 1) = \beta_0 + \beta_1 R_{i,-j} + \beta_2 G_i + \beta_3 (R_{i,-j} \times G_i),$$

where $G_i = 0$ for humans and $G_i = 1$ for chatbots, and $R_{i,-j}$ is the per-item rest score used as an ability proxy. Nested model comparisons allow us to test separately for *uniform DIF* (via the main group effect β_2) and *non-uniform DIF* (via the interaction term β_3). In cases where the interaction implies different directions across the ability range, we label the effect as *Sign-change DIF*. From the fitted logistic function we compute likelihood-ratio p -values (m_0 vs. m_1 , m_1 vs. m_2), McFadden’s $\Delta R^2 = R_{m_2}^2 - R_{m_0}^2$ with $R_m^2 = 1 - \ell(m)/\ell(m_{\text{null}})$, and at-median odds ratios, which indicate how much more or less likely chatbots are to answer correctly relative to humans at the median ability level after trimming. We adopt $\Delta R^2 \geq 0.035$ as the threshold for a meaningful effect size [9], and use these measures together to detect and characterize DIF effects.

3. METHODOLOGY

This section describes how the psychometric measures above were applied to design a statistical method that identifies items that function differently for humans and chatbots (RQ1), and also the process through which DIF items were subjected to qualitative analysis by subject matter experts to characterize human–chatbot DIF behavior (RQ2).

3.1 Method Design

Task Definition and Modeling. Our task is to develop a statistically principled method that detects and characterizes multiple-choice items exhibiting systematic response differences between human learners and LLM-based chatbots. Adopting the DIF terminology, we refer to humans as the reference group and chatbots as the focal group.

Design Process. To develop a method for identifying true item-level differences between humans and chatbots, we followed a stepwise process, which is outlined in Fig. 1. Below, we briefly describe the key steps (the implementation and results are detailed in Section 4):

Step 1 – DIF Analysis: Given response data from humans and chatbots (see data description below), we applied the two DIF methods defined above (MH- and LR-DIF). These methods allow us to identify items where humans and chatbots with the same ability level differ in the probability of providing a correct response (here, *ability* refers to respon-

dents’ overall proficiency on the instrument). We interpret the items flagged in this step as candidate DIF items.

Step 2 – Negative Control Analysis: To control for false positives and validate the overall stability of the DIF analysis, we performed a negative control analysis (Placebo Test [7]). For that, Step 1 was reapplied to 50 null datasets per instrument, where the chatbot group was replaced with random samples of human responses (drawn without replacement). On each null dataset, both MH- and LR-DIF were rerun, and we recorded which items were flagged as DIF to compute their false-positive rates.

Step 3 – Applying Psychometric Diagnostics: Next, we analyzed the items flagged by the placebo test in Step 2 using psychometric validation measures, with the purpose of identifying whether these items have certain psychometric characteristics (e.g., low ITC) that can explain their instability and be used to filter them upfront.

3.2 Instruments and Data

We used two assessment instruments from two very different contexts. The first was a chemistry diagnostic test administered as a formative assessment activity in preparation for the high school matriculation exam. It consisted of 22 multiple-choice items answered by 931 students. The second was the quantitative section of a psychometric entrance exam used for higher-education admissions, containing 40 items with over 4,800 respondents. Both instruments featured multi-modal items with figures, images, and formulas.

To generate chatbot data, we collected responses from six LLM chatbots representing three distinct model families. For each family, we included both a previous version and the most recent release available at the time of study, specifically: OpenAI’s GPT-4o and GPT-5.2; Google’s Gemini 1.5 Pro and Gemini 3 Pro; and Anthropic’s Claude 3.5 Sonnet and Claude 4.5. Via the chat models’ web interfaces, we prompted each model to generate a list of final answers to the instrument, which we attached as a PDF file in the prompt. The models typically returned numbered answer lists, which we exported to CSV. This process was repeated 20 times per model, simulating 20 “artificial students” for each chatbot. Overall, for the analysis, we had 120 responses from the six chatbots pooled together. Runs were conducted in separate sessions to prevent memory effects, and variation across responses was introduced by using a non-zero default temperature setting and default parameters set. Since both assessments included multimodal content, it was important that all models supported visual inputs. When a model skipped an item or did not provide a valid option, the response was counted as incorrect (similar to how student responses were treated). The combined human and chatbot responses were then balanced by down-sampling the more abundant data source to construct a dataset at a 1:10 ratio.

The ability distributions of the chatbots and human learners on both instruments are presented in Fig. 2. The bimodal distribution of the chatbots’ ability originates from the division between previous and recent chatbot versions (to examine the robustness of the method to model generations, we validated the method reported below also on a subset containing only the previous versions). Table 1

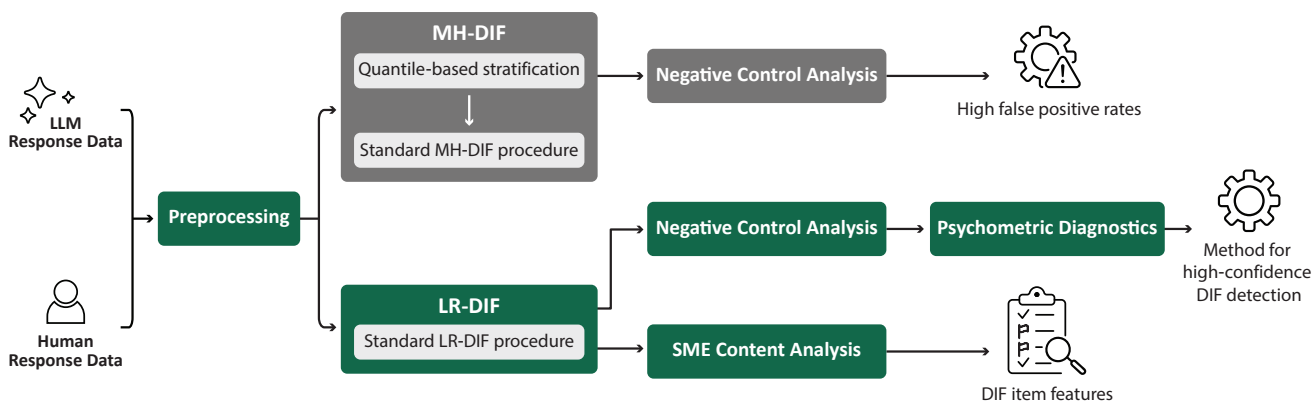


Figure 1: Implemented methodological framework for human-chatbot DIF analysis.

shows the fraction correct of each group on the chemistry items (which were further analyzed by the subject-matter experts). As shown, although the overall success rate of both groups is similar, the chatbots have higher success rates on most items. These sophisticated relations can be handled by the MH- and LR-DIF, as we later demonstrate.

3.3 Qualitative Analysis by Domain Experts

The qualitative analysis was conducted only for the items flagged for DIF by the LR-DIF method in the chemistry instrument. Two subject-matter experts participated in the review process. One expert leads the national team responsible for the annual analysis of the high-stakes chemistry matriculation examinations. The experts received the instrument together with a table summarizing the DIF properties of the flagged items (e.g., POS or NEG; see Table 3) and were asked to qualitatively analyze the DIF items. The two experts independently reviewed, examined, and characterized the items. The two experts then engaged in a joint discussion to compare their evaluations and resolve discrepancies until consensus was reached regarding the characterization of each item.

4. EXPERIMENTS AND RESULTS

This section applies the pipeline outlined in Section 3 and illustrated in Fig. 1. The computational part was applied to the psychometric and chemistry data. The qualitative part was applied to the chemistry dataset, and is described in Subsection 4.5. We then conclude with the resulted procedure for detecting human-chatbot DIF items.

4.1 Preprocessing

The following properties were computed *per item* j . Let $X_{i,j} \in \{0, 1\}$ denote correctness for respondent i on item j .

(i) **Ability proxy.** We defined an *ability proxy* for respondent i as the *rest score* $R_{i,-j}$, the respondent’s sum of correct answers across all items *except* item j . Using rest score for respondent i , as opposed to using the global score of respondent i , avoids contamination from the target item and provides a consistent matching variable (i.e., a covariate used to condition comparisons on respondents’ overall proficiency).

(ii) **Trimming non-overlapping ability ranges.** Following DIF

analysis guidelines to ensure that both groups were compared only within the overlapping strata of the ability proxy, we removed the human responses in the ability ranges not populated by the chatbots.

(iii) **ITC.** We computed the items’ ITC values, which serve as indicators of item quality, and are later used as a diagnostic filter in the validation stage (see Subsection 4.4). ITC values were computed on the human-only, pre-trimmed responses, as restricting to human data ensures item quality is evaluated relative to the reference group, while using pre-trimmed data preserves the full range of human ability for a fair discrimination estimate.

4.2 RQ1: DIF Technique for Identifying Items on Which Humans and Chatbots Diverge

We applied MH-DIF and LR-DIF on the data, as defined in Subsection 2.2.1, with humans and chatbots as reference and focal groups, respectively.

1. **MH-DIF.** For each item, we: (a) Stratified the trimmed responses by rest score into up to four quantile bins (or fewer when necessary). (b) Applied `difMH` to compute the MH statistics retrieved below. (c) Retrieved the common odds ratio α_{MH} , the chi-square statistic, and its p -value. (d) Flagged items as DIF if $p < 0.05$ and the effect size category (defined in Section 2.2.1) was B or C. (e) Assigned direction: $\alpha_{MH} < 1 \Rightarrow$ POS (chatbot advantage), $\alpha_{MH} > 1 \Rightarrow$ NEG (chatbot disadvantage).

2. **LR-DIF.** For each item, we: (a) Fit three nested LR models: (i) Ability-only model. (ii) Uniform DIF model: add group membership as a predictor (iii) Non-uniform DIF model: add the ability \times group interaction. (b) Computed likelihood-ratio test p -values for uniform (p_{uniform}) and non-uniform ($p_{\text{nonuniform}}$) DIF. (c) Calculated McFadden’s ΔR^2 (fit improvement of the full model over the ability-only baseline). (d) Flagged items as DIF if

$$(p_{\text{uniform}} < 0.05 \text{ or } p_{\text{nonuniform}} < 0.05) \text{ and } \Delta R^2 \geq 0.035.$$

(e) Summarized flagged items by: (i) Log-odds contrast at median ability. (ii) Odds ratio = $\exp\{\text{log-odds contrast}\}$. (iii) Probability gap between humans and chatbots at median ability. (f) Classified direction as POS, NEG, or *sign-*

Table 1: Fraction correct per item for students and chatbots (chemistry).

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22
Students	0.836	0.531	0.603	0.699	0.921	0.918	0.968	0.968	0.924	0.944	0.740	0.563	0.820	0.827	0.783	0.429	0.484	0.487	0.374	0.552	0.672	0.658
Chatbots	0.979	0.792	0.917	0.781	0.802	0.479	0.667	0.990	1.000	0.979	0.510	0.865	0.979	0.458	0.365	0.917	0.979	0.688	0.802	1.000	0.604	0.240

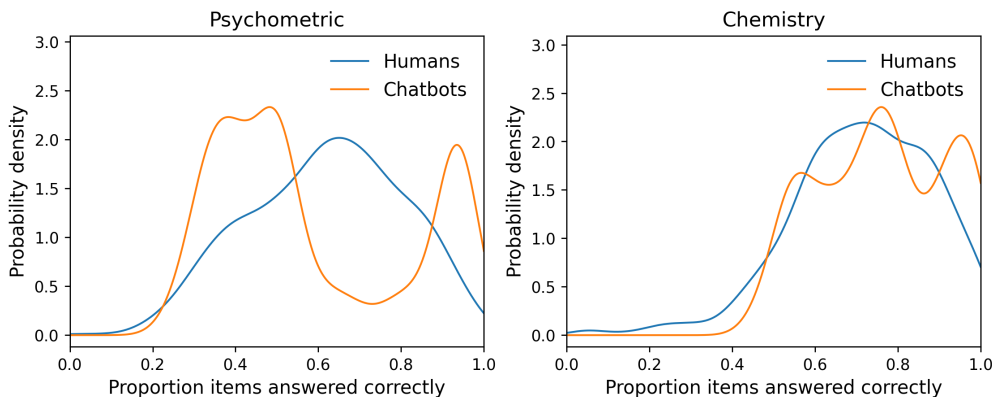


Figure 2: Human-Chatbot ability distribution.

Table 2: MH-DIF: Summary of DIF-flagged items.

Inst.	POS	NEG
Psych.	4-6, 9-10, 12, 16, 21-22, 25, 28, 30-33	11, 13-15, 17, 20, 23-24, 34-38, 40
Chem.	5-7, 11, 14-15, 21-22	1-3, 12-13, 16-19

Table 3: LR-DIF flagged items: \tilde{Q} : uniform; \tilde{Q} : non-uniform.

Inst.	POS	NEG
Psych.	$\tilde{1}$, $\tilde{13}$ - $\tilde{15}$, $\tilde{34}$, $\tilde{40}$	$\tilde{16}$, $\tilde{21}$, $\tilde{25}$, $\tilde{31}$
Chem.	$\tilde{12}$, $\tilde{16}$, $\tilde{17}$, $\tilde{19}$	$\tilde{5}$ - $\tilde{7}$, $\tilde{14}$, $\tilde{15}$, $\tilde{22}$

change if the group effect changed sign at different ability levels.

The items flagged as DIF by both methods are presented in Table 2 and Table 3. As can be seen, MH-DIF flagged a large set of items as DIF, and particularly, a large number of items as POS (chatbot advantages conditioned on ability), including items on which the students scored higher when compared only on raw averages (e.g., items 6-7 in the chem. instrument; see average group scores in Table 1). The LR-DIF was much more conservative, as expected, flagging less items overall. Its more sophisticated modeling enabled it also to identify items with non-uniform DIF (DIF magnitude changes across the ability ranges; e.g., chem. item 12).

These patterns indicate that chatbot-human differences are not only widespread but also vary by item, by direction, and by ability levels, highlighting the diagnostic power offered by DIF analysis. However, in light of the considerable disagreement between the methods (e.g., chemistry item 1, which was flagged as POS by the LR-DIF and NEG by the MH-DIF), we conducted Negative Control Analysis to dis-

criminate true DIF from statistical artifacts.

4.3 Negative Control Analysis

For both datasets, we created $R = 50$ null datasets in which the chatbot group ($n=60$) was replaced by a random sample of human respondents (without replacement). We then reran the MH-DIF and LR-DIF pipelines on each null dataset. Table 4 reports the false-positive counts of items that were flagged at least once under the 50 placebo test runs. Two patterns stand out: (1) MH-DIF exhibits diffuse background noise: many items are flagged once or twice, and several reach 8–10% false-positive rates. This noise profile appeared on both instruments, indicating that, at least in our experiment, MH-DIF alone cannot yield reliable DIF signals without additional filtering. (2) LR-DIF is far more stable. On the psychometric instrument it produced no null flags across all 50 runs, demonstrating excellent specificity. On the chemistry instrument, LR-DIF was generally stable but showed a small cluster of unstable items (e.g., Item 5).

These findings support LR-DIF as the primary DIF detection method. Yet the presence of a small unstable chemistry items cluster, which showed elevated false-positive rates in the null runs, needs additional item-level analysis and filtering (see Section 4.4).

4.4 Psychometric Diagnostics and Results

We incorporated psychometric diagnostics to analyze and refine the candidate DIF set flagged by LR-DIF. As shown in Table 4, Negative Control Analysis confirmed LR-DIF’s stability, though chemistry items 5, 9, and 10 exhibited false-positive rates above 10%. To investigate these items further, we examined their ITC, and found that each had very low ITC values (Item 5 = 0.11, Item 9 = 0.08, Item 10 = 0.04). In contrast, chemistry items with ITC values above 0.20 were consistently stable across null runs. These results reinforced ITC as a reliable diagnostic for DIF analysis, using

Table 4: Negative Control Analysis: The items on each *false_positive_count* level (0 omitted).

Counts	Psych.(MH-DIF)	Chem.(MH-DIF)	Psych.(LR-DIF)	Chem. (LR-DIF)
1	6-7, 10, 12-15, 17, 19-21, 24, 27-28, 33, 38	2-3, 6, 9, 13-14, 17-19, 22	—	14
2	9, 18, 22, 29-32, 37, 39	7, 16	—	6-8
3	2, 8, 16, 26, 34	11, 20, 21	—	—
4	25, 35	—	—	—
5	23, 40	12	—	—
6-7	—	—	—	—
8-11	—	—	—	5, 9-10

the standard 0.20 as a cutoff for ‘acceptable’ [20].

4.5 RQ2: Domain Expert Analysis

As described in Subsection 3.3, the qualitative analysis of the DIF items was conducted by two subject-matter experts who received the instrument and a table that listed the properties of the LR-DIF items, taken from Table 3.

As shown in the Table 3, Items 12, 16, 17, and 19 had POS DIF, meaning that chatbots had *higher* likelihood of answering them correctly than humans at the same ability level. In 12, the system successfully circumvented a common alternative conception according to which sodium chloride is composed of atoms. Instead, it correctly distinguished between the ionic lattice structure of NaCl and the behavior of its constituent particles in solution, emphasizing that dissolution involves the separation of Na⁺ and Cl⁻ ions rather than the breaking of atoms or the formation of atomic species. In 16, which involves a sequence of solution manipulations, the system effectively maintained consistency by monitoring the relationship between solute moles, solution volume, and concentration across several experimental stages. In 17 demonstrated a scientifically appropriate interpretation of the law of conservation of matter by situating the assessment of mass change at the level of the entire experimental system, rather than restricting the analysis just to the original metal. Item 19 further highlighted the system’s capacity to apply algorithmic reasoning: the multiple-choice structure facilitated accurate assignment of oxidation states, identification of electron transfer, and correct designation of oxidizing and reducing agents. Collectively, these examples underscore the system’s strength in tasks that prioritize rule-governed reasoning and conceptual clarity, where adherence to canonical chemical principles prevents susceptibility to intuitive but scientifically inaccurate interpretations.

In contrast, items 6, 7, 14, and 15, had NEG DIF, meaning that chatbots had *lower* likelihood of answering them correctly than humans at the same ability level. These items place greater demands on visual interpretation, linguistic nuance, and the execution of complex, multi-step procedures. Items 6-7 require learners to interpret partial microscopic models, such as diagrams omitting solvent molecules, or to map symbolic labels (A–C) onto microscopic and macroscopic representations of aqueous solutions. These forms of representation, which depend heavily on implicit visual conventions, posed substantial challenges for the system, whereas students were able to draw upon prior representational.

Questions 14 and 15 illustrate a further point of divergence. Beyond the computational demands involved, these ques-

tions also require sensitivity to linguistic nuance, namely the ability to interpret subtle wording, implied meanings, and context-dependent scientific language. For example, in Question 15, students needed to recognize that the statement “the volumes were measured under the same conditions” implicitly justifies the use of proportional relationships between gas volumes and moles according to Avogadro’s law. Similarly, Question 14 requires interpreting the phrase “the volume of 1 mole of gas is 25 liters” as applying specifically to gaseous substances in the reaction, while distinguishing between gaseous and liquid species within the balanced equation. Such interpretive demands rely not only on factual chemical knowledge, but also on sensitivity to how scientific information is linguistically framed. Question 14 further involves a multi-stage stoichiometric calculation, proceeding from gas volume to moles, through mole ratios, and finally to mass determination, with each step contingent on the correctness of the one before it. Such extended procedures are particularly vulnerable to intermediate errors in algorithmic reasoning. Question 15 compounds this difficulty, as errors originating in the previous calculation can propagate forward, whereas students often display the metacognitive awareness required to detect and correct inconsistencies in their own work.

Taken together, these findings suggest that the chatbots excel in domains governed by formalized conceptual rules. By contrast, students retain a relative advantage in tasks that demand flexible interpretation of visual representations, sensitivity to linguistic subtleties, and careful monitoring of extended quantitative reasoning processes.

Resulting Method. The conclusion from the experimental results, which both validated the computational process statistically, and evaluated the meaningfulness of the LR-DIF results with subject matter experts on a sample dataset (chemistry), yielded that applying LR-DIF to items with $ITC \geq 0.2$ is a reliable method for identifying items that, with high confidence, exhibit differential behavior across humans and chatbots.

5. DISCUSSION

The LR-based DIF approach developed and piloted in this research provides a method and conceptual framework for identifying assessment items that show differential behavior between humans and chatbots. In two assessment contexts and on leading chatbots, we demonstrated that the method provides reliable and stable results. A key observation that DIF, and this research, makes, is that differential behavior is a more nuanced property than merely looking at overall performance gaps between groups, since the DIF analysis controls for ability level, and can identify items as

positive/negative DIF (chatbots over/under-perform learners of the same skill) even if the difference between the mean group performance is in the opposite direction. Our LR-DIF based method offers a modeling that also enables the detection of non-uniform DIF patterns in which either the magnitude or direction of group differences vary across ability levels. Our experiments with both the MH- and LR-DIF highlighted that the MH-DIF, typically the default choice due to its simplicity, produces considerable noise and a high false positive rate. The LR-DIF excelled in this aspect as well. Its downside is its relative complexity and the fact that it typically requires larger sample sizes for stable estimates.

Using the DIF-method to deliver content analytics to the subject-matter experts (RQ2) drove an analysis that yielded interesting insights. As reported in Subsection 4.5, analyzing the POS DIF items, we found that the chatbots managed to circumvent distractors aimed at surfacing alternative conceptions (sometimes referred to as ‘misconceptions’) commonly held by students. Given that LLMs can be cognitively biased due to their training data [6], which likely included student data also representing wrong or incomplete knowledge [27], the fact that the chatbots dodged these designated ‘traps’ is interesting, and is in disagreement with previous work that found moderate alignment between LLMs and student misconceptions [12]. On the contrary, experts’ analysis of the NEG DIF items revealed that the chatbots underperformed on items requiring visual interpretation and connecting visual and textual information, items that their wording required understanding linguistic nuances, and those requiring performing multi-step problem solving procedures. These findings reinforce previous work about GenAI problem solving in STEM domains [24, 26].

More generally, this research demonstrates a fruitful application of assessment and measurement theory to educational data mining, with the purpose of establishing theory-grounded approaches for understanding GenAI capabilities, and for developing assessment in the GenAI era.

Limitations. This research is the first, to our knowledge, to apply DIF for analyzing chatbot assessment data and, as such, naturally has several limitations. In terms of internal validity, the chatbots’ bi-modal performance (fraction correct; see Figure 2) may impact the DIF stratification step, potentially reducing the effective size of the data and thus the method’s robustness. The very high ability that the newer models exhibit may reflect data leakage or prior exposure during training, rather than genuine ability, and may falsely cause them to be identified as positive DIF (models outperform humans). A further limitation of our study is that human and chatbot ability estimates were based on a binary scoring model rather than a polytomous one. More nuanced scoring models may capture partial knowledge or intermediate reasoning patterns, and could therefore affect the ability estimates and the resulting DIF patterns. A related limitation is that DIF methods typically assume a relatively uni-dimensional construct and comparison along a common ability continuum [8, 28]. Given the heterogeneous task types included in the instruments, such as visual versus algorithmic reasoning, as well as the use of multiple LLM architectures, some of the observed DIF effects may reflect multidimensionality or within-group variability rather

than item-level bias alone. In addition, the SME analysis focused only on LR-DIF flagged items, which were shown to favor either humans or chatbots, and did not compare them against non-DIF items. Such a comparison could provide a clearer interpretation of the features that distinguish DIF from non-DIF items and is left for future work.

The main limitation to external validity is that the results are based on a small number of instruments and specific GenAI tools. Additionally, the SME analysis was conducted only for the Chemistry instrument; extending this analysis to the quantitative reasoning instrument would be important for broadening observed patterns to additional domains.

Contribution and future work. The main contribution of this work is the introduction of a theory-driven and statistically sound method for detecting items that exhibit differential behavior between chatbots and human learners, and the presentation of evidence of its ability to yield meaningful analytics for subject-matter experts who seek to integrate GenAI considerations into assessment design. In future work, we plan to build on this foundation for (1) studying the task dimensions of Human–GenAI DIF items, to better understand the capabilities of this technology, (2) incorporating Human–GenAI DIF analysis into assessment design, and (3) Use DIF analysis to evaluate GenAI progress.

Ethics statement. The research was approved by the Institutional Review Board No. 3331-2.

Data and code availability. The materials needed to reproduce the analyses reported in this paper, including the code, sample data, and prompts used to generate the chatbot data, are available in the GitHub repository: ([link](#))

6. ACKNOWLEDGMENTS

This work was supported by the Knell Family Institute for Artificial Intelligence, Israel. The authors thank the National Institute for Testing and Evaluation for providing access to psychometric exam data.

7. REFERENCES

- [1] N. Akbari. The AI cheating crisis: Education needs its anti-doping movement, 2024. Retrieved from <https://www.edweek.org/technology/opinion-the-ai-cheating-crisis-education-needs-its-anti-doping-movement/2024/02>.
- [2] T. Barnes. The q-matrix method: Mining student response data for knowledge. In *American association for artificial intelligence 2005 educational data mining workshop*, pages 1–8. AAAI Press, Pittsburgh, PA, USA, 2005.
- [3] B. Borges, , et al. Could chatgpt get an engineering degree? evaluating higher education vulnerability to ai assistants. *Proceedings of the National Academy of Sciences*, 121(49), 2024.
- [4] D. R. E. Cotton, P. A. Cotton, and J. R. S. and. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2):228–239, 2024.
- [5] R. J. De Ayala. *The theory and practice of item response theory*. Guilford Publications, 2013.

- [6] J. M. Echterhoff, Y. Liu, A. Alessa, J. McAuley, and Z. He. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024.
- [7] A. C. Eggers, G. Tuñón, and A. Dafoe. Placebo tests for causal inference. *Am. J. Polit. Sci.*, 68(3):1106–1121, 2024.
- [8] P. W. Holland and H. Wainer. *Differential item functioning*. Routledge, 2012.
- [9] M. G. Jodoin and M. J. Gierl. Evaluating type i error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14(4):329–349, 2001.
- [10] M. Khalil and E. Er. Will ChatGPT get you caught? Rethinking of plagiarism detection. In *International Conference on Human-Computer Interaction*, pages 475–487. Springer, 2023.
- [11] K. R. Koedinger, E. A. McLaughlin, and J. C. Stamper. Automated student model improvement. *International Educational Data Mining Society*, 2012.
- [12] N. Liu, S. Sonkar, and R. Baraniuk. Do llms make mistakes like students? exploring natural alignments between language models and human error patterns. In *International Conference on Artificial Intelligence in Education*, pages 364–377. Springer, 2025.
- [13] Y. Liu, S. Bhandari, and Z. A. Pardos. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052, 2025.
- [14] P. Lu et al. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- [15] D. Magis, S. Beland, F. Tuerlinckx, and P. De Boeck. A general framework and an r package for the detection of dichotomous differential item functioning. *Behavior research methods*, 42(3):847–862, 2010.
- [16] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4):719–748, 1959.
- [17] P. Martinková, A. Drabinová, Y.-L. Liaw, E. A. Sanders, J. L. McFarland, and R. M. Price. Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE Life Sci. Educ.*, 16(2), 2017.
- [18] A. Prothero. New data reveal how many students are using AI to cheat, 2024.
- [19] H. J. Rogers and H. Swaminathan. A comparison of logistic regression and mantel-haenszel procedures for detecting differential item functioning. *Applied psychological measurement*, 17(2):105–116, 1993.
- [20] A. Shete. Item analysis: An evaluation of multiple choice questions in physiology examination. *J. of Contemporary Med. Educ.*, 2015.
- [21] B. Sorenson and K. Hanson. Identifying generative artificial intelligence chatbot use on multiple-choice, general chemistry exams using Rasch analysis. *Journal of Chemical Education*, 101(8):3216–3223, 2024.
- [22] A. Strugatski and G. Alexandron. Applying IRT to distinguish between human and generative AI responses to multiple-choice assessments. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference, LAK 2025, 2025*, pages 817–823. ACM, 2025.
- [23] T. Susnjak and T. R. McIntosh. ChatGPT: The end of online exam integrity? *Education Sciences*, 14(6), 2024.
- [24] K. D. Wang, E. Burkholder, C. Wieman, S. Salehi, and N. Haber. Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. In *Frontiers in Education*, volume 8, page 1330486. Frontiers Media SA, 2024.
- [25] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang. Scibench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. *arXiv preprint arXiv:2307.10635*, 2023.
- [26] E. Yacobson, Y. Schleifer, Z. Bar-Dov, S. Rap, R. Blonder, and G. Alexandron. Benchmarking ai on standard chemistry exams: Llms still underperform compared to high school students. *Journal of Science Education and Technology*, pages 1–18, 2026.
- [27] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *Br. J. Educ. Technol.*, 55(1):90–112, 2024.
- [28] B. Zumbo. A handbook on the theory and methods of differential item functioning (dif). 01 1999.
- [29] R. Zwick. A review of ets differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement, 2012.