Assessing AI skills: A washback point of view

Meirav Arieli-Attali[1]*, **Beata Beigman Klebanov**[2]*, **Tenaha O'Reilly**[2], **Diego Zapata-Rivera**[2], **Tami Sabag-Shushan**[3], and **Iman Awadie**[3]
[1]Fordham University, USA & One Assessment, Israel
[2]ETS Research Institute, USA
[3]National Authority for Testing and Evaluation, Israel

Abstract

The emerging dominance of AI in the community's perception of skills of the future makes assessing AI skills necessary to help guide learning. Creating an assessment of AI skills shares some challenges with other assessments but also poses new ones. We examine these from the point of view of washback and exemplify using two exploration studies conducted with 9th grade students.

1 Introduction

Washback is the impact of testing on curriculum, teaching, and learning. Positive washback occurs when practicing for the test results in wholesome learning. Negative washback occurs when preparing for the test results in a narrowing of the learning, such as the exclusion of important practices to focus on the skills targeted by the test or on test-taking strategies (Hughes, 1989).

In principle, the moment one makes choices regarding inclusion or exclusion of materials on a test, one sets things up for washback. Any assessment is going to focus on a particular construct – the skill that is the target of the assessment – and necessarily leave out other related skills. Thus, in the context of assessing writing proficiency, the choice of an essay task may have encouraged more attention to this genre; Burstein et al. (2014) found "an extraordinary prevalence of the essay" in USA K-12 context. Furthermore, once one designs a task – making decisions about timing, format, and scoring – there might be a further narrowing of the practice towards what the scoring rubric demands or implies. The debate about the artificial nature of a five-paragraph persuasive essay as a "test genre" is an example. The fact that the assessment task isn't identical to professional or vocational practice does not necessarily undermine validity; persuasive writing for a test, for example, shares important rhetorical elements with OpEd writing for the New York Times (Beigman Klebanov et al., 2019). The divergence between test and non-test contexts is nothing new, in itself. However, we believe assessment of AI skills comes with some new challenges.

2 Moving target

One challenge is specifying exactly what the target construct for the assessment is, within the general umbrella of AI skills. To articulate the construct, one would typically consult the relevant skills frameworks (such as OECD (2025) or UNESCO (2024)) and stakeholders (such as employers, educators, test takers) and model an assessment task on a common application of the target skill. Since AI tools are evolving, their typical and effective uses likewise continuously evolve. Moreover, the pace of the changes in AI tools and capabilities is such that it is likely to outpace the assessment development cycle of ideating, implementing, piloting, revising, and administering an assessment. Thus, the common tasks of today may change by tomorrow, so the assessment, when operationalized, would focus on the skills of yesterday and thus open up the possibility of negative washback, where preparing for the test would entail practicing an outdated use case of the AI technology. To address the challenge of assessing a 'moving target', one could (a) focus on fundamental elements that are less likely to change fast, such as understanding the implications of training on huge amounts of text data, and/or focus on enabling skills such as critical thinking; and (b) implement mechanisms to continuously review and revise the frameworks, constructs, and tasks, using the principles of evidence-centered design (Mislevy et al., 2003).

We note, in addition, that continuously updating the assessment to stay close to the current real-life

*Corresponding authors: mattali@fordham.edu and bbeigmanklebanov@ets.org

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 274–280

use comes with its own challenges. While student engagement and positive washback potential can thereby be improved, the very verisimilitude of the assessment may introduce into the performance variance that is authentic – it is part of real-life activities being mimicked – but might be construct-irrelevant from the point of view of the assessment. We will leave this issue for a more concrete discussion in the context of the exploration studies.

3 Broadening the construct

The second challenge is related to the consequences of selecting the focus of an assessment to be a particular set of AI skills. As explained, any selection would entail a de-selection of other skills that are not part of the target construct. To address this challenge, one could imagine a suite of assessments targeting various AI skills and rotation or sampling of the different assessments, as needed. However, picking specific AI skills does not only de-focus other AI skills. It could also de-focus skills that are not necessarily tied to AI. This is because much of the change brought about by AI is about using AI for tasks people did without AI before; with AI, these can be done faster and perhaps better; in any case, they are done differently.

Making the AI-way to perform a task the target of an assessment may de-focus the non-AI way of doing it. For example, using AI to help brainstorm an idea for a project is an increasingly common use of AI; yet people also think about project ideas themselves and brainstorm with other people. If a task on an assessment of AI skills asks students to brainstorm using AI, would that create washback potential whereby students will *have to* use AI for brainstorming to learn to do it more effectively and efficiently, at the expense of brainstorming with others or thinking creatively for themselves?

The issue of technology phasing out some human skills isn't a new consideration in the broader landscape of technological innovation. Weaving machines and calculators got integrated into the cultural fabric, largely displacing hand-weaving and calculating large sums on paper. What gives us pause is that the skills to be displaced might turn out to be those that are fundamental for the well-being of individuals and societies, such as creativity and thinking together with other people. Would it be wise, or, indeed, ethical, to impact the ongoing societal debate about the importance of skills like creativity by setting up an assessment that opens a possibility for washback against these skills, the element of choice in negative or positive washback notwithstanding?

Discussing assessment validity and washback in language testing, Messick (1996) argued that in order for washback, either positive or negative, to be tied to an assessment, one needs to show that it happened as a result of the assessment. Given the large variety of factors that go into curriculum design and educator choices of how to implement relevant learning and practice, Messick argued that it is more fruitful for assessment developers to consider issues in the assessment that could produce negative washback. Chief among these is construct under-representation. If, in order to succeed on an assessment, one needs to exercise only some of the skills that go into the real-world version of the target construct, the assessment under-represents the construct. Applying this reasoning to the discussion above, perhaps we should consider AI skills as part of broader constructs. Creativity can be exercised with AI or without AI; focusing on only one of these – either one – is likely to under-represent the construct in its current real-life use. Ergo, an assessment that includes brainstorming with AI should also include brainstorming without AI.

4 Exploration studies

In what follows, we exemplify the points raised above via two exploration studies we conducted with 9th grade students from public high schools in Israel; the students came from middle-high SES background in both studies. The first was conducted in June 2024, the second – in March 2025. The first study took place when students and teachers in Israel were only beginning to explore generative AI tools, most haven't tried them yet at all. By the time of the second study, all schools in Israel have introduced AI tools via an "AI month" – activities that included introducing to students several AI tools via specific tasks they needed to complete and submit. Teachers and students alike took part in these activities. The purpose was to show how AI can benefit education and to encourage teachers to implement AI in their teaching. Awareness of cyber security and AI's fake information and hallucinations were also part of this month's activities.

The first study included 10 students in one-on-one cognitive labs, where an experimenter observed a student working on the task. The goal was to explore how students react to and interact with the

emerging generative AI. The second study included 72 students in 3 classrooms, working independently on a structured task on computers and smartphones. According to the regulations by the Israeli Department of Education, high school students (grades 9-12) are allowed to access the internet, including generative AI, directly, with parental consent and with a teacher's mediation. The task for the second study was developed based on the first study's insights and our evolving conceptual framework, inspired by the ECD (Mislevy et al., 2003). We implemented revisions to the framework in an iterative way following the "moving target" of the evolving capabilities of AI.

4.1 Study1 – June 2024

The task in this study consisted of three phases: (1) pre-task planning; (2) information gathering; and (3) preparing a 'product'. The task asked students to plan a 2-day class trip, guiding students through the necessary elements, e.g., choose a site, plan the arrival, choose or plan activities in the site's vicinity, find appropriate places to sleep nearby (hostel or camping) and eat (restaurants or takeouts). The 'product' students were asked to prepare was a trip brochure, one that can be published or sent out to the trip participants (their classmates).

The pre-task planning phase included planning verbally or in writing in front of the experimenter. For the information gathering phase, students were referred to ChatGPT (the experimenter created a free account for them) and asked to fact-check its responses using a search engine (e.g., Google). For the third phase of preparing the trip brochure, students were given Word or PowerPoint templates they could fill with pictures and the information they gathered. Students were told to use critical thinking and creative thinking, as well as to imagine that the trip they are planning could be a trip they take with their classmates. In other words, the task aimed to resemble an authentic use of AI tools and the internet to plan a trip, where it is necessary to verify the information given by the AI (e.g., correct names and details of sites or activities) and ascertain the feasibility of the plan (e.g., the distances between sites can be covered within the allocated time). Each student worked on the task while an experimenter sat beside them. The experimenter gave instructions at the start of the task and took the observer role with little interference unless needed, following the guidelines of cognitive labs in educational measurement (Arieli-Attali et al., 2023). The time allocation was 90 minutes.

4.2 Study 2 – March 2025

Based on our evolving Media & AI literacy framework and insights from Study 1, we designed a computerized scenario-based task, where students followed a storyline in which they were asked to help a tour guide plan a trip for a youth group. The task was structured such that the tour guide was the one who is planning the trip step-by-step, posing questions or needs in which he is requesting students' help in gathering the information for him. Thus, students were not asked to do the planning themselves nor did they have freedom in directing it; rather, they were requested to gather bits of information at each step to help the tour guide with his planning. Media & AI literacy items were incorporated as part of the information gathering process; critical and creative thinking items were incorporated as part of the storyline. Students had access to ChatGPT by opening a different tab on their device; this activity was not logged. The time allocated to the task was 90 minutes.

We now report on some insights from both studies to illustrate our main arguments above.

5 Discussion: Moving target

5.1 Changing AI capabilities

The target construct of AI literacy was composed primarily from the three previously well-researched constructs of (1) digital literacy; (2) media information literacy; and (3) critical thinking. One needs basic digital skills to operate digital tools on a computer or a smartphone in order to perform any AI literacy task. Media information literacy is needed not only in order to understand how and where to search for online information and identify its sources, but also to create, share or publish information to achieve one's goals. As information online is not always reliable, students need to apply their critical thinking skills in any online interaction, including when using generative AI.

As discussed above, one challenge is to define the skills of *today*, as AI capabilities change rapidly. Examining students' work in the two studies, less than one year apart, illustrated this point. In June 2024, the LLMs in Hebrew (e.g., ChatGPT, Gemini, Claude) were providing numerous fake details (or hallucinations) and make mistakes in phrasing in Hebrew. Thus, in the first study, we could focus on asking students to fact-check and edit the AI re-

sponse, exhibiting their critical thinking skills. For example, some of the AI responses at that time included restaurants that do not exist, fake distances that would take more than eights hours' drive between the breakfast site and the lunch site (it takes less time than that to cross the country north-to-south or east-to-west), wrong details about the sites or the activities available at the sites. Less than a year later, the LLM responses were almost entirely accurate and very well phrased. Thus, while it is still the case that one would need to check the important details before embarking on the trip – in the manner of "measure twice, cut once" prudent planning – the editing trace of the final assessment product would be unlikely to contain substantial revisions. From the point of view of washback, it was no longer the case that students would truly grapple with fake information through this task, so using the assessment task to help set them on a course towards developing and practicing a critical attitude was no longer a viable option, at least not if one were using the publicly available AI tools as-is, without, for example, intentionally introducing incorrect information through engineering a prompt that would mediate between the student and the generative AI tools.

5.2 Challenges of approximating real-life use

Mirroring a real-life use case of AI can help set things up for positive washback through real-life applications of the practices encouraged by the assessment. Authenticity was thus a leading aspect of task development.

In the first study, the task was to plan a trip from scratch, with little guidance and only a few constraints. The task asked student to imagine that they are really going to invite their friends to this trip. As students were performing the task, our experimenters were watching them thinking out loud, and documenting their actions. Examining the trip brochures as the task "product" each student submitted and comparing those to the experimenters' protocols on student actions during task performance yielded the surprising insight that while it was evident that the task required and the students exhibited their critical and creative thinking skills in the process of planning the trip, the products were poor evidence of these processes. Some of the more critical and creative students ended up with a relatively poor brochure, due to poor digital or graphic skills or poor decision-making skills.

For example, one student's brochure was a paragraph describing the trip she planned, having no pictures, links, maps or any visuals or arguments that may persuade her friends to join the trip. In addition, some of the information was not fully accurate. The product would receive a low score. However, the experimenter protocol of that student revealed a thorough search, taking into account different conflicting considerations, and validating the information in many cases except one or two cases where she failed to do so; unfortunately, the latter found their way into the final brochure. Although the critical skills were not executed to the best, only the worst of them were evident in the outcome, masking all the other cases where they were executed correctly. It seemed that it was easier to find traces of mistakes in the final products rather than of correct conduct.

In addition, going through the protocols revealed that some students failed the task completely due to poor decision-making skills. They ended up hesitating and trying out different routes, and although they exhibited good media and AI skills and even good critical thinking – they finished the task without any product at all. Thus, while this aspect could suggest strong student engagement with the task and its authenticity in that a one-hour planning activity might not yield any plan that satisfies the traveller's standards, it introduced decision-making as one of the constructs assessed, which we judged to be outside of our desired assessment focus.

Finally, we found that the difficulty of the task depended very much on how students decided to approach it. Some students chose an "easy" trip, part of which was already stated in webpages of the chosen sites, while others chose to take a more challenging route of creating everything from scratch, trying to come up with their own creative combinations, some of which turned out to not be feasible at all. It was the case that those who chose the easier task performed better – in terms of the quality of the final product – than those who chose the harder task. Thus, the task itself was not comparable between students, creating an additional challenge from the point of view of scoring. Even putting this issue aside and examining the products themselves convinced us that creating a common scoring rubric would be extremely challenging due to variation across the submitted brochures.

The issue of variation in both process and product that comes with a closer approximation to real-life is not a unique challenge for assessing AI skills. However, the sheer extent of possibilities for vari-

ation may be a hallmark of real life in the era of fast-paced advances in AI technology. That is, the fact that one could quickly come up with, check, and discard a lot of different ideas is related to the strengths of generative AI in idea generation and in instant provision of a wealth of relevant information on almost any conceivable topic. Similarly, the possibility of a large variation in the quality of the visual designs of the brochures created *in mere minutes* may have come about due to the generative AI-induced amplification of differences in the student's independent design skills: It isn't only the visual artists among the students who could come up with visually compelling brochures in a matter of minutes, but also those students who could articulate the imagined designs in a textual prompt.

Based on the results of study 1, we developed a much more structured version of the task for the second study, so that students had less freedom in deciding on the type of trip and more opportunities to show clear evidence of their media and AI skills, as well as more direct focus on their critical and creative skills alongside their media and AI skills.

Scenario-based assessment is a promising paradigm for structured tasks where multiple aspects of a skill can be targeted through different elements of the scenario, thus potentially supporting both standardization and authenticity (Sabatini et al., 2020). In scenario-based assessment, the different discrete items, each targeting an aspect of the skill, are integrated into a thematically coherent whole, where the storyline resembles enough the real-world situation that it allows for representing more aspects of the target real-world skill.

In the second study, we designed a storyline where a tour guide needed to plan a trip for his youth group and is asking for assistance in the planning. The guide needs information which he asks the students to search for and verify. Although this task lacked some of the agency and authenticity of the original where students did the full planning themselves, this task included discrete items within the storyline aimed to elicit student critical thinking skills. For example, as part of the tour preparation, the guide is looking to create a post about the site the youth are going to visit, finding in social media a slogan stating a specific (incorrect) fact about the site. The tour guide then asks students whether he should share that slogan. This discrete item within the scenario elicits student critical thinking of fact-checking information before sharing.

In principle, an alternative to structuring the task could be to use not only the final product but also process data as a basis for assessment. For example, students could be asked to submit preliminary ideas, the LLM prompts and search queries they used, and/or reflections as they move through the task; the richer data could potentially support giving students credit for exhibiting the target thinking patterns even if the outcomes are not clearly reflected in the product. To allow for drawing evidence of skills from process data, one would need to clearly articulate the target construct and think through the extent to which fruitful thinking patterns in the context of the task can be reliably identified irrespective of the quality of the final product.

6 Discussion: Broadening the construct

As part of media-information literacy and critical thinking skill, an essential skill that was particularly identified as needed for AI literacy is "prompt engineering", that is, the ability to write to the LLM an appropriate request that will yield the desired response. Generally speaking, as students are more accurate and detailed in their request to the LLM, they may get a better response. Specifically, in the trip planning task, if the assessment focuses on the prompt engineering aspect of the task and lets the gen-AI do the planning, we might neglect the human – non-AI – skill of planning. We sought to learn about pre- and post–AI–use planning by designing a task where students first plan without AI and later plan with the aid of AI. We examined this issue in both studies.

In the first study, students were first asked to plan a blueprint of a class trip verbally while talking to the experimenter, and the experimenter documented what students said. This first stage was primarily aiming to elicit students' planning skill, while it also served as an engagement means to ease the transition to the AI task. The task instructions included several restrictions to the desired class trip so as to give some structure to the open-ended task, yet left it open enough to allow for students' planning. The instructions were: "You should plan a two-day class-trip for your class to a historic site in the vicinity of... ; you should find a place to spend the night (hostel or camping) and activities only for the first day. The activities should be appropriate for a group of students your age. You should plan for one morning activity and one afternoon activity, and lunch in-between. You should ignore for now budget or security considerations." After students

finished telling the experimenter their plan and said that they are satisfied with it, they were then asked to open ChatGPT and ask it to detail or improve their blueprint. At this point they were facing the computer, and the experimenter was sitting behind them documenting their actions.

The main observation reported by the experimenters was that there was almost no connection between the initial blueprint plans students described at the first stage and the second part of the task. This was manifested in two main ways: (1) Students lacked good prompt engineering skills, that is, they made general requests from the LLM, ignoring what they already came up with; (2) The LLM response took a different route and students continued with the LLM route, forgetting their own. Thus, although students did invest creativity and effort in generating initial ideas, they either did not feel committed to these plans enough to pursue them with the help of LLM, or did not know how to do that and opted for the more generic suggestions by the LLM in response to a general prompt. We inferred from this experience that unless the students' initial ideas were elicited effectively and recorded to serve as part of the assessment data, the post-AI version is unlikely to reflect these ideas. This disjoint nature of the two brainstorming experiences is a challenge in designing a coherent scenario-based task that would cover both effectively.

In the second experiment, although the task was much more structured and required less overall planning, we did preserve the aspect of planning on a smaller scale. At some point it the task, the tour guide asks students to plan a one-hour activity around the theme of the site they're visiting. The students were given instructions (or constraints) about the activity – the theme of the site, the timeframe of the activity (one hour), the materials they have (ancient coins), and that the activity needs to be a group activity to suit a group of students of grade 9. After students submitted their planned activity, they were asked to open ChatGPT and now ask the LLM to detail or improve their plan, and the write down the AI response. There followed a question asking students to compare the AI activity plan to their own initial activity plan. The results largely replicated those of the first study – only 21% of the students actually made a thoughtful comparison, explaining in detail what AI added and why it was a better plan (14%) or why they decided to reject AI's elaboration and stay with their plan (7%). The rest of the students did not engage in the activity as intended: 30% said either AI's or their own plan was better, without explanation, while 45% said they did not know or did not answer the question at all. The remaining four students said that they used ChatGPT to help them come up with the original idea to begin with, therefore no comparison was necessary. While disengaged responses or disconnected own and AI plans dominated, we did obtain, from the 21% of the students, the intended behavior, where the two plans were connected and a meaningful evaluation and comparison were conducted. We are considering ways to encourage this behavior, both during test and in the form of washback – a learning activity where students can practice having their own ideas meaningfully elaborated by AI.

Limitations. In the current discussion, we exemplified assessment of AI skills in a stand-alone, non-disciplinary way. Additionally, the focus was on practical skills rather than on understanding how AI works or on AI ethics. We leave a discussion of contextualization and of ethics to future work.

7 Conclusion

Considering the emerging need to assess AI skills, we present some challenges related to fixing an AI-skills-focused construct to target in an assessment. One challenge is the rapid evolution of AI capabilities, which may lead to assessing today AI skills of yesterday; the other is the hazard of under-representing a broad, human-centered construct by focusing on the AI-reliant way to exercise the relevant skill. We illustrated via two exploration studies the need to revise and refine both the conceptual framework and the tasks themselves, in order to capture the changes in AI capabilities and AI practices and experiences.

We consider washback – the impact of testing on teaching and learning – to be an important motivation, keeping in mind that an assessment task can be used as a model by teachers to prepare students. Grounding both assessment and instruction of AI skills in common AI Literacy frameworks provides the first part of the bridge between assessment and instruction. Beyond this common ground, we have an opportunity to support positive washback by creating tasks that do not only provide good assessment data but have sufficient richness to capture a broad construct and enough authenticity to engage students in relevant practice – with the caveat that "relevant practice" in the age of AI might require frequent construct revision and updating.

References

Meirav Arieli-Attali, Irvin R Katz, and Gabrielle Cayton-Hodges. 2023. The many faces of cognitive labs in educational measurement. *ASK: Research and Methods*, 32(1):91–120.

Beata Beigman Klebanov, Chaitanya Ramineni, David Kaufer, Paul Yeoh, and Suguru Ishizaki. 2019. Advancing the validity argument for standardized writing tests using quantitative rhetorical analysis. *Language Testing*, 36(1):125–144.

Jill Burstein, Steven Holtzman, Jennifer Lentini, Hillary Molloy, Jane Shore, Jonathan Steinberg, Meg Vezzu, and N Elliiot. 2014. Genre research and automated writing evaluation: Using the lens of genre to understand exposure and readiness in teaching and assessing school and workplace writing. In *Annual Meeting of the National Council on Measurement in Education (NCME), Philadelphia, PA*.

Arthur Hughes. 1989. *Testing for language teachers*. Cambridge university press.

Samuel Messick. 1996. Validity and washback in language testing. *Language testing*, 13(3):241–256.

Robert J Mislevy, Russell G Almond, and Janice F Lukas. 2003. A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1):i–29.

OECD. 2025. Empowering learners for the age of AI: An AI literacy framework for primary and secondary education (review draft). Technical report, OECD, Paris.

John Sabatini, Tenaha O'Reilly, Jonathan Weeks, and Zuowei Wang. 2020. Engineering a twenty-first century reading comprehension assessment system utilizing scenario-based assessment techniques. *International Journal of Testing*, 20(1):1–23.

UNESCO. 2024. AI competency framework for students. Technical report, UNESCO, France.