ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2024; 15(Special issue); 361-377

Generative AI in K12: Analytics from Early Adoption

Brad BOLENDER* Sara VİSPOEL** Geoff CONVERSE ***
Nick KOPROWİCZ**** Dan SONG***** Sarah OSARO******

Abstract

The integration of generative AI in K12 education and assessment development holds the potential to revolutionize instructional practices, assessment development, and content alignment. This article presents analytical insights and findings from early adoption studies utilizing AI-powered tools developed by Finetune—Generate and Catalog. Generate enhances the efficiency of assessment item development through customized natural language generation, producing high-quality, psychometrically valid items. Catalog intelligently tags and aligns educational content to various standards and frameworks, improving precision and reducing subjectivity. Through three comprehensive case studies, we explore the practical applications, benefits, and lessons learned from employing these AI systems in real-world educational settings. The purpose of this series of studies was to investigate the ways generative AI is currently being used in practical applications in test development to improve processes and products. The studies demonstrate significant reductions in time and costs, enhanced accuracy, and consistency in content alignment, and improved quality of educational and assessment materials. The findings underscore the substantial benefits and critical importance of customized AI systems, rigorous training for both AI models and users, and adopting appropriate evaluation metrics. With the use of off-the-shelf generative AI models expanding rapidly, it is vital that the effectiveness of AI systems that are highly customized through collaborations with measurement experts be presented, in order to maximize benefits and uphold the fundamental principles and best practices of test development.

Keywords: Generative AI, Assessment Development, Content Alignment, Educational Measurement

Introduction

Over the past decade natural language processing (NLP), machine learning, and artificial intelligence (AI) have grown steadily as tools to increase efficiency across industries, including education and assessment. The uses for these tools have been wide-ranging, from developing tests (Gierl & Haladyna, 2013) through automated scoring (Yan & Rupp, 2020) of short-answer constructed-responses (Burrows et al., 2015) and essays. These technologies have spread through the industry at a steady pace, as applications have been coded and refined (Attali & Burstein, 2006), validity arguments for their use have been developed and defended (Bennett & Zhang, 2015), and data analyses have been executed and presented that support their judicious implementation into testing organization's pipelines. However, in the past few years Generative AI has exploded onto the scene, and it has the power to dramatically alter educational instruction, test development, content analysis, and curriculum alignment methods.

Generative AI represents a transformative field in artificial intelligence focused on the autonomous generation of data. Central to this innovation are Large Language Models (LLMs), inherently complex neural networks optimized to understand, generate, and manipulate human language. These models,

* Principal Measurement Scientist, Finetune Learning, Iowa-City, US, bbolender@finetunelearning.com, ORCID ID: 0009-0000-1521-6136
** Chief Assessment and Learning Officer, Finetune Learning, Iowa-City, US, sara@finetunelearning.com, ORCID ID: 0009-0002-8159-8210
*** AI Scientist, Finetune Learning, Iowa-City, US, geoff@finetunelearning.com, ORCID ID: 0000-0001-8764-9950
**** Senior AI and Machine Learning Applied Scientist, Finetune Learning, Iowa-City, US, ORCID ID: 0009-0005-1548-6116
***** Graduate Research Assistant, University of Iowa, Iowa-City, US, dan-song@uiowa.edu, ORCID ID: 0000-0002-7466-6150
****** Research Assistant at University of Iowa, Iowa-City, US, ORCID ID: 0009-0008-9264-9072

To cite this article:
Bolender et al. (2024). Generative AI in K12: Analytics from Early Adoption. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special Issue), 361-377. https://doi.org/10.21031/epod.1539710

Received: 28.08.2024
Accepted: 25.11.2024

often based on transformer architectures, such as GPT (Generative Pre-trained Transformer) and its derivatives, leverage vast amounts of text data to learn linguistic patterns including syntax, semantics, and context. Despite their remarkable achievements, LLMs are not without challenges, such as considerable computational requirements and the propensity for generating contextually inappropriate or biased content, reflecting biases present in training data. Ongoing research addresses these issues through model distillation, ethical frameworks, and improved dataset curation, improving generative AI's alignment with human values, fairness, and inclusiveness.

With the rapid development of use-cases for Generative AI, concerns have also been raised about potential implications for education and assessment, including protecting the validity of assessments and avoiding the introduction of fairness and bias issues. As a result, several groups composed of researchers and testing organizations have convened and published guidelines for responsible use of the technology (Bolender et al., 2023; Hao et al., 2024; Ho, 2024). The focus of these guidelines has been to protect the validity and reliability of educational assessments, to ensure fair testing practices for test takers from all demographic backgrounds, and to specify methods for protecting the privacy and security of all test data including individually identifiable data from test takers. The guidelines also provide recommendations for ensuring transparency and accountability surrounding the use of Generative AI in the test development process, so stakeholders will be fully informed of the ways in which AI was used to aid in development of the test instruments, but also what measures were taken to protect validity.

Finetune has developed two AI-supported systems to assist with tasks related to K12 education and assessment, called Generate and Catalog. Due to the novelty of generative AI, not many studies exist of incorporation into real-world processes, especially those that focus on assessment. This paper will serve as an additional contribution to the Finetune research agenda (Khan et al., 2021a; Khan et al., 2021b). Since Finetune AI scientists and psychometricians were given early access to generative AI models as far back as 2018, multiple years of real-world data from use, as well as lessons learned on what generative AI does well, what it does not, and what is required for efficacy, have been integrated into these test development tools.

Finetune's applications and research contribution differs from most AI research using LLMs, due to a direct emphasis on customizing these AI tools specifically for assessment purposes. Incorporating best practices for AI and assessment and custom-building to customer requirements is significantly different than simply submitting a query to an LLM API. In this paper, we will share data from real users across multiple content areas and contexts to share critical information about how SMEs are using AI-assisted technologies and the degree to which best practices are being followed when using generative AI.

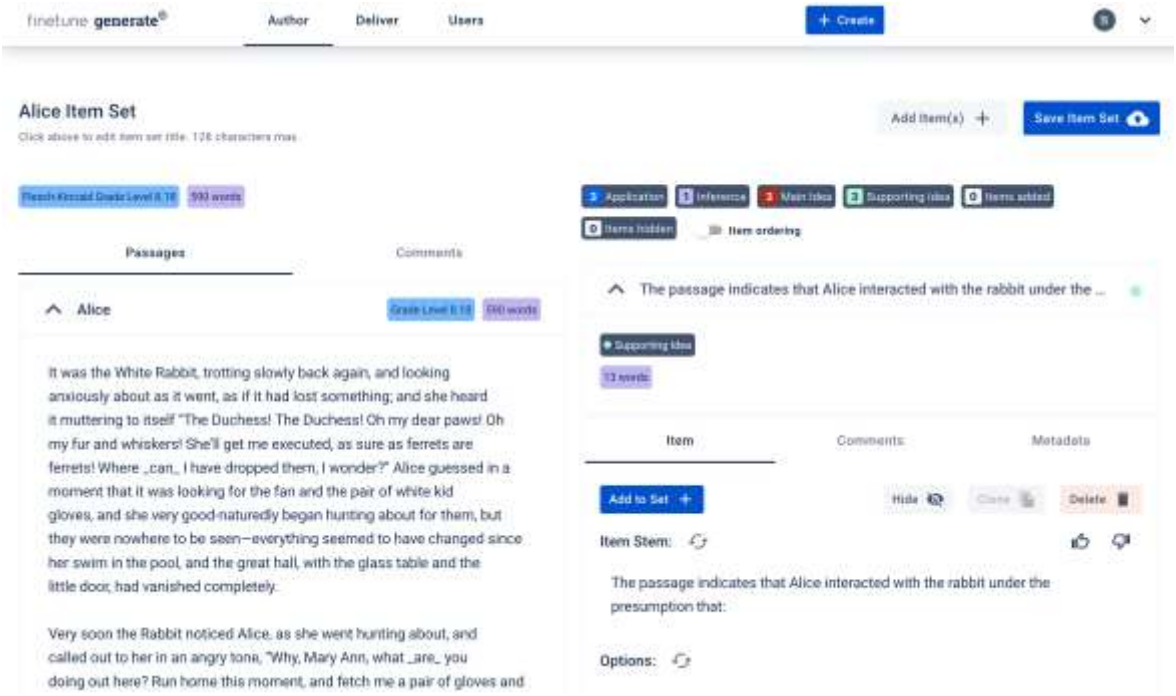
In this paper data will also be shared on using customized AI features in Finetune Catalog to automatically tag and align educational content such as items and learning materials, to standards, learning objectives, and cognitive complexity levels across content areas. This process involves harnessing both generative AI as well as more conventional machine learning techniques. Additionally, a natural language rationale can be generated explaining why an item is tagged with a particular standard. Outcome data will also be shared on how using this AI application can decrease the amount of subjectivity in tagging.

Generate

Finetune Generate (Khan et al., 2021b) is an AI-assisted system designed to enhance the efficiency and scalability of assessment content generation for educational purposes. The system leverages state-of-the-art natural language generation (NLG) methodologies in conjunction with the domain-specific expertise of assessment developers to facilitate the creation of a large volume of highly customized and psychometrically valid assessment items. Central to Finetune Generate is the Transformer architecture, which, through extensive pretraining on diverse text corpora, is adept at producing sophisticated, context-sensitive text that serves as a foundation for item generation.

Figure 1

The Finetune Generate AI-powered test development system.



When building a Generate model, items are intentionally tailored to meet requirements determined uniquely by test developer needs. Various sources of content are acquired from the user, including test blueprints, learning objectives, and cognitive complexity frameworks to build selectable sub-models that target constructs of interest in accordance with test specifications. Additionally, implicit reinforcement is provided relative to the user's style of item writing, as well as influence how an item aligns with a construct, by utilizing exemplar items provided by the user. With LLMs, there is often a trade-off between creativity—or randomness—and factual correctness—or determinism. To strike a balance, multiple sources of randomization are introduced to give a sense of variety in items, while still rooting the core content in the user-provided data. Additional features are integrated that connect AI-generated content to factual source material. This can be done in either a pre-generation manner, through Generate using a textbook passage as inspiration, or at post-generation where, given some generated content, a search is executed for a relevant textbook passage to serve as reference.

The Generate user-experience is customizable to different use-cases. At a baseline, selectable sub-models are provided that align to relevant constructs that drive the AI-generated content. Users may also input key words / key phrases to guide the AI, input a custom passage to use as a fact-base or inspiration, or input a reading comprehension stimulus passage to create item sets. Aside from content generation, AI solutions have been built into other post-hoc features, such as identifying a correct answer to an MCQ item, finding a citation for an item, and creating a rationale for correctness of the key option(s) and rationales for falsifiability of distractors.

This approach is unique in how AI scientists and measurement scientists work together to integrate the information in specifications and guidelines to develop a customized generative AI model. The first process of AI-enhanced item development involves partners providing details about their assessments including test purposes, test specifications, descriptions of constructs, test blueprints, item types, cognitive complexity requirements, references they want to include, and item writing guidelines. The resulting model is deployed in the Generate application so high-quality item drafts that meet requirements are produced.

A noteworthy unique feature about this generative AI application is that the capability of interacting with the customized AI model persists throughout the entire item development process. SMEs develop the items within the application both by editing stimuli, stems, and answer options directly, and also by

regenerating portions of the items with additional requests to the customized AI model. If users have reference materials that they want to use, we upload those materials into the application so SMEs can employ features like using the AI-assisted references to find citations that provide evidence for a key. In addition, content-, bias-, and committee-review processes can be completed within the application. Reviewers are able to share comments on items and can access the AI for assistance in generating possible fixes as they make subsequent revisions until the content is considered to be in its final state. At that point, the full set or a subset of items can be exported in multiple formats including plain-text, csv, and QTI-compliant XML.

Catalog

The second AI system involves applying generative AI to the task of associating learning materials to frameworks. In K-12 especially, in order for any learning material to be used flexibly and adaptively, associated metadata must be accurate, including tagging information to frameworks describing learning objectives, competencies, and cognitive complexity levels. For this task, a different application was developed: Finetune Catalog (Khan et al., 2021a). This system has been used to complete projects that entail tagging hundreds of items to larger projects tagging more than 50,000 items to various frameworks. Additionally, the Catalog engine has been used to provide AI-authored rationale statements for all tags assigned.

Figure 2

The Finetune Catalog AI tagging and alignment system.

Catalog employs a comprehensive AI-driven methodology to intelligently match educational content with relevant tags, serving as an expert across diverse educational domains. The process begins with the conversion of different types of educational materials into a format that is suitable for advanced analysis. An innovative framework is utilized to identify deep semantic relationships between content and tags, ensuring that diverse educational content and standards are aligned into a cohesive semantic space. This approach addresses a critical industry challenge and enhances the precision of our tagging process.

Emulating the expertise of subject matter experts, Catalog deploys a multi-level analytical approach tailored to interpret the educational intent behind each piece of content. The system navigates through multiple stages to determine the most relevant tags, incorporating mechanisms to validate its decisions at each step. Additionally, the process is customized based on user input and iterative refinement, allowing the AI system to adapt and align closely with the user's needs. This rigorous yet flexible methodology ensures that Catalog delivers highly accurate and contextually appropriate tagging.

Catalog uses a pipeline of varied techniques including embeddings with similarity measures, LLM prompting techniques such as few-shot prompting, chain-of-thought (CoT) prompting, self-reflection, and multi-turn interactions, along with hierarchical search to most effectively recognize correct associations between content.

The purpose of this article is to share three case studies illuminating how real users are currently interacting with AI-powered systems designed to streamline education and assessment processes including assessment item development, tagging of assessment items, and gap analysis to determine how well curricular materials cover learning objectives measured on summative assessments.

Case Study 1: Generate To Support Assessment Item Development

Research Questions

In what manner, and to what degree, do SMEs edit AI-generated item drafts before moving them to the next phase of item review?

A customized instance of Generate was built to assist with the development of a high school English Language Arts test. The Generate instance was a typical item set model that enables SMEs to paste a reading stimulus passage into a text box to use as the basis for a set of items testing various reading comprehension constructs, such as detail retrieval, inference making, overall text understanding, and understanding of language features. Additionally, the model was designed to support paired-text stimuli, so multiple texts could be entered, and synthesis items could be developed that require test takers to draw conclusions based on information in both texts.

The item writer guidelines document, the test blueprint, and example items were used by the Finetune AI team to develop a custom model that would generate many item drafts resembling the user's existing content, but that would not be based on any templates or copy any items already on their exams. A secure, unique instance of the Generate application was provided to the users – a team of 7 SMEs who would be using the model in their development cycle to help write roughly 60 items for an upcoming set of test forms. Training was provided to the user team to show them how to import stimulus passage, generate item drafts, revise and edit the items, and save them in Generate's interface to folders identifying the item sets as ready for review in the next phase of the project. Users were instructed that Generate is intended to be a human-in-the-loop AI-supported system for item development, so they as SMEs were expected to treat the outputs as item drafts, and their expertise was necessary to refine the items into their final forms.

Methods

After Generate training was done SMEs engaged in a typical development process, except rather than starting from a blank slate and having to come up with ideas for items, they used the customized Generate model to create item drafts. SMEs would then browse the item drafts and make decisions about which item drafts to work further on and which items to reject. SMEs would then consider the items in relation to the reading stimulus and make any minor edits necessary to the stem to ensure best measurement of the construct of interest. If needed, the key was edited to make sure there was one clear correct answer, and distractors were edited as needed to ensure they were clearly incorrect. SMEs interacted with the custom AI model throughout this procedure, by editing stimulus materials and using AI to regenerate stems, and by further editing the stem and regenerating answer options. Additionally, the custom model underwent continuous improvement through SMEs using a "thumbs-up" button to identify the best generated items, which were then integrated into examples for the model to consider in subsequent item generation, and a "thumbs-down" button to identify poor items to discourage the model from generating more like them. All of these efforts contribute to the custom generative model getting better and better as it is used by the SME in the process. When SMEs were satisfied with the items, then they saved them to a folder for further review.

To investigate the research question for this data set, the database underlying the application was accessed, as each item generated is assigned a unique ID that stays with the item through all versions as it is edited or revised during the test development process. The unique IDs for all 58 items that were saved for further review were queried, and the original AI-generated versions of the items were exported from the database so they could be compared to the versions that were saved in the review folder.

To compare the original AI-generated item drafts to the versions SMEs moved into the review folder, the Levenshtein edit distance (Levenshtein, 1966) was calculated between the two versions. Levenshtein edit distance (from here on simply referred to as edit distance) is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change the original version of an item into the saved version. This is an established method to make quantitative comparisons between language strings. Although character edits may not be especially intuitive in terms of imagining the extent to which an item has been changed, it may help to consider that some testing organizations use a concept called "standard word count" to refer to the length of content, which is total number of characters (including punctuation and spaces) divided by 6. So an edit distance of 48 could be imagined to correspond roughly to 8 words being changed.

After edit distances were recorded for both stems and options, results were summarized by grouping the items by the SME who worked with them, and the mean distances were calculated. This gives an idea of the typical amount individual SMEs edited the AI-generated items, both stems and options, prior to saving them for review, and it also gives an idea of the variation in editing behavior between SMEs on this project.

Additionally, it was determined how many out of the 58 saved items had 0 stem edits and 0 answer option edits, in order to understand how often portions of the AI-generated item drafts were satisfactory in their original states to move forward to the review phase of development.

Case Study 1: Results And Discussion

Table 1 shows mean edit distances between AI-generated item stems and review versions of item stems by SMEs who worked on them. Table 2 shows mean edit distances between AI-generated item options and review versions of item options by SMEs who worked on them. Table 3 shows the frequency of edit distance of 0 between AI-generated item stems and review versions of item stems overall, indicating items where the stems were satisfactory to be moved forward to the review process without additional editing. Table 4 shows the frequency of edit distance of 0 between AI-generated answer options and review versions of answer options overall, indicating items where the answer options were satisfactory to be moved forward to the review process without additional editing.

Table 1
Mean Edit Distance Between AI-Generated Stems and SME-Revised Stems

Subject Matter Expert	Mean Edit Distance Stem
SME_1	29.7
SME_2	32.0
SME_3	51.4
SME_4	59.5
SME_5	65.0
SME_6	6.1
SME_7	41.2

Table 2
Mean Edit Distance Between AI-Generated Answer Options and SME-Revised Answer Options

Subject Matter Expert	Mean Edit Distance Options
SME_1	54.2
SME_2	66.0
SME_3	90.3
SME_4	129.0
SME_5	75.0
SME_6	31.4
SME_7	167.0

Table 3
Edit Distance Frequency – Stem

Edit Distance	Freq.
0	5
>0	53

Table 4
Edit Distance Frequency – Options

Edit Distance	Freq.
0	9
>0	49

These results show that SMEs were active in working with the AI-generated item drafts. Although there was some variation between SMEs, most of them moved items into the review phase with at least 29 edit distance or more between drafts and review versions. The exception is SME 6 whose review items had a mean edit distance of 6.1 from the AI-generated items, which was quite a bit lower than the other SMEs. However, referring to Table 2 we see that the answer options for SME 6's review items had a mean edit distance of 31.4 from the AI-generated answer options. So it is possible that SME 6 was satisfied with the AI-generated stems, and they spent relatively more time working on refining the answer options.

On the whole, SMEs worked extensively with the answer options of AI-generated items, producing mean edit distances between 31.4 and 167. A future line of research could involve investigating the amount and type of editing that was done to keys, in order to enhance construct measurement or to make correctness certain, versus editing done to distractors, which may have been done to make items easier or harder or to introduce common errors and misconceptions. Results from that research could be used to inform further advancement in AI model and system development in terms of continuing to integrate best measurement practices into AI systems.

Out of the 58 items, there were 5 instances in which no edits were made to AI-generated stems, and 9 instances in which no edits were made to AI-generated answer options. Again, this is evidence that SMEs were generally active in working with the draft materials, but that in some cases the AI-generated materials were of sufficient quality to move them forward to the peer review phase of development.

Case Study 2: Catalog To Support Tagging

Another use case for customized AI systems specifically for pedagogical and assessment-related insights to execute alignments. That is, using customized AI systems to align assessment materials such as items, instructional materials, and texts from across all content areas to frameworks such as competencies, CEFR, national standards, learning objectives, Bloom's taxonomy, and assessment blueprints. This is particularly exciting due to current methods being used in the field to conduct alignments and the well-known problems that each brings. At present, alignments of educational material are typically done using one of four methods: manually, by keyword, by semantic similarity, or by crosswalk.

The manual process involves having subject matter experts (SMEs) read the original materials and make a personal decision about which aspect(s) or standard(s) of a framework the item content is aligned with, and then manually match these up and tag them as being associated. Unfortunately, when using this approach, the alignment results from individuals inevitably vary even though the materials and framework do not. Whether due to differences of opinion in expert judgment, inconsistent interpretation of a framework, or lack of attention, alignments done by multiple or even single SMEs do not tend to provide repeatable results.

Another conventional approach for executing alignments is to use Natural Language Processing (NLP) technology and applying keyword searches. This method involves identifying specific words to search for in the content, and establishing rules for assigning specific tags based on the search results. Using a keyword approach typically results in overly focusing on the content or topic and fails to consider other critical process or behavioral aspects of an item that elicit what an examinee should know and be able to do. The keyword approach also requires a considerable upfront investment of SME resources to identify and cross-check potential keywords that are representative of every standard without triggering too many false positive matches with the wrong standards.

Another alignment approach involves a computational linguistic strategy of calculating the semantic similarity between the object being aligned (assessment item, learning content) and framework elements (state standards, learning objectives, etc.). Using text embeddings, an individual piece of content such as an assessment item can be compared to every framework element such as learning standards, and values can be computed that represent the similarity in meaning between all pairs. Afterward, the standard with the top similarity value can be assigned, or a SME could select the best standard from the top several options.

For some domains semantic similarity may be a useful approach, specifically when items and standards are expected to be highly semantically similar and similar in content, such as science standards that focus on core science ideas and recall-type items (e.g., both refer to "phases of the moon"). However, this method does not work for other domains or situations where standards and items are not expected to be semantically similar. For example, consider a reading standard that says, "Read closely to determine what the text says explicitly/implicitly and make logical inferences from it," and an item that asks, "Why did the narrator choose a particular course of action?" Semantic similarity is not a strong approach when working with rich assessment and learning tasks that go beyond knowledge of content topics.

A fourth alignment/tagging approach involves using crosswalks of frameworks (e.g., state standards, test blueprints, learning objectives, cognitive complexities). The crosswalk approach focuses on relating all elements within one framework to a different framework, then using the transitive property to infer resultant mappings. This process only involves the frameworks and does not directly involve the material being aligned. Step one is to associate each element in Framework 1 to the most similar element that can be found in Framework 2 (e.g., a state standard in one state associated as nearly the same as a similar state standard in a different state). Then, the reasoning is that any assessment task or lesson that had been associated with that element/statement in Framework 1 should now be considered aligned with the statement in Framework 2 that had been identified as aligning with Framework 1.

Assessment and educational experts use the crosswalk approach in order to try to save time and resources. Without carefully considering the assessment task, the crosswalk approach enables alignment based solely on SME ideas on relationships among the statements of the frameworks. The crosswalk approach is particularly notorious due largely to the limitation of not using the primary source of text of the materials that are being aligned. Without looking at the actual tasks and lessons, subtleties for why they had been aligned to a particular element of a framework may be missed. Additionally, since

frameworks rarely align directly or use the same language in the same way, associations must be interpolated at best, rather than interpreted from direct evidence. Another problem with the crosswalk approach is that errors are cascaded and proliferated throughout the project. If one element is not truly similar to another it had been associated with, then everything coming in and out of that relationship contributes to errors.

Given the heavy lift required, sometimes this alignment work is outsourced to an external group who uses one or more of these methods. Inevitably, external SMEs lack the insight of internal SMEs about the materials themselves, which hinders the ability to align accurately. Also, the external group may or may not have the requisite experience with the desired framework to infer necessary interpretations of how the framework is intended to be operationalized in the relevant assessment or learning context. Ironically, outsourced alignments must be audited and reviewed by the very SMEs whose time was meant to be protected by outsourcing the work in the first place.

Regardless of method, any error or inconsistency in tagging introduces significant consequential error when assembling an assessment or when providing remediation recommendations for improvement. Tagging tasks must be as error-free as possible. In addition, given how quickly things are changing in educational settings educators and assessment developers should be assuring that all tagging is consistent and up to date thus enabling instruction and assessment to be more consistent and accurate across all materials.

As discussed previously, current alignment strategies (manual, keywords, semantic similarity, and crosswalks) are fraught with known problems. This specific study features multiple aspects of using customized AI to perform alignments.

Research Questions

First is answering the question: how much time is saved. For that inference, we will look at a use case of a K-12 educational service center in Texas that had to maintain and align an item bank comprising 90,000 test questions to state standards amidst evolving educational trends and the introduction of technology-enhanced items (TEIs). The next question is, can a customized AI model align items to multiple frameworks and provide evidence-based justifications all at one time. Finally, the third question is answering the most common question that is asked about the customized AI tagging technology: namely, how good is it? In this case, hundreds of assessment tasks had each been aligned with multiple frameworks so preexisting tags for each item and each framework were available for comparison to assess quality.

Methods

Each of these studies involved developing a customized AI model to align materials and provide evidence-based rationales justifying the application of each tag. The first use case involved 90,000 items. The model was designed to align items to Webb's DOK framework which gives inferences about the cognitive complexity of each assessment task. The next case study involved developing a customized AI model to be able to align and provide evidence-based rationales for nearly 600 assessment items to six different frameworks simultaneously.

Each framework in this study focused on a different aspect of the construct. One framework consisted of task descriptions that were highly technical in nature, focusing primarily on the content of the assessment task. Another framework focused on competencies that could be measured by executing the task. A framework focused on inferences that could be made about the examinee's social-emotional or foundational/durable skills. Another framework required inferences about the examinee's proficiency with respect to different process skills. The final framework required inferences about the level of cognitive complexity executed by the examinee during the task. SMEs were then asked to review the tags and provide feedback about accuracy.

Results And Discussion

Regarding the first case study of 90,000 items, results included an 88% reduction in item alignment time, and 85% cost savings over manual methods. Additional quality metrics included a 96% accuracy rate in content alignment. In the second use case, 600 items were aligned successfully to 6 frameworks

resulting in roughly 3600 alignment decisions each with a customized evidence-based justification for each tagging decision. Regarding the effectiveness, the assessment items and their tags were provided back to SMEs along with previous tags. Initial agreement of the AI-assigned tags compared to the previous tags can be seen in Table 5.

Table 5

Initial Agreement Between AI-assigned and SME-assigned Tags

Framework	Initial Agreement
Technical content	41.4%
Competencies	64.0%
Discipline	59.1%
Foundational Skills	60.8%
Process Skills	65.6%
Thinking Skills	72.6%

SMEs were then told to review the quality of the newly assigned AI tags and provide any corrections. Table 6 shows the proportion of mismatches where the SME agreed with the AI-assigned tag over the original tag assigned by an SME.

Table 6

Mismatched Tags: SME Agreement With AI-assigned Tags Over Original SME-Assigned Tags

Framework	Agreement
Technical content	88.1%
Competencies	75.1%
Discipline	74.1%
Foundational Skills	71.8%
Process Skills	64.3%
Thinking Skills	49.3%

Table 7 shows the level of SME agreement with the AI-assigned tags. Note that this is before this feedback was taken into account and used to recalibrate the custom model. Each SME provided a rationale for the disagreement.

Table 7

SME Agreement With AI-assigned Tags After Updates but Before Model Recalibration

Framework	Agreement
Technical content	93.0%
Competencies	89.8%
Discipline	89.8%
Foundational Skills	89.8%
Process Skills	87.7%
Thinking Skills	86.1%

These results demonstrate SME agreement across frameworks ranging from 86% to 93% agreement before calibration. That is, the feedback provided for those areas of nonagreement was used to recalibrate the AI model, therefore improving tagging performance and increasing agreement even further in the next round.

All of these results suggest that using customized AI systems could significantly decrease time for alignment tasks, increase the consistency of tags and evidence supporting each tag, as well as demonstrating very high levels of accuracy according to SMEs across content areas and frameworks.

Case Study 3: Catalog For Gap Analysis

Another disruptive use case features applying customized AI systems to execute gap analyses about how well current items and assessment materials cover and align with test blueprints and requirements. Typically, the same basic manual and keyword approach described above are also used to conduct gap analyses. Once again, executing this manually takes a great deal of time and ends up being inconsistent due to the role of opinions in alignment decisions. Additionally, executing by keyword often results in an overreliance on content topic and subsequent undervaluing of the skills described by requirements.

Importantly, given the inordinate amount of time and resources the current typical process takes, stakeholders must be judicious with how often such an analysis can be performed. Using customized AI for this purpose is not only faster and more repeatable, but analyses can also be run to provide additional levels of insight. For example, in a typical gap analysis, an item bank can be queried for coverage. However, queries are lacking sufficient insight into how a particular item truly measures a particular construct. Outputs of analyses will be shared regarding whether the material is considered to be a direct match or less-direct match to the framework elements. Additionally, if something is missing, customized AI can provide specific insight into what is covered and what is not covered.

Typically, a gap analysis is used as a summative activity. The result is used to evaluate the final item pool against the framework or test blueprint after assessment development is largely or entirely complete. The goal of the evaluation is, as with most summative assessments, to get a passing grade—i.e., to have the assessment judged as 'covered' with the standards it is intended to assess and to serve as one piece of evidence in a validity argument for use in decision-making.

In this approach, robust information about the alignment will be provided to the SME reviewers and provided much earlier in the assessment development process. For example, an initial set of items may be drafted to cover only one aspect of the framework. That set can be submitted to the evaluation system immediately, to see whether the system agrees with the coverage estimation, whether it uncovers other aspects of the framework also assessed in the set, and whether it adequately covers the selected part of a given learning standard. Here, the alignment system can provide information about what learning standards are covered—and what parts are not—for each item to be routed back to the assessment developer. The AI system can provide evidence in a narrative format, explaining why the item was associated with a particular framework element. Feedback on this analytic evidence set from the SMEs may be given back to the AI scientists and psychometricians so that updates and refinements to the AI model can be made, making future iterations of the evaluation steadily more precise. As more items are added to the set and more framework coverage is assumed, they can be rapidly verified by the alignment system.

Using AI to execute these analyses can be repeated as many times as desired with consistent results, where repetition with subject-matter experts is time-consuming and costly, in addition to the likelihood of disagreement between SMEs. The ability to check coverage repeatedly, rapidly, accurately, and easily will ensure that the final product is fully aligned to the relevant learning objectives, with no gaps or weak points of coverage. Routinely reviewing accuracy and breadth of coverage should improve the assessment development process, while also making it faster and more efficient. This partnership optimizes the combination of strengths from human expertise with automated system consistency, speed, and accuracy.

This third and final use case comes from the need in primary through upper secondary educational settings to understand systematic and rigorous coverage of educational concepts vertically as well as horizontally. Documenting where and when prerequisite skills are taught offers insights and should provide scaffolding for learner pathways. Currently, this is typically approached by coordinating teacher panels and collecting their professional opinions about scope and sequence. The challenge comes from how much time these analyses take and, again, how much opinion may vary among teachers. Additionally, the world is changing so fast, desired skills and learning standards are updated frequently, and educators must keep pace in order to make sure students are well prepared for college and beyond.

Unfortunately, any significant change in curriculum immediately evokes the need for a new analysis. Given how long and how much time scope and sequence documents take to develop, teachers are

restricted from making changes at the risk of introducing problems into educational progression. Instead, examination of results from applying AI customized for assessment insights related to multiple scope and sequence situations, and it will discuss what insights were able to be uncovered with respect to covering material in optimal pathways, identifying gaps in instruction, and identifying whether prerequisite skills were indeed covered adequately. This customized approach goes beyond text embeddings to employ the latest techniques to encode content in a context-attentive fashion, thus enabling valid and repeatable capture of deep conceptual and contextual relations in item content as well as in the educational/workforce frameworks, facilitating alignment of those materials to each other.

This study features a use case of applying customized AI technology in order to obtain K-12-specific inferences with respect to identifying potential instructional issues and opportunities to improve student performance. The study involves one large U.S. public K-12 school district made up of over 40,000 students and over 2000 teachers, distributed across more than 30 different schools.

School district leadership had analyzed student test data across grades and subjects from previous school years. The content area identified as having the biggest deviation from desired performance was in a specific Algebra course taught at multiple schools within the district. One restriction of the study was that student performance data would not be available. Therefore, the decision was made to execute AI-assisted analysis of the instructional materials with a particular focus on how well these materials covered the requisite knowledge, skills and abilities described in detail in the benchmark statements and descriptors found in the state standards.

Materials eligible for analysis included primarily instructional artifacts. These materials included the state specific scope and sequence documentation, state standards and benchmarks, assessment blueprints and unit tests, lessons and student-facing instructional materials. Benchmark level descriptions and evidence found in the state standards totaled around 70 different statements. The student-facing instructional materials consisted of 74 lessons. The assessment blueprints and sample assessments were at the lesson level for the Algebra course.

Research Questions

The research questions guiding this study were, could we use customized AI tools to help make evidence-based inferences on how well the current instructional materials covered desired topic areas? And could any instructional gaps be identified that might potentially account for lower student performance?

Methods

The first step of the study involved developing the customized AI system (Catalog) such that it would provide multiple insights. The most critical inferences were having the customized model provide primary tags relating to the benchmark level of specificity and secondary tags when appropriate, along with evidence-based rationales for those tagging decisions. The customized AI model also needed to provide prerequisite skills in terms of the benchmarks from lower grades. Having prerequisites identified for the instruction enables educational leaders to see any places in the current curriculum that students might struggle if they are not at sufficient proficiency at the start of the instruction. In those places, adding specific remediation strategies at the start of these lessons might increase the student access to the current instruction and increase engagement.

Once the AI model was developed, next steps included executing multiple analyses of the various instructional materials. The first analysis was a unit-level analysis of over 70 instructional units of student-facing instruction, problems, activities, and practice problems. All of the various student-facing materials and teacher plans were provided to the customized AI model. The model analyzed each complete unit separately and provided primary and secondary tags at the benchmark level along with evidence-based justifications for each of the tagging decisions. Each instructional unit had been previously tagged to benchmark level by an unidentified source, but that information was not used in the analysis. In addition, prerequisite skills, as articulated by the benchmarks from previous grades, were provided for each lesson. The rationale for having this information again is that if students were lacking

sufficient proficiency in prerequisite skills and knowledge, they might not be able to fully engage and benefit from the instruction.

The second analysis was at the sub-unit level of the curricular materials. That is, each of the Algebra units were broken into 15-20 subunits, totaling 874 subunits. In this analysis, the customized AI provided primary and secondary tags at the benchmark level of specificity for each of the subunits. Note that this more specific level of analysis had not been completed prior to this analysis. Tagging at the sub-unit level provides a more detailed view of the instruction and coverage within each unit, revealing benchmark coverage learning during the days and hours within the entire unit.

Final steps of the study included executing multiple comparisons of information obtained by the AI-enabled analyses to current documentation of instructional and assessment coverage.

Results

Results from tagging at the unit level are presented first. Primary and secondary tags and their evidence-based justifications provided along with prerequisite skills and knowledge required for each of the lessons. The primary and secondary tags from the customized AI tagging were then compared to the preexisting tags provided for the lessons from a different source. Results showed 81% agreement between the AI-tagging and the preexisting tags for the primary benchmarks covered by the course. The remaining 19% of the primary tags were different. Some of the AI-identified primary benchmarks identified were quite similar to the previously identified tags, however, others were quite different. The evidence-based justifications made the task of validating the AI-provided tags easy and direct. Unfortunately, as is commonly the case, the preexisting tags were not accompanied by any information about how they were decided or justification for the tags.

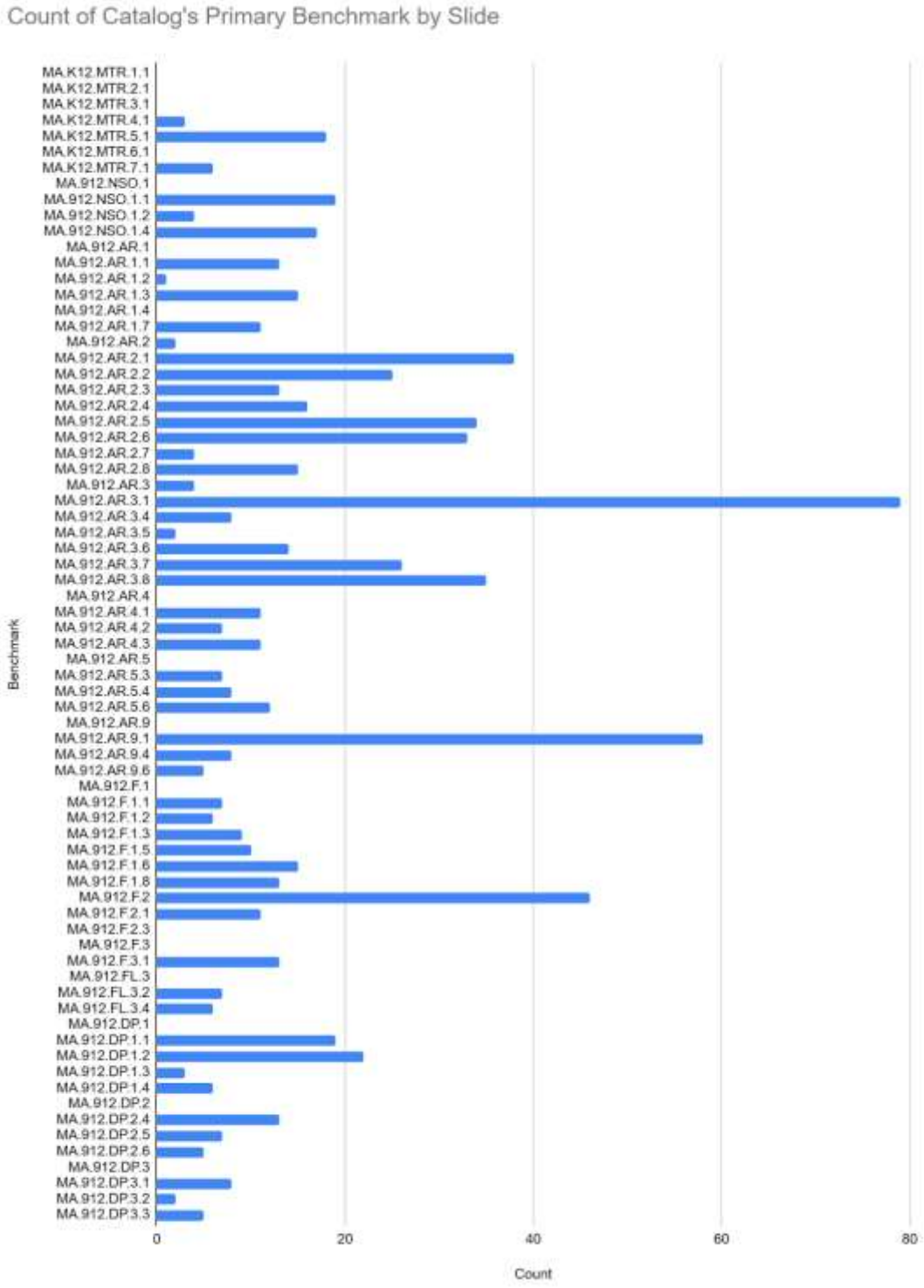
Lessons that had a mismatch of AI-assigned tags to preexisting tags were pointed out for educational leaders to consider. For example, one of those mismatches revealed that a multiple-day instructional unit was actually designed to cover the content for a particular kind of function. In fact, the standards and benchmarks required for this particular course did not require primary coverage of that topic. In this case, those multiple days might be better spent not covering that lesson, but instead covering something else more important.

Second, the distributions of primary benchmark coverage at the sub-unit level were analyzed. These results are shared in Figure 1. This analysis provided insight into the hourly and daily coverage within the instructional subunits so that the district could easily make a judgment about whether all benchmarks were being sufficiently covered. The prerequisite skills for each subunit of instruction were also listed in case known issues in previous proficiency could be responsible for impeding efficacy of the instruction.

Some of the most compelling results were the comparisons of the primary content coverage (benchmark tags) as identified by AI to the assessment blueprint for particular instructional lessons. The analysis for the first unit revealed that 33.3% of the blueprint was not covered by primary instruction according to the AI. The analysis for the second unit assessment revealed 50% of the assessment were benchmarks that did not receive primary unit instruction according to the AI. When shared with district leaders, SMEs confirmed that this analysis actually confirmed their suspicion about the instructional misalignment with the assessment, but they had previously lacked the data to support it.

Overall, the customized AI model was able to provide consistent, evidence-based alignments for multiple levels of instruction that districts and teachers lack time and resources to complete manually. This study demonstrated the power of being able to perform multiple levels of analyses efficiently and accurately in to be able to answer various questions about instructional coverage.

Figure 3
Count of Catalog's primary benchmark by slide.



Discussion

Generative AI has great potential in the K-12 space, including for instruction and assessment. The novel, real-world applications presented in these studies have demonstrated great promise as well as shed light on some lessons learned when working with generative AI.

First, all the real-world studies in these studies have used generative AI systems customized for applications in education and assessment by teams of AI scientists, psychometricians, and experts in measurement and education. As mentioned previously, using customized AI systems are not the same as simply prompting a Large Language Model (LLM) that lacks additional information, training, and expertise in assessment and instruction. Therefore, the gains and efficiencies of the customized systems are not expected to be reproducible using a general LLM, unspecific to a particular domain. The generative system is most effective at outputting high quality material when examples, descriptions, knowledge, and elaborations from the domain experts are integrated into the pipeline.

A second lesson learned is the importance of training, both for the AI model and for the SMEs using the customized system. If a customized model is being developed, ensuring high-quality exemplars are featured as the majority of the training set can improve the quality of initial drafts of items coming from the customized model. Many times, test developers will tend to want to use higher numbers of examples of items that they have available to customize their model rather than choosing fewer, but higher-quality questions. The problem with large numbers of items is that the likelihood is greater for those items to be ones that the organization does not deem as high quality. When a model is trained with lower quality inputs, then the greater the likelihood for lower quality drafts being produced by the customized model. The higher quality the training set, the better the customized model will be. Just as important as the training set is the SMEs chosen to interact with the customized system. The best SMEs for working with the AI system are people that are eager to use the technology and have a positive attitude about the potential of doing things slightly differently. They should be the kinds of SMEs that are motivated to use all of the features actively. Taking actions like regenerating and editing stems and options that are not ideal gives very helpful and actionable feedback so that the model improves much more quickly and efficiently.

A third lesson learned is to stress that this is a new way of doing things so therefore the outputs of interest are slightly different. For example, when talking about developing a customized AI model for item generation, the output of interest is not “number of items” as much as a customized AI model that is able to produce a high-quality draft at any level of specificity, cognitive level, of any kind across the entire test blueprint. Similarly, when considering AI-enabled alignment, the output is not just a single alignment as much as a customized model specific to the framework and materials provided that can be validated, calibrated and reused producing extremely reliable and consistent results.

A fourth lesson learned is to be careful when choosing metrics to evaluate the quality of the AI tools. As we have seen, when it comes to alignment, mismatches should be investigated fully and not just presupposed to be due to either the AI or the SME being incorrect. Many organizations will want teachers and SMEs to be the arbiters of quality. Many SMEs have developed their own heuristics and notes to save time when aligning materials. Unfortunately, many times those heuristics may not work as well as taking a fresh look at each item and each framework element as the AI is doing. Similarly, the AI model should not be over- or under-rated. The AI model needs to be checked to be sure inferences are being made appropriately according to evidence.

Conclusion

The research undertaken on the application of generative AI within the K12 educational setting highlights significant potential for these technologies to impact and enhance various educational processes. From assessment item development utilizing Finetune Generate to the intelligent tagging and alignment of content with Finetune Catalog, our findings present robust evidence supporting the efficiency and efficacy of customized AI systems. The case studies underscore the tangible benefits these technologies can offer, such as substantial reductions in time and costs, marked increases in consistency and accuracy, and improvements in the quality of educational content and its alignment with standards and learning objectives.

Key lessons have emerged from these studies, the foremost being the irreplaceable role of customization in achieving high-quality output from AI systems. Off-the-shelf LLMs, while powerful, do not match the efficacy of models tailored specifically to the nuances of educational content and assessment requirements. This customization involves crucial input from domain experts, high-quality training data, and continuous interaction and feedback from SMEs to refine model performance. This underscores a new paradigm in AI application where the fusion of advanced computational techniques and human expertise yields superior results. Therefore, researchers and practitioners, alike, should not settle for using off-the-shelf, general models to generate draft-quality assessment content for further refinement by SMEs, as the results will be lacking in terms of how well generated content adheres to specific style and structure specifications, and also how well it upholds the fundamental principles of assessment. The best results will come from collaborative system-building done through cooperation between content experts, psychometricians, measurement scientists, and AI scientists.

Furthermore, our research highlights the importance of rigorous training for both AI models and human users. The effectiveness of AI systems in generating and refining educational content greatly depends on the quality of the training data fed into the models and the proficiency of SMEs in leveraging these systems. The active engagement of motivated and knowledgeable SMEs in using AI tools ensures that the outputs are continuously improved, and the AI systems evolve to meet the specific needs and standards of educational contexts. This collaborative approach not only enhances the AI's performance but also fosters greater acceptance and utilization of technology among educators. It should not be expected that SMEs who are simply given access to AI-powered tools will figure out the best way to accomplish efficiency and quality gains. Specific training on how to use the AI-powered systems is a must, and providing time for learning the systems is critical.

Finally, it is critical to adopt appropriate metrics for evaluating AI systems. Traditional measures may fall short in capturing the nuanced improvements AI can bring to educational processes. For instance, merely calculating metrics like agreement percentage between the AI and existing tags misses the opportunity to consider evidence for a fresh approach to tagging decisions. Researchers and practitioners should strongly consider that specific instances may occur where AI-assigned tags could be as accurate, or more accurate, than SME-assigned tags—whether due to real advantages of AI analysis of associations between content and tags, or due to possibilities such as fatigue on the part of human SMEs. A better performance measure than agreement with existing tags may be SME agreement with AI-generated rationales explaining why certain tags were assigned. This holistic approach to evaluation will help stakeholders better understand and appreciate the profound impacts of generative AI in education, ultimately driving forward its integration and advancement.

An additional note is warranted for researchers and practitioners who would use generative AI for assessment and education applications: AI models are continuously and rapidly evolving, as well as learning new information, which means previously generated assessments and constructs could be called into question as new versions of models are released and developed (Kaldaras et al., 2024). This is a suggestion against using AI generated assertions and materials directly in production-level applications, and a suggestion for continuing to have SMEs retain final control, to refine and smooth over implicit assertions and choices made by AI models that could be inconstant as new versions are rolled out.

As we look ahead, the future of AI-assisted test development and AI-assisted tagging work is bright. With continuous advancements in AI, particularly in the development of even more sophisticated and contextually aware models, we can anticipate continued enhancements in the precision, speed, and creativity of educational content creation. The capacity for seamless, real-time alignment of educational materials to evolving standards and the personalized adaptation of learning resources to meet individual student needs will revolutionize instructional practices. Our initial studies are promising, and we envision a future where educators are empowered with AI tools that not only relieve them of repetitive tasks but also open up new horizons for innovative teaching strategies, enabling a richer, more responsive educational environment for all learners.

Declarations

Gen-AI use : The authors of this article declare (Declaration Form #: 212241835) that Large Language Models (LLMs), were used for writing and editing text in up to 10% of the article. They further affirm that all content generated by GenAI has been carefully reviewed, and they assume full responsibility for its inclusion.

Conflict of Interest: None

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Bennett, R., & Zhang, M. (2015). Validity and Automated Scoring. *Technology and Testing*, 142–173. Routledge. https://doi.org/10.4324/9781315871493-8

Bolender, B., Foster, C., & Vispoel, S. (2023). The Criticality of Implementing Principled Design When Using AI Technologies in Test Development. *Language Assessment Quarterly*, *20*(4-5), 512–519. https://doi.org/10.1080/15434303.2023.2288266

Burrows, S., Gurevych, I., & Stein, B. (2015). The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Gierl, M. J., & Haladyna, T. M. (2013). *Automatic Item Generation: Theory and Practice*. https://doi.org/10.4324/9780203803912

Hao, J., Alina, Yaneva, V., Lottridge, S., Matthias von Davier, & Harris, D. J. (2024). Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI. *Educational Measurement*. https://doi.org/10.1111/emip.12602

Ho, A. D. (2024). Artificial Intelligence and Educational Measurement: Opportunities and Threats. *Journal of Educational and Behavioral Statistics*, *0*(0). https://doi.org/10.3102/10769986241248771

Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing Valid assessments in the Era of Generative Artificial Intelligence. Frontiers in Education (Vol. 9, p. 1399377). https://doi.org/10.3389/feduc.2024.1399377

Khan, S., Rosaler, J., Hamer, J., & Almeida, T. (2021a). Catalog: An educational content tagging system. In Hsiao, I., Sahebi, S., Bouchet, F., Vie, J. (Eds.), *Proceedings of the International Conference on Educational Data Mining*, 736-740. International Educational Data Mining Society.

Khan, S., Hamer, J., & Almeida, T. (2021b). Generate: A NLG system for educational content creation. In Hsiao, I., Sahebi, S., Bouchet, F., Vie, J. (Eds.), *Proceedings of the International Conference on Educational Data Mining*, 741-744. International Educational Data Mining Society.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*(8), 707-710.

Yan, D., Rupp, A. A., & Foltz, P. W. (2020). *Handbook of Automated Scoring*. CRC Press.