IRSTI 14.01.85 https://doi.org/10.63597/UTO3105-4161.2026.6.2.006

N. Dieteren
Dutchtestologist, The Netherlands
e-mail: dutchtestologist@outlook.com

ORCID 0009-0009-7887-3745

TESTOLOGY AND TECHNOLOGY: HOW THE HUMAN FACTOR CAN LEVERAGE AND ENHANCE THE USE OF AI IN MAKING GOOD TEST ITEMS ON HIGHER ORDER THINKING SKILLS

Technology is becoming more dominant and leading innovations in educational assessment. The introduction of the use of Artificial Intelligence (AI), more specifically the application of General purpose engines for writing test items, is considered by many experts as a paradigm shift in assessments. The use of AI in general has clear advantages when creating test items automatically and more efficient. The way in which AI-powered algorithms analyze existing item banks with test questions is generating items in a much higher speed and efficiency. But, the use of General purpose AI models also has some serious drawbacks, in terms of resources, costs, reliability, ethical concerns and legal issues. Customisable AI models might offer an alternative that is worth serious consideration for development of educational assessment tools. Whatever AI model used, the human factor cannot be missed. The use of AI in item generation will change the role of the test expert, but a human check will still be needed. In many large scale assessments that are used for making valid judgments to inform decision making, it is essential to find a good combination of innovative item development and responsible human decision making. In complex cognitive domains, assessing higher order thinking skills, the human expert review is particularly important.

Keywords: Testology, Automated item generation, Generative AI models, Large Language Models.

Introduction

When it comes to modern methods for test development, the role of technology is becoming more dominant and leading innovations in educational assessment. The introduction of the use of Artificial Intelligence (AI), more specifically the application of general-purpose AI engines like ChatGPT or DeepSeek for writing test questions, is considered by many experts as a paradigm shift in assessments. “In fact, it’s about a total makeover of our approach to assessing from the ground up” [1].

Materials and research мethods

This article will give a selective, not exhaustive but representative, overview of recent developments in the use of AI for generating good quality test items, specifically in a context of using such items in high-stakes assessments that put high demands on validity, reliability, security, generalizability and justification of test results and decisions taken for individual test takers. The article is not based on any empirical research or originated data, but aims to serve as a guide and support for practitioners who work in the field of high-stakes educational measurement.

Results and Discussion

1. How the use of AI will transform assessment tools

All professional companies that provide educational assessment tools, services and consultancy, are developing new strategies, tools and expertise that can support customers in the field of educational assessment with making use of AI in a way that balances the technological advantages of Artificial Intelligence with the unique expert skills of item writers, test developers and educators. Most companies offer computerized, digital or online authoring systems for developing test items and building item banks. Examples are AQA [2-7]. All companies are working towards a transformation and expansion of such Item authoring and Item banking systems with modules, strategies and capacity building services that include the use of AI.

1.1 Terms and definitions: "what's in a name?"

Allowing myself to make a small side-step to the world of Shakespeare and his famous play 'Romeo and Juliet' , I want to explain shortly what some quite similar-looking acronyms and terms mean when talking about modern technology that can assist item writing in educational assessment.

First of all there is the difference between *Automated Item Generation*[1] (AIG) and *Generative AI models* (GAI). See how the confusion starts when looking to the acronyms.

Automated Item Generation can be seen as the overarching central concept, in contrast with a traditional human-based writing of test items one-by-one [8, 9]. Before the introduction of AI, automated item generation was enabled by replacing numbers or words in predefined models or templates, aiming to assess the same underlying construct [10]. These models were based on the use of previously developed and assessed test items. This AIG approach showed some disadvantages, like a high demand on programming expertise of the test developers and a risk of getting 'more of the same' in tests, what can reduce the reliability of test results once such tests become more predictable and therefore 'trainable'.

Generative AI models work differently when it comes to generating (large amounts of) test items, as they are based on the use of Large Language Models and machine learning, what enables a test developer to produce very large numbers of test items in an iterative process that demands little to none IT-expertise. In short, we can state that the use of AI enables a very scalable AIG [10].

Another important distinction we have to make is between *General* AI models and *Generative* AI models [11]. General models analyse existing data with a certain purpose, like classifying, selecting, predicting. Take for example facial recognition as one of the security measures in high-security contexts, like exam centers or airports. The objective of *General* AI models is to make or support decisions based on data the model has been trained on. This is sometimes called *'Discriminative AI'* [12]. The objective of *Generative* AI models is to create entirely new content, based on patterns learned from training data. One of the best known examples is ChatGPT. So the goals and outputs of both models are different. Generative AI models require much larger datasets to understand patterns for generation, whereas General AI models can work with limited and focused data sets, like a set of features of human faces can serve face recognition purposes. Also, Generative AI models are less transparent in their 'decision-making' processes and can be seen as 'black boxes', whereas General AI models are often more interpretable.

General AI (models) should not be confused with Artificial *General* Intelligence (AGI). Once more the use of acronyms can be confusing and conceal the real meaning of terms in writing, reading and discussing. AGI is an overarching concept that holds the characteristics of both General AI and Generative AI.

In the context of this article, focusing on the development of good test items with the help of AI, I will further explore the use of *Generative* AI models. It is important to make a distinction between General purpose generative models and Customisable (or 'customised') generative models.

2.1 General purpose AI-models

In most cases the integration of artificial intelligence in test development is using so-called *'General-purpose AI models'.* These models – also called generic models - are AI systems that can handle many different tasks. Examples of such models are ChatGPT, DeepSeek, Copilot and Gemini [12].

These generic AI models are based on the use of large-language models (LLMs) and have some clear advantages, also in the field of educational assessment [14]:

1. High level of language understanding, because trained to handle queries in human languages.

[1] Also sometimes called Automatic IG; in this article I use the term Automated IG

2. Pre-trained on vast text corpora, what makes them versatile in answering questions on almost every topic.
3. Very cost-effective once trained, as they can be used in many different contexts.
4. Automation on the basis of huge volumes of data that can be processed.

However, these General purpose AI / LLM models have some serious drawbacks [14]:

1. Speed and cost: training and using these generic models requires vast data processing, demanding much time, resources and money. The environmental impact can be too high and not realizable in certain contexts, regions, scenarios.
2. Hallucinations: generic models may generate incorrect, nonsensical or nonethical information.
3. Knowledge cut-off: incorrect information generated (see 2.), can be structural because of the probabilistic generation mechanisms used. LLMs generate outputs through probabilistic sampling from pre-trained data's distributions, which makes their responses inherently non-deterministic. The output distribution is fixed at the end of pre-training, so models do not automatically incorporate new knowledge, leading to an unavoidable knowledge cut-off.
4. Challenge of control: output from generic models is still human-based and can contain biased information. How to check this?
5. Generalisations: output from generic AI models are based on large data sets on all domains possible; they mostly lack specialized expertise in a specific domain.

2.2 Customisable AI-models

An alternative approach that can integrate AI into automated systems for item authoring and item banking in educational assessment, can be called *'Customisable AI models'*, also based on the use of LLMs.

A customizable AI model is one that can undergo fine-tuning to suit specific tasks. This process of fine-tuning involves training the model on fine-tuned domain-specific data, enabling it to generate text, understand context, and provide more precise and tailored responses for specialized applications or industries. Since these LLMs are trained on fine-tuned domain-specific data, they are enabled to provide more domain-specific and contextually aware responses. This is the primary reason for their popularity across industries[13].

An example in the field of educational assessment is *Finetune Generate*® (Finetune learning, (14), an patented custom AI model designed by assessment professionals for assessment professionals and subject matter experts (SMEs). This engine continuously improves from simple SME feedback during item development, permeating every stage of the assessment development cycle. Finetune Generate applies advanced AI to empower authors to design, edit, and create items, passages and content in novel ways and with a high level of efficiency and accuracy when it comes to high-quality results that do not need extensive prompt revisions and optimisation. The developers call this Generate engine an *AI-human hybrid model,* that aims to be(come) an Automated Item Generation (AIG) model that combines the best qualities of technology and human expertise [15].

2.3 How to combine AI-models with human expertise

The use of such Customizable AI model for writing test items and developing assessments in many different formats and for various purposes, can revolutionise the practice of educational assessment and could become a next milestone in Testology innovation. However, the development and use of Customisable AI models, like Finetune Generate, is not without challenges [16]. The advantages of generic AI models can be seen, to a certain extent, as the main drawbacks for customizable models:

1. Data dependency: these models rely heavily on quality and quantity of data.
2. Continuous maintenance: customization leads to a constant need for updates and maintenance, what can become even more costly than generic models, as the developments in AI technology are following up each other in an unprecedented speed and scope.
3. Lack of transferability or adaptability: customizable models are often less flexible in adapting new demands, technologies or contextual challenges. This is not to say that these models are 'domain-exclusive' and cannot perform broader linguistic and reasoning competencies; broader

than what was specifically trained by the customized data. When customization is done well and strategically smart, the generated new content might even be more widely applicable than a poorly built general purpose AI model. Here lies a challenge for LLM and machine learning engineers when customizing AI models.

3. Item generation and AI : importance of human-in-the-loop

When writing test items for higher order thinking skills and complex cognitive domains, the expert review by humans is of critical importance. It will , however, fundamentally change the role of that expert, who traditionally has been seen as someone who combines subject matter(s) expertise with general understanding of what makes a good test [10]. The way in which AI-powered algorithms analyze existing item banks with test questions is generating items in a much higher speed and efficiency. A generic model application like ChatGPT – certainly when using the more sophisticated plus versions under paid subscription – can find patterns and create new questions based on these patterns in a way that human beings will not be able to match. But, does this imply that the human factor can be missed? Certainly not, in my opinion. The use of AI in item generation will change the role of the test expert, but a human check will still be needed. In many large scale assessments that are used for making valid judgments to inform decision making, it is essential to find a good combination of innovative, technology-based item development and responsible human decision making.

Even with the use of AI technology in the most advanced ways possible, the development of good quality items for educational assessment, remains a creative process. With this article I want to pledge for the use and further development of customisable AI-human hybrid models – like Finetune Generate – for specific use in educational assessment.

4. STAIR-AIG: an example of a ‘human-in-the-loop’ approach

In paragraphs 2.2 and 2.3 I have explained the potentials, but also the drawbacks of development of and working with customizable AI models for item writing. The initial investments needed, to build a customized AI model for the specific purpose of item writing, can be high. Certainly when such model cannot build on previously developed models in very common language areas, like English, Spanish or Chinese, it might be ‘out of reach’ for assessment bodies in smaller language areas, like for example Russian or Kazakh. I called this a lack of transferability. But also once developed, a customized AI model for item writing needs maintenance and regular updates. [16].

An alternative approach, that combines a general purpose AI model for item writing and test development, with the advantages of the ‘human-in-the-loop’, is presented by STAIR-AIG, what stands for Systematic Tool for Assessment Item review in Automatic Item Generation) [10]. This tool is specifically developed for the use in test development processes for high complex domains, like higher-order thinking skills. In such domains an additional review by a human expert is important because the definition of measurement goals and development of adequate assessments questions is challenging.

Anyone who works in the field of educational measurement and development of high-stakes tests in large scale contexts, knows how time consuming and to a certain extent ‘subjective’ the review processes can be. Hybrid frameworks that combine automation

of items development with human intervention and supervision, offer an attractive alternative. This approach is also known as ‘Human-AI-collaboration’(HAIC) [10].

STAIR-AIG is an iterative framework in contrast to the static character of most AIG models. It incorporates AI-based reviews with the human reviewers’ feedback. The outcomes of human-based feedback are used as training data to refine the behavior of the LLMs.

This iterative process is illustrated in figure 1 below.

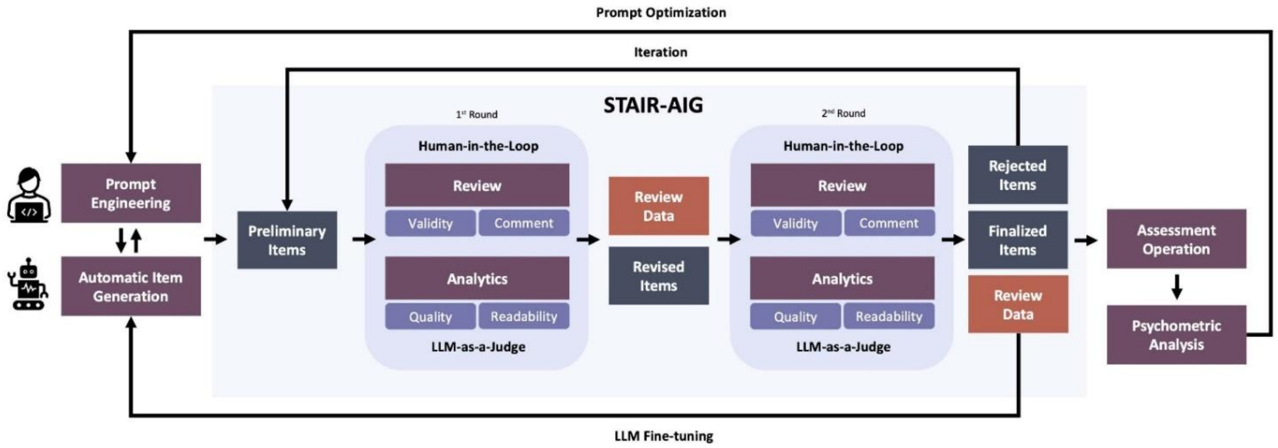

Figure 1 - A human-in-the-loop framework for AI-generated test item development and psychometric validation

What makes STAIR-AIG different from comparable frameworks, is that the human-based review output is used for optimization processes both upstream and downstream. Upstream it will be used to optimize the prompts, what improves the next cycles of item generation. Downstream, the accumulated data from reviews can be used to train the LLMs, in order to produce new items that are better aligned to the objectives of the assessment and to the quality criteria as used by the human experts.

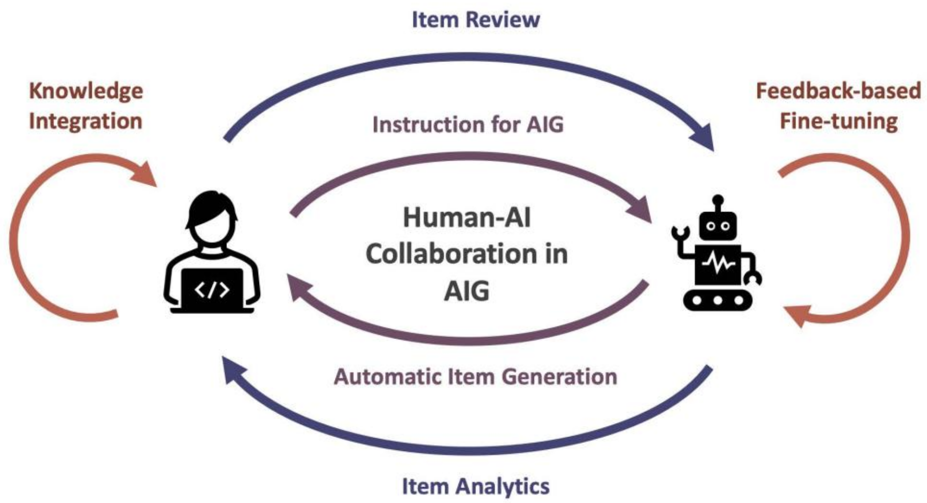

Figure 2 - A human-AI collaboration framework for automatic item generation, review, analytics, and feedback-based fine-tuning

For a more detailed description and explanation on the STAIR-AIG framework, I refer to the very informative and inspiring article by Euigyum Kim and others [10].

Conclusion

The rapid developments in AI technology are certainly influencing the world of Testology. The way we develop test items, organize test assembly, delivery, administration, evaluation and results reporting, will change. Even more, it is already changing in an unprecedented way and speed. For educational assessment in high stakes contexts, general purpose AI models offer promising perspectives but also some drawbacks and possible risks. To a certain extent, customizable AI models may offer a better solution for specific contexts and applications of educational assessment. Whatever use of AI technology will be chosen, the human factor cannot be missed [17]. Building, developing and maintaining knowledge and skills in the field of Testology will be needed in the future. Experienced and well-trained test experts will have a role, although that role will be different from before. Their task will shift from item writing and production tasks, to reviewing, checking, justifying and accounting for quality.

Notice: this article is a revised and expanded version of a full paper as written, submitted and used for a presentation on the 41th Annual Conference of the Association for Educational Assessment in Africa, 2025, in Addis Ababa/Ethiopia. It has not been published in any other medium.

Special thanks go to Mr. Euigyum Kim, Graduate School of Education, Sogang University, Seoul/Korea, for his valuable feedback and additional recommendations.

References

1. Huseyn V. AI in educational assessments: balancing innovation with responsibility // E-assessment Association. – 2024. – https://www.e-assessment.com/news/ai-in-educational-assessments-balancing-innovation-with-responsibility/
2. Scharaschkin A. The role of AI in educational assessment: AQA webinar, February 14, 2024. – London : AQA, 2024.
3. Leigh A. AI Question Generation: The Risks and Alternatives // Grademaker News. – 2024. – https://www.grademaker.com/news/ai-question-generation-the-risks-and-alternatives/
4. Grademaker: official website. – 2024. – https://www.grademaker.com/
5. Vretta Solutions: official website. – 2026. – https://www.vretta.com/solutions/
6. TAO Testing: official website. – 2026. – https://www.taotesting.com/
7. Assessment Systems: official website. – 2026. – https://assess.com/
8. Gierl M. J., Lai H. Instructional Topics in Educational Measurement (ITEMS) module: Using automated processes to generate test items // Educational Measurement: Issues and Practice. – 2013. – Vol. 32, №3. – P. 36–50. – https://doi.org/10.1111/emip.12018
9. Ghader Kurdi G., Leo J., Parsia B., Sattler U., Al-Emari S. A systematic review of automatic question generation for educational purposes // International Journal of Artificial Intelligence in Education. – 2020. – Vol. 30. – P. 121-204. – https://doi.org/10.1007/s40593-019-00186-y
10. Kim E., Li S., Khalil S., Shin H. J. STAIR-AIG: Optimizing the Automated Item Generation Process through Human-AI Collaboration for Critical Thinking Assessment // Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications. – Association for Computational Linguistics, 2025. – P. 920-930.
11. Dhamani N., Engler M. Introduction to Generative AI. – 2nd ed. – New York : Simon and Schuster, 2026. – 350 p. – ISBN 9781633434882.
12. Jain A. Generative AI vs Discriminative AI // Medium. – 2024. – https://medium.com/@abhishekjainindore24/generative-ai-vs-discriminative-ai-fab0e12f6025
13. Sarkar S. General-purpose vs. customizable LLMs: weighing in on the debate // Gnani.ai Blog. – 2024. – https://www.gnani.ai/resources/blogs/general-purpose-vs-customizable-llms-weighing-in-on-the-debate/
14. Finetune Generate: official website. – 2026. –https://finetunelearning.com/finetune-generate/
15. Kahn S. M., Almeida H. J. Generate: A NLG system for educational content creation // Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021). – 2021. – P. 45-52.
16. Kamani V. Custom AI models vs. Off-the-Shelf Tools: Choosing the Right Path // LinkedIn Pulse. – 2025. – https://www.linkedin.com/pulse/custom-ai-models-vs-off-the-shelf-tools-choosing-right-varun-kamani-bogsf/
17. Alina A. von Davier, Burstein J. AI in the assessment ecosystem: A human-centered AI perspective // Artificial Intelligence in Education: The Intersection of Technology and Pedagogy / ed. by P. Ilic, I. Casebourne, R. Wegerif. – Cham : Springer, 2025. – Vol. 261. – (Intelligent Systems Reference Library). – P. 15-38.

Н. Дитерен

ТЕСТОЛОГИЯ ЖӘНЕ ТЕХНОЛОГИЯЛАР: ЖОҒАРЫ РЕТТІ ОЙЛАУ ДАҒДЫЛАРЫНА АРНАЛҒАН САПАЛЫ ТЕСТ ТАПСЫРМАЛАРЫН ЖАСАУДА АДАМ ФАКТОРЫ ЖИ ҚОЛДАНУДЫ ЖӘНЕ ОНЫ ЖЕТІЛДІРУДІ ҚАЛАЙ ІСКЕ АСЫРА АЛАДЫ

Технологиялар білім беруді бағалауда басымдыққа ие болып, инновацияларды басқарып келеді. Жасанды интеллектті (ЖИ) қолдануды енгізу, дәлірек айтқанда, тест тапсырмаларын жазуға арналған әмбебап қозғалтқыштарды қолдануды көптеген сарапшылар бағалаудағы парадигмалық өзгеріс ретінде қарастырады. Жалпы ЖИ-ді қолдану тест тапсырмаларын автоматты және тиімдірек жасауда айқын артықшылықтарға ие. ЖИ-мен жұмыс істейтін алгоритмдердің тест сұрақтары бар қазіргі тапсырма банктерін талдау тәсілі тапсырмаларды әлдеқайда жоғары жылдамдықпен және тиімділікпен жасайды. Бірақ әмбебап ЖИ модельдерін қолдану ресурстар, шығындар, сенімділік, этикалық мәселелер және құқықтық мәселелер тұрғысынан кейбір елеулі кемшіліктерге де ие. Баптауға болатын ЖИ модельдері білім беруді бағалау құралдарын дамыту үшін маңызды болып табылатын балама ұсынуы мүмкін. Қандай ЖИ моделі қолданылса да, адам факторын жоққа шығаруға болмайды. Тапсырма жасауда ЖИ-ді қолдану тест сарапшысының рөлін өзгертеді, бірақ адамның тексеруі әлі де қажет болады. Шешім қабылдауға негізделген жарамды пайымдаулар жасау үшін қолданылатын көптеген ауқымды бағалауларда инновациялық тапсырмаларды әзірлеу мен адамның жауапты шешім қабылдауын жақсы үйлестіру маңызды. Жоғары ретті ойлау дағдыларын бағалау үшін күрделі танымдық домендерде адам сарапшысының шолуы ерекше маңызды.

Түйін сөздер: тестология, автоматты тапсырма генерациясы, генеративті ЖИ модельдері, үлкен тілдік модельдер.

Н. Дитерен

ТЕСТОЛОГИЯ И ТЕХНОЛОГИИ: КАК ЧЕЛОВЕЧЕСКИЙ ФАКТОР МОЖЕТ ИСПОЛЬЗОВАТЬ И СОВЕРШЕНСТВОВАТЬ ПРИМЕНЕНИЕ ИИ ПРИ СОЗДАНИИ КАЧЕСТВЕННЫХ ТЕСТОВЫХ ЗАДАНИЙ НА НАВЫКИ МЫШЛЕНИЯ ВЫСШЕГО ПОРЯДКА

Технологии становятся все более доминирующими и лидируют в инновациях в области оценивания образовательных результатов. Внедрение использования искусственного интеллекта (ИИ), а точнее применение универсальных движков для написания тестовых заданий, рассматривается многими экспертами как смена парадигмы в оценивании. Использование ИИ в целом имеет явные преимущества при автоматическом и более эффективном создании тестовых заданий. То, как алгоритмы, основанные на ИИ, анализируют существующие банки заданий с тестовыми вопросами, приводит к созданию заданий с гораздо большей скоростью и эффективностью. Но использование универсальных моделей ИИ также имеет некоторые серьезные недостатки с точки зрения ресурсов, затрат, надежности, этических проблем и юридических вопросов. Настраиваемые модели ИИ могут предложить альтернативу, которая заслуживает серьезного рассмотрения для разработки инструментов оценивания образовательных результатов. Какая бы модель ИИ ни использовалась, человеческий фактор не может быть упущен. Использование ИИ в генерации заданий изменит роль эксперта по тестам, но человеческая проверка все еще будет необходима. Во многих крупномасштабных оцениваниях, которые используются для принятия обоснованных суждений для информирования о принятии решений, крайне важно найти хорошее сочетание инновационной разработки заданий и ответственного принятия решений человеком. В сложных когнитивных областях, оценивающих мыслительные навыки более высокого порядка, экспертная оценка человека особенно важна.

Ключевые слова: тестология, автоматическая генерация заданий, генеративные ИИ-модели, крупные языковые модели.

References

1. Huseyn, V. (2024). AI in educational assessments: Balancing innovation with responsibility. *E-assessment Association*. https://www.e-assessment.com/news/ai-in-educational-assessments-balancing-innovation-with-responsibility/
2. Scharaschkin, A. (2024). *The role of AI in educational assessment: AQA webinar, February 14, 2024*. AQA.
3. Leigh, A. (2024). AI question generation: The risks and alternatives. *Grademaker News*. https://www.grademaker.com/news/ai-question-generation-the-risks-and-alternatives/
4. Grademaker. (2024). *Official website*. https://www.grademaker.com/
5. Vretta Solutions. (2026). *Official website*. https://www.vretta.com/solutions/
6. TAO Testing. (2026). *Official website*. https://www.taotesting.com/
7. Assessment Systems. (2026). *Official website*. https://assess.com/
8. Gierl, M. J., & Lai, H. (2013). Instructional Topics in Educational Measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice, 32*(3), 36-50. https://doi.org/10.1111/emip.12018
9. Ghader Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*, 121-204. https://doi.org/10.1007/s40593-019-00186-y
10. Kim, E., Li, S., Khalil, S., & Shin, H. J. (2025). STAIR-AIG: Optimizing the automated item generation process through human-AI collaboration for critical thinking assessment. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 920-930). Association for Computational Linguistics.
11. Dhamani, N., & Engler, M. (2026). *Introduction to generative AI* (2nd ed.). Simon and Schuster.
12. Jain, A. (2024). Generative AI vs discriminative AI. *Medium*. https://medium.com/@abhishekjainindore24/generative-ai-vs-discriminative-ai-fab0e12f6025
13. Sarkar, S. (2024). General-purpose vs. customizable LLMs: Weighing in on the debate. *Gnani.ai Blog*. https://www.gnani.ai/resources/blogs/general-purpose-vs-customizable-llms-weighing-in-on-the-debate/
14. Finetune Generate. (2026). *Official website*. https://finetunelearning.com/finetune-generate/
15. Kahn, S. M., & Almeida, H. J. (2021). Generate: A NLG system for educational content creation. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)* (pp. 45-52).
16. Kamani, V. (2025). Custom AI models vs. off-the-shelf tools: Choosing the right path. *LinkedIn Pulse*. https://www.linkedin.com/pulse/custom-ai-models-vs-off-the-shelf-tools-choosing-right-varun-kamani-bogsf/
17. Von Davier, A. A., & Burstein, J. (2025). AI in the assessment ecosystem: A human-centered AI perspective. In P. Ilic, I. Casebourne, & R. Wegerif (Eds.), *Artificial intelligence in education: The intersection of technology and pedagogy* (Vol. 261, pp. 15-38). Springer.

Information about the author:
Nico Dieteren – Dutchtestologist, The Netherlands, dutchtestologist@outlook.com

Автор туралы мәлімет:
Нико Дитерен – Dutchtestologist, Нидерланды, dutchtestologist@outlook.com

Сведения об авторе:
Нико Дитерен – Dutchtestologist, Нидерланды, dutchtestologist@outlook.com