Delft University of Technology

Toward Personalised Learning Experiences
Beyond Prompt Engineering

Kruis, Joost; Pera, Maria Soledad; Napel, Zoë ten; Landoni, Monica; Murgia, Emiliana; Huibers, Theo; Feskens, Remco

DOI
10.1145/3628516.3659367

Publication date
2024

Document Version
Final published version

Published in
Proceedings of ACM Interaction Design and Children Conference

Citation (APA)
Kruis, J., Pera, M. S., Napel, Z. T., Landoni, M., Murgia, E., Huibers, T., & Feskens, R. (2024). Toward Personalised Learning Experiences: Beyond Prompt Engineering. In *Proceedings of ACM Interaction Design and Children Conference: Inclusive Happiness, IDC 2024* (pp. 644-649). (Proceedings of ACM Interaction Design and Children Conference: Inclusive Happiness, IDC 2024). Association for Computing Machinery (ACM). https://doi.org/10.1145/3628516.3659367

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher
is the copyright holder of this work and the author uses the
Dutch legislation to make this work public.

Toward Personalised Learning Experiences: Beyond Prompt Engineering

Joost Kruis
joost.kruis@cito.nl
Cito Institute for Educational Measurement
Arnhem, Netherlands

Maria Soledad Pera
m.s.pera@tudelft.nl
Web Information Systems - TU Delft
Delft, Netherlands

Zoë ten Napel
zoe.tennapel@cito.nl
Cito Institute for Educational Measurement
Arnhem, Netherlands

Monica Landoni
monica.landoni@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

Emiliana Murgia
emilianamurgia@gmail.com
University of Genova
Genova, Italy

Theo Huibers
t.w.c.huibers@utwente.nl
University of Twente
Enschede, Netherlands

Remco Feskens
remco.feskens@cito.nl
Cito Institute for Educational Measurement
Arnhem, Netherlands
University of Twente
Enschede, Netherlands

ABSTRACT

We discuss the foundation of a collaborative effort to explore AI's role in supporting (teachers and) children in their learning experiences. We integrate principles of educational psychology, AI, and HCI, and align with best practices in education while undertaking a human-centered focus on design and development that puts the student at the centre and keeps the expert-in-the-loop. Initially, we study assessment items—questions or tasks tied to a learning target. These items vary in complexity, serve as indicators of students' grasp of specific concepts and spotlight areas where support may be needed. This preliminary analysis will help us outline a framework to guide the design and evaluation of AI technology for K-12 education. Such a framework would ensure that assessment item generation technology goes beyond the current one-dimensional approach by incorporating multifaceted, adaptable perspectives that consider the variegated landscape of learners' needs, subject matter complexities, and pedagogical goals.

CCS CONCEPTS

• **Social and professional topics → Student assessment**; *K-12 education*; **Children**; • **Human-centered computing**;

KEYWORDS

Generative AI, Learning, Human-centered Design

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
IDC '24, June 17–20, 2024, Delft, Netherlands
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0442-0/24/06
https://doi.org/10.1145/3628516.3659367

ACM Reference Format:
Joost Kruis, Maria Soledad Pera, Zoë ten Napel, Monica Landoni, Emiliana Murgia, Theo Huibers, and Remco Feskens. 2024. Toward Personalised Learning Experiences: Beyond Prompt Engineering. In *Interaction Design and Children (IDC '24), June 17–20, 2024, Delft, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3628516.3659367

1 LEARNING & ASSESSMENT IN THE ERA OF AI

Discussions surrounding Artificial Intelligence (**AI**) in education have become prominent among researchers, developers, industry professionals, and educators. However, the discourse has often taken siloed stances, focusing either on new technology development or the challenges and opportunities AI brings to education. For instance, we find new techniques to generate educational questions automatically [6, 17], alongside works raising concerns on plagiarism-related issues models like ChatGPT can bring to the classroom [9, 21] or advocating for embracing AI for teaching and learning [1, 15]. Recognising AI's potential to sustain the many facets of education, e.g., personalised tutoring, learning resource recommendation, and lesson planning [2, 18, 33, 39, 40, 42], we argue that meaningful outcomes in this field require a comprehensive, interdisciplinary exploration of AI's role.

In the education spectrum, let us zoom in on **evaluation**. Originally, evaluation referred to "giving value", a pedagogical meaning that should remain essential in the AI era. As evaluation is integral to teaching and learning, the challenge lies in constantly creating high-quality `assessment items`, such as test questions and other learning objects, to help students (and teachers) identify knowledge gaps to address. Crafting such items—integral in the educational evaluation machinery—to enable accurate measurement of student learning and facilitate effective feedback mechanisms, is non-trivial. This process demands expertise in item construction as well as a deep understanding of the concept to be examined. The complexities involved in providing items are fair, free from differential item functioning, and embody desirable characteristics (e.g., reliability and validity), are magnified by the diverse levels of education and the varying abilities of students. Each educational setting requires items that both assess knowledge accurately and guarantee equity and (cultural) inclusivity for all learners. Thus, the task extends beyond creating items that are pedagogically sound also to ensure they are adaptable to the myriad of learning environments and student needs.

Although `assessment item` generation has largely been the responsibility of experts, the emergence of generative AI (**GAI**) has prompted research and industry initiatives to offer solutions that promise to assist this intricate process. These solutions predominantly leverage APIs that interact with established AI models like OpenAI [27], employing slightly optimised prompts to produce educational content [4, 11, 13, 20, 22]. These advancements represent significant strides forward. Yet, they overlook a fundamental principle of education: the need for customisation to cater to diverse learning needs and contexts. In today's 'superdiverse' society [41], education is inherently not a one-size-fits-all endeavour. Neither is evaluation, hence the demand for technologies that are as versatile as the populations they aim to serve.

Here, we discuss the preliminary stage of an ongoing collaboration focused on establishing a holistic framework to guide the design and evaluation of technologies that leverage GAI in constructing effective `assessment items` for K-12 education. Anchored in human-centered principles and keeping the expert-in-the-loop, this framework aims to ensure that item generation goes beyond prompt manipulation [7, 16], and accounts for diverse learners' needs, the nature of the subject matter and its different levels of complexities, and pedagogical goals. In turn, items would align with specific learning outcomes and be engaging and reflective of societal values. For example, when studying the states of liquids, tailoring the content to reflect cultural backgrounds and experiences might lead Italian students to think about boiling water for pasta; British students might relate it to boiling water for tea. Integrating insights from various disciplines and blending theoretical exploration and empirical research (including iterative design and field testing with educators, content developers, and children), we explore the integration of AI in educational content creation. In Sections 2 and 3, we discuss practical examples showcasing scenarios where GAI can aid `assessment item` generation and identify areas requiring further exploration. Informed by emerging insights and prior experience, we suggest open research directions as the next steps that can contribute to identifying the components of the framework we advocate for.

2 A PRELIMINARY EXPLORATION OF GAI FOR ASSESSMENT ITEM GENERATION

Exploring how GAI can improve knowledge-building practices [19] and enhance the teaching-and-learning process [34] is in its infancy. While the idea of using GAI to foster culturally grounded and personalised learning shows promise [3, 14], putting it into practice is challenging, particularly in producing learning content that caters to an individual's current educational needs. Given GAI's proficiency in generating coherent and contextually appropriate text, it is conceivable that GAI could facilitate the arduous task of `assessment item` construction. Despite its sophistication, GAI lacks the nuanced understanding required to meet the detailed requirements of educational item construction—it often overlooks the subtleties needed to ensure that content is educationally valid and reliable. Further, the premise that a singular approach could meet the diverse needs inherent in education is fundamentally flawed [23, 32].

When constructing an `assessment item` there are important criteria the item must meet to be suitable to administer [26]. Firstly, the question should be easy to comprehend and clearly convey its intention, ensuring it is unambiguous and contains sufficient information to determine the correct answer. If prefaced by a stimulus (text and/or image), the question should directly relate to the stimulus, with explicit references. The expected format of the item also imposes constraints. For multiple-choice items, for example, regardless of the subject (e.g., science or literature), criteria regarding the quality of the alternatives also apply. In this case, all alternatives should be clear, unambiguous, plausible, mutually exclusive, approximately of equal length, with a single correct answer.

To enable our work in this space, we created an *Assistant*—an instance of an AI model setup for a specific purpose; here, item construction—in the OpenAI platform using the GPT-4 model. Through a trial-and-error approach, we engaged with the *Assistant* to probe different settings and gain insights into the potential and limitations of GAI for `assessment item` construction. We provided specific instructions to guide the generation of questions and alternatives. For brevity, we omit the ($\sim$ 20) precise characteristics that the questions and alternatives should adhere to (see [26] for a description of these rules in Dutch) as well as the formatting details of the expected output. We tested this prompt with `assessment items` from the Programme for International Student Assessment (PISA) test, which assesses the ability of 15-year-olds to apply their reading, mathematics and science knowledge and skills to real-life challenges [30]. In this initial stage of our work, we address learners as a group to pave the way to further explore the implications of item personalisation, i.e., items that meet the needs of individual learners.

```
Instructions OpenAI Assistant:
You will receive a prompt with the following structure:
{{subject: name of a subject}}
{{compentency: the ability to be tested}}
{{stimulus: stimulus prefacing a question}}
{{question: the question that is asked or empty}}
{{n_alt: the number of alternatives to generate}}
{{population: the population characteristics for the group of test-takers}}

Act as an expert in test and item construction, educational measurement and
the {{subject}} that is provided in the prompt.

if ({{question}} part of the prompt is empty) {
    Generate a question based on the {{stimulus}} that measures the
    specified {{competency}} and is suitable for the {{population}}.
    Ensure that the generated {{question}} adheres to the following principles:
        1. ...
} else {
    Ensure that the provided {{question}} adheres to the following principles:
        1. ...
}
Generate {{n_alt}} response alternatives for the given/generated {{question}}.
Ensure that the generated {{alternatives}} adhere to the following principles:
    1. ...
```

```
Return latex output enclosed in a code block with the following format: ...
```

Bird migration. For our first example, we probed the *Assistant*'s ability to produce alternatives for an existing question. Specifically, the second question about Bird migration [29] from the 2015 PISA Science test, an open-ended item. We tasked the *Assistant* with generating four response alternatives with this input:

- **Subject**: Science
- **Competency**: Evaluate and design scientific enquiry
- **Stimulus**: Bird migration is a seasonal large-scale movement of birds to and from their breeding grounds. Every year volunteers count migrating birds at specific locations. Scientists capture some of the birds and tag their legs with a combination of coloured rings and flags. The scientists use sightings of tagged birds together with volunteers' counts to determine the migratory routes of birds.
- **Population**: 15 year old high-school students
- **Question**: Identify a factor that might make the volunteers' counts of migrating birds inaccurate, and explain how that factor will affect the count.

As per PISA correction guidelines, one of these factors should be considered in correct responses: (1) The observers may miss counting some birds because they fly high; (2) If the same birds are counted more than once, that can make the numbers too high; (3) For birds in a large group, volunteers can only estimate how many birds there are.

The *Assistant* first generated these four alternatives:

- **Alternatives**: **(A)** The counted number might be higher due to birds repeating the same route. It affects the count by inflating the actual number of migratory bird population. **(B)** Birds may fly too high to be seen. This can significantly underestimate the actual bird counts. **(C)** Inclement weather may affect bird paths causing the volunteers to count fewer birds than usual. **(D)** The absence of volunteers at certain hours of the day, such as early morning when birds typically migrate. This could lead to an underestimation of the bird counts.
- **Correct Answer**: All options are plausible.

This response violated two instruction rules, i.e., there should be exactly at least, and no more than, one correct response. Attempts to regenerate additional alternatives with the same input sometimes led the *Assistant* to return a single option as the **Correct Answer**. After closer inspection, we still found multiple correct answers among alternatives.

- **Alternatives**: **(A)** Bird watchers might count the same bird multiple times as it circles over the count area, leading to overestimations in bird populations. **(B)** Some rare bird species might camouflage well in the environment, making their counts inaccurate. **(C)** The scorching mid-day sun might cause bird watchers to miss sightings, leading to underestimations in bird counts. **(D)** The windy weather might change the usual migration paths, causing the bird watchers to count fewer birds than there actually are.
- **Correct Answer**: A.

When the *Assistant* was alerted to the fact that it violated some of the rules, it responded as follows:

- **Assistant's Note**: The original question was open-ended and complex for a multiple choice question. It was simplified to focus on a single factor for the inaccuracy and formatted for multiple-choice.
- **Question**: What effect might poor weather conditions have on the volunteers' counts of migrating birds?
- **Alternatives**: **(A)** The poor weather might deter the birds from migrating, leading to an undercount. **(B)** The poor weather might enhance visibility, leading to a more accurate count. **(C)** The poor weather might impact the visibility, making it hard for volunteers to see and count the birds accurately. **(D)** The poor weather might attract more birds to migrate, leading to an overcount.
- **Correct Answer**: C.

The new question seemed better suited to generate alternatives with a single correct answer. However, we noted multiple alternatives that could be correct. Thus, we adapted the question to be even more specific and unambiguous.

- **Question**: What is a factor that would lead the volunteers to undercount the number of migrating birds?
- **Alternatives**: **(A)** The birds fly at a height that is difficult to observe. **(B)** The birds migrate at a speed faster than estimated. **(C)** The birds change their migratory routes. **(D)** The birds do not wear their tags on the migration journey.
- **Correct Answer**: A.

Although the new alternatives were less ambiguous, multiple ones would lead to undercounting the migrating birds (only D is clearly incorrect). Indeed, the *Assistant* could produce suitable alternatives. Yet, doing so in a single pass—a must for automated deployment—was unlikely. Consequently, the *Assistant* might help generate candidate alternatives, but expert intervention would still be needed to select suitable ones that adhere to all requirements for good items. Interestingly, when the *Assistant* was not provided with a question and had to generate one from scratch, its initial response appeared much better, despite exhibiting notable areas for improvement.

- **Generated Question**: If you were to plan a scientific investigation to study bird migratory routes, what data would you require from the volunteers and what additional data might you collect to enhance the precision of your study?
- **Alternatives**: **(A)** Counts of migrating birds and additional data on bird species **(B)** Counts of migrating birds and additional data on weather conditions **(C)** Counts of migrating birds and additional data on date and time of observations **(D)** Counts of migrating birds and additional data on location of bird sightings
- **Correct Answer**: D.

Analysing questions and alternatives, we saw that the first part of the question ("what data would you require from the volunteers") is redundant, as it yielded the same answer ("Counts of migrating birds"). Including the location of bird sightings in the correct response is vital. Thus, an expert might take the response from the *Assistant* and modify it as:

- **Question**: If you were to plan a scientific investigation to study bird migratory routes, what data other than counts of migrating birds would you require from the volunteers?
- **Alternatives**: **(A)** the bird species **(B)** the weather conditions **(C)** the date and time of observations **(D)** the location of bird sightings
- **Correct Answer**: D.

We assessed a scenario where we retained the original input but adjusted the population age range to 5-year-old primary-school students, recognising the original question's unsuitability for this age. The *Assistant* typically flagged this matter with a message (below) and generated a new question with response alternatives. While these often revealed improvement areas, there were instances where proper combinations of questions and alternatives were produced.

Example feedback. The population is very young and the question expects a high level of cognitive ability that could be challenging for 5-year-old primary- school students. At this age, children might struggle to understand the concept of bird migration, tagging, or the need for additional data for scientific investigation. Therefore, the question should be modified to be more appropriate for the age group.

- **Population**: 5-year-old primary-school students
- **Assistant's Note**: The original question was revised as it was too complex for the population. The revised question is more basic and appropriate, focusing on recognising the purpose of using the coloured rings and flags for the birds.
- **Question**: Why do you think scientists put coloured rings and flags on the legs of the birds?

- **Alternatives**: (A) To make the birds look pretty. (B) Because they want to play with the birds. (C) To track the birds and find out where they go. (D) Because the birds like to wear them.
- **Correct Answer**: C.

Overall, for these types of factual/knowledge questions, although not consistently, GAI can generate appropriate questions and alternatives; which we expected given the inherent nature of this language model.

Sailing ships. We shift the focus to a mathematics item to study how the *Assistant* handles calculations. For this, we use a 2012 PISA question about Sailing ships (question 1/3) [28], a multiple-choice item with five choices. Per PISA correction guidelines, the correct response is 30 km/h. We input the question and requested five response alternatives:

- **Subject**: Mathematics
- **Competency**: Apply calculation of percentage within a given real-world situation
- **Stimulus**: 95% of world trade is moved by sea, by roughly 50 000 tankers, bulk carriers and container ships. Most of these ships use diesel fuel. Engineers are planning to develop wind power support for ships. Their proposal is to attach kite sails to ships and use the wind's power to help reduce diesel consumption and the fuel's impact on the environment.
- **Population**: 15-year-old high-school students
- **Question**: One advantage of using a kite sail is that it flies at a height of 150 m. There, the wind speed is approximately 25% higher than down on the deck of the ship. At what approximate speed does the wind blow into a kite sail when a wind speed of 24 km/h is measured on the deck of the ship?

The *Assistant* stated that all rules of the question were followed and generated these five alternatives:

- **Alternatives**: **(A)** 28 km/h **(B)** 29 km/h **(C)** 30 km/h **(D)** 31 km/h **(E)** 32 km/h
- **Correct Answer**: C.

This response adhered to all rules. While the alternatives are plausible, they are not probable. When asked to guarantee that the incorrect alternatives reflect common mistakes, the *Assistant* produced:

Some students might forget to add the increase back to the original speed, resulting in an answer of just 6 km/h. Alternatively, others might interpret "25% higher" incorrectly and multiply the original speed by 25% (giving 18 km/h) rather than adding on 25%. Alternatively, they might misinterpret 25% higher as half of the original speed, leading to alternatives like 36 km/h. A random answer like 50 km/h can be included to account for guesses not based on mathematical calculations.

- **Alternatives**: **(A)** 6 km/h **(B)** 18 km/h **(C)** 30 km/h **(D)** 36 km/h **(E)** 50 km/h
- **Correct Answer**: C.

The explanations of the mistakes were consistent with the alternatives and plausible; two alternatives were used by PISA. Thus, it appears reasonable to improve the *Assistant* by integrating this general rule:

Let's assume common misconceptions/mistakes are:

- misunderstanding of percentage calculation as 25 km/h instead of percent increase.
- confusion with percentage points leading to a result lower than 24 km/h, i.e., 22 km/h.
- misunderstanding of percentage application leading to double the original speed, i.e., 48 km/h.
- a random plausible distractor around average speed of wind, i.e., 20 km/h.

None of the alternatives follow from likely calculation mistakes, so we introduced a new rule (below), resulting in other alternatives. Note that the correct answer should be D. Some alternatives did not reflect common mistakes. When asked about the correct response, the *Assistant* noticed the mistake and improved itself.

5. Make sure the incorrect alternatives follow from likely mistakes for the target population related to the compentency

- **Alternatives**: **(A)** 18 km/h **(B)** 25 km/h **(C)** 28 km/h **(D)** 30 km/h **(E)** 48 km/h
- **Correct Answer**: C.

This example underscores the need for human intervention to generate good alternatives. The *Assistant* seems to struggle to correctly calculate the answer and generate likely mistakes, ensuring the alternatives follow from these mistakes. Since the *Assistant* can do one of the tasks and improve itself, a collaboration among multiple assistants could be a feasible approach in this scenario.

3 INSIGHTS & NEXT STEPS

Reflecting on our discussion thus far, it is clear that GAI can assist in automating the `assessment item` generation process. However, designing GAI-powered technology to support educators, teaching, and children's learning, rather than solely automating educational processes, requires a holistic approach beyond prompt engineering. Drawing from our expertise in Information Retrieval, Child-Computer Interaction, AI, Educational Assessment, and Expert Item Construction, and building from our prior work [5, 10, 12, 24, 25, 31, 35–37], we suggest several open research directions. These are intended to inform the structure of a framework that can ultimately guide the design and evaluation of K-12 AI technology—using GAI-driven `assessment item` generation as a use case.

We first emphasise the importance of defining what constitutes a "good" `assessment item`. Key considerations include balancing challenge levels to prevent boredom or frustration, identifying misconceptions, ensuring inclusive and politically correct language, and addressing the inherent biases of algorithms, [38]. We also stress the need to empirically define the limitations of GAI in the primary school environment, which involves investigating whether large language models align with competencies and subject matters taught at different educational levels, as preliminary work suggests otherwise [24]. Our earlier example on calculations further underscores this need. Beyond suitability, prompt engineering, and model analysis, we posit that moving on from addressing groups of learners, as in our examples, to exploring GAI's abilities to produce personalised items requires considering other multiple perspectives. While interest and readability are pivotal, determining other factors, such as cultural and social alignment, is equally crucial yet more complex to automate. Ethical considerations should be carefully managed to ensure the responsible design and use of AI technologies in educational settings, especially when catering to a vulnerable user group, such as children [8, 43].

We argue that effective `assessment items` can pinpoint individual children's learning needs. To ensure these solutions genuinely reflect children's needs, it is imperative to engage children and teachers (the experts-in-the-loop) as part of the research and design process. This collaborative approach will be vital in creating `assessment items` that enrich the learning experiences of all stakeholders involved, starting with children.

REFERENCES

[1] Ibrahim Adeshola and Adeola P. Adepoju. 2023. The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments* (2023), 1–14. https:

//doi.org/10.1080/10494820.2023.2253858

[2] Vincent Aleven, Richard Baraniuk, Emma Brunskill, Scott Crossley, Dora Demszky, Stephen Fancsali, Shivang Gupta, Kenneth Koedinger, Chris Piech, Steve Ritter, Danielle R. Thomas, Simon Woodhead, and Wanli Xing. 2023. Towards the Future of AI-Augmented Human Tutoring in Math Learning. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, Ning Wang, Genaro Rebolledo-Mendez, Vania Dimitrova, Noboru Matsuda, and Olga C. Santos (Eds.). Springer Nature Switzerland, Cham, 26–31. https://doi.org/10.1007/978-3-031-36336-8_3

[3] Fx. Risang Baskara. 2023. Fostering Culturally Grounded Learning: Generative Ai, Digital Storytelling, And Early Childhood Education. *International Conference of Early Childhood Education in Multiperspectives* 2 (2023), 351–361.

[4] Aras Bozkurt and Ramesh C. Sharma. 2023. Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education* 18, 2 (2023), i–vii. https://doi.org/10.5281/zenodo.8174941

[5] Karin Bügel and Piet F. Sanders. 1998. *Richtlijnen voor de ontwikkeling van onpartijdige toetsen.* Cito Instituut voor Toetsontwikkeling, Arnhem, The Netherlands.

[6] Dhawaleswar Rao CH and Sujan Kumar Saha. 2020. Automatic Multiple Choice Question Generation From Text: A Survey. *IEEE Transactions on Learning Technologies* 13, 1 (2020), 14–25. https://doi.org/10.1109/tlt.2018.2889100

[7] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. arXiv:2310.14735 [cs.CL]

[8] Jennifer Chubb, Sondess Missaoui, Shauna Concannon, Liam Maloney, and James Alfred Walker. 2022. Interactive storytelling for children: A case-study of design and development considerations for ethical conversational AI. *International Journal of Child-Computer Interaction* 32, Article 100403 (2022), 14 pages. https://doi.org/10.1016/j.ijcci.2021.100403

[9] Debby R.E. Cotton, Peter A. Cotton, and J. Reuben Shipway. 2024. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International* 61, 2 (2024), 228–239. https://doi.org/10.1080/14703297.2023.2190148

[10] Eva de Schipper, Remco Feskens, and Jos Keuning. 2021. Personalized and Automated Feedback in Summative Assessment Using Recommender Systems. *Frontiers in Education* 6 (2021), 11 pages. https://doi.org/10.3389/feduc.2021.652070

[11] Paul Denny, Viraj Kumar, and Nasser Giacaman. 2023. Conversing with Copilot: Exploring Prompt Engineering for Solving CS1 Problems Using Natural Language. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1.* Association for Computing Machinery, New York, NY, USA, 1136–1142. https://doi.org/10.1145/3545945.3569823

[12] Brody Downs, Aprajita Shukla, Mikey Krentz, Maria Soledad Pera, Katherine Landau Wright, Casey Kennington, and Jerry Fails. 2020. Guiding the selection of child spellchecker suggestions using audio and visual cues. In *Proceedings of the Interaction Design and Children Conference.* Association for Computing Machinery, New York, NY, USA, 398–408. https://doi.org/10.1145/3392063.3394390

[13] Scott F. Gregory. 2024. Empowering Teaching With Prompt Engineering: How to Integrate Curriculum, Standards, and Assessment for a New Age. In *Transforming Education With Generative AI: Prompt Engineering and Synthetic Content Creation*, Ramesh C. Sharma and Aras Bozkurt (Eds.). IGI Global, Hershey, PA, USA, 239–260. https://doi.org/10.4018/979-8-3693-1351-0.ch012

[14] Ariel Han and Zhenyao Cai. 2023. Design implications of generative AI systems for visual storytelling for young learners. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference.* Association for Computing Machinery, New York, NY, USA, 470–474. https://doi.org/10.1145/3585088.3593867

[15] Jussi S. Jauhiainen and Agustín Garagorry Guerra. 2023. Generative AI and ChatGPT in School Children's Education: Evidence from a School Lesson. *Sustainability* 15, 18, Article 14025 (2023), 22 pages. https://doi.org/10.3390/su151814025

[16] Nimrita Koul. 2023. *Prompt Engineering for Large Language Models.* Nimrita Koul.

[17] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2019. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education* 30, 1 (2019), 121–204. https://doi.org/10.1007/s40593-019-00186-y

[18] Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence* 6, Article 100210 (2024), 10 pages. https://doi.org/10.1016/j.caeai.2024.100210

[19] Alwyn Vwen Yen Lee, Seng Chee Tan, and Chew Lee Teo. 2023. Designs and practices using generative AI for sustainable student discourse and knowledge creation. *Smart Learning Environments* 10, 59 (2023), 1–18. https://doi.org/10.1186/s40561-023-00279-1

[20] Unggi Lee, Haewon Jung, Younghoon Jeon, Younghoon Sohn, Wonhee Hwang, Jewoong Moon, and Hyeoncheol Kim. 2023. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies* (2023), 1–33. https://doi.org/10.1007/s10639-023-12249-8

[21] Chung Kwan Lo. 2023. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences* 13, 4, Article 410 (2023), 15 pages. https://doi.org/10.3390/educsci13040410

[22] Leo S. Lo. 2023. The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship* 49, 4, Article 102720 (2023), 3 pages. https://doi.org/10.1016/j.acalib.2023.102720

[23] Louis Major, Gill A. Francis, and Maria Tsapali. 2021. The effectiveness of technology-supported personalised learning in low- and middle-income countries: A meta-analysis. *British Journal of Educational Technology* 52, 5 (2021), 1935–1964. https://doi.org/10.1111/bjet.13116

[24] Emiliana Murgia, Zahra Abbasiantaeb, Mohammad Aliannejadi, Theo Huibers, Monica Landoni, and Maria Soledad Pera. 2023. ChatGPT in the Classroom: A Preliminary Exploration on the Feasibility of Adapting ChatGPT to Support Children's Information Discovery. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization.* Association for Computing Machinery, New York, NY, USA, 22–27. https://doi.org/10.1145/3563359.3597399

[25] Emiliana Murgia, Maria Soledad Pera, Monica Landoni, and Theo Huibers. 2023. Children on ChatGPT Readability in an Educational Context: Myth or Opportunity?. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization.* Association for Computing Machinery, New York, NY, USA, 311–316. https://doi.org/10.1145/3563359.3596996

[26] José Noijons. 2017. Het construeren van gesloten vragen. In *Toetsen op School*, Piet F. Sanders (Ed.). Stichting Cito Instituut voor Toetsontwikkeling, Arnhem, The Netherlands, 95–108. https://cito.nl/media/avhgajek/toetsen-op-school-algemeen.pdf

[27] OpenAI. 2024. OpenAI Platform. https://platform.openai.com/. Accessed: 20-03-24.

[28] Organisation for Economic Co-operation and Development (OECD). 2012. PISA 2012 Released Field Mathematics Items - Question: Sailing Ships [PM923Q01]. https://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf. Accessed: 2024-03-21.

[29] Organisation for Economic Co-operation and Development (OECD). 2015. PISA 2015 Released Field Trial Cognitive Test Items - Question: Bird Migration [S656]. https://www.oecd.org/pisa/PISA2015Questions/platform/index.html?user=&domain=SCI&unit=S656-BirdMigration&lang=eng-ZZZ. Accessed: 2024-03-21.

[30] Organisation for Economic Co-operation and Development (OECD). 2024. Programme for International Student Assessment (PISA). https://www.oecd.org/pisa/. Accessed: 20-03-24.

[31] Maria Soledad Pera, Katherine L. Wright, Casey Kennington, and Jerry A. Fails. 2023. Children and Information Access: Fostering a Sense of Belonging. In *Joint Proceedings of the IUI 2023 Workshops: HAI-GEN, ITAH, MILC, SHAI, SketchRec, SOCIALIZE (CEUR Workshop Proceedings, Vol. 3359)*, Alison Smith-Renner and Paul Taele (Eds.). CEUR, Sydney, Australia, 254–257. https://doi.org/10.1145/3563359.3596996

[32] Daniel D. Pratt. 2002. Good Teaching: One Size Fits All? *New Directions for Adult and Continuing Education* 2002, 93 (2002), 5–16. https://doi.org/10.1002/ace.45

[33] Sherry Ruan, Jiayu He, Rui Ying, Jonathan Burkle, Dunia Hakim, Anna Wang, Yufeng Yin, Lily Zhou, Qianyao Xu, Abdallah AbuHashem, Griffin Dietz, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2020. Supporting children's math learning with feedback-augmented narrative technology. In *Proceedings of the Interaction Design and Children Conference.* Association for Computing Machinery, New York, NY, USA, 567–580. https://doi.org/10.1145/3392063.3394400

[34] Lena Ivannova Ruiz-Rojas, Patricia Acosta-Vargas, Javier De-Moreta-Llovet, and Mario Gonzalez-Rodriguez. 2023. Empowering Education with Generative Artificial Intelligence Tools: Approach with an Instructional Design Matrix. *Sustainability* 15, 15, Article 11524 (2023), 20 pages. https://doi.org/10.3390/su151511524

[35] Piet F. Sanders (Ed.). 2017. *Toetsen op School.* Stichting Cito Instituut voor Toetsontwikkeling, Arnhem, The Netherlands. https://cito.nl/media/avhgajek/toetsen-op-school-algemeen.pdf

[36] Piet F. Sanders and Theo J.H.M. Eggen. 1993. Inleiding. In *Psychometrie in de praktijk*, Theo J.H.M. Eggen and Piet F. Sanders (Eds.). Stichting Cito Instituut voor Toetsontwikkeling, Arnhem, The Netherlands, 1–16.

[37] Cornelis Sluijter. 1998. *Toetsen en beslissen.* Cito Instituut voor Toetsontwikkeling, Arnhem, The Netherlands.

[38] Catherine Stinson. 2022. Algorithms are not neutral: Bias in collaborative filtering. *AI and Ethics* 2, 4 (2022), 763–770. https://doi.org/10.1007/s43681-022-00136-w

[39] Danielle R. Thomas, Shivang Gupta, and Kenneth R. Koedinger. 2023. Comparative Analysis of Learnersourced Human-Graded and AI-Generated Responses for Autograding Online Tutor Lessons. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, Ning Wang, Genaro Rebolledo-Mendez, Vania Dimitrova, Noboru Matsuda, and Olga C. Santos (Eds.). Springer Nature Switzerland, Cham, 714–719. https://doi.org/10.1007/978-3-031-36336-8_110

[40] Jessica Van Brummelen, Viktoriya Tabunshchyk, and Tommy Heng. 2021. "Alexa, Can I Program You?": Student Perceptions of Conversational Artificial Intelligence Before and After Programming Alexa. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference*. Association for Computing Machinery, New York, NY, USA, 305–313. https://doi.org/10.1145/3459990.3460730

[41] Steven Vertovec. 2023. *Superdiversity: Migration and social complexity*. Taylor & Francis, New York, NY, USA. https://doi.org/10.4324/9780203503577

[42] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2022. Informing Age-Appropriate AI: Examining Principles and Practices of AI for Children. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 536, 29 pages. https://doi.org/10.1145/3491102.3502057

[43] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2024. Challenges and opportunities in translating ethical AI principles into practice for children. *Nature Machine Intelligence* 6, 3 (2024), 265–270. https://doi.org/10.1038/s42256-024-00805-x