

Examining ChatGPT responses to TPACK assessment items

Peter Gonschwerowski¹, Edith Lindenbauer² and Benjamin Rott³

¹University of Cologne, Mathematics & Natural Sciences, Germany; pgonsche@uni-koeln.de

²University College of Education Upper Austria, Austria; edith.lindenbauer@ph-ooe.at

³University of Cologne, Mathematics & Natural Sciences, Germany; brott@uni-koeln.de

The development and, thus, the objective and valid assessment of the skill of selecting digital learning material (dLM) is important for pre-service teachers. In this paper, we compare the responses of ChatGPT 3.5 with responses from pre-service teachers to items for assessing this skill to gain insights into the capabilities of ChatGPT and the longevity of the items for its assessment. The results reveal that, for one, ChatGPT 3.5, so far, cannot analyze dynamic dLMs, and second, it does not make a decision on the use of the dLM but provides predominantly TPK but appropriate reasoning for using or not using it. ChatGPT TPK responses are comparable with pre-service teachers' responses, and further studies are required to understand its impact fully. Still, the presented results support the projected effects of ChatGPT on assessments in teacher education and the evaluation of the skill of selecting dLM.

Keywords: Artificial intelligence, performance-based assessment, teacher education, pre-service teachers, teacher evaluation.

Introduction

Selecting digital learning materials (dLMs) is a crucial skill for educators due to the possibilities dLMs offer in teaching and the varying quality of the many freely available dLMs. Therefore, this skill needs to be developed in teacher training, and valid and objective assessment instruments are required to assess the success of such development processes (König et al., 2022; Redecker & Punie, 2017). In this context, the most frequently used instruments so far are self-assessment instruments, which are based on the TPACK framework (Mishra & Koehler, 2006), and only a few valid and objective TPACK assessment instruments have been published, one of them using open-text items developed by Gonschwerowski et al. (in review). With the free public availability of text-based natural language processing and multimodal artificial intelligence (AI) models, examining such assessment instruments and respective items becomes increasingly important because we must understand how AI chatbots respond to such items and potentially undermine their assessment purpose. Therefore, this paper examines the ability of ChatGPT 3.5 to respond to items developed and validated in Gonschwerowski et al. (in review) for assessing the (pre-service) teachers' skill of selecting dLM by having to reason for or against the use of a given dLM for specific learning content in the context of a specific learner age and potential special educational needs.

Existing multimodal AI systems process text and images and either generate new images, modify images, or categorize images (Livberber & Ayvaz, 2023). However, such systems do not currently interpret dynamic dLMs and their functionality or intended learning goal. We chose ChatGPT 3.5, although it is a text-based AI model and cannot directly analyze or interpret images because of its

popularity (Livberber & Ayvaz, 2023) and free use. We developed a textual description of a dLM and its functionality and provided it to ChatGPT to answer the following research questions.

RQ1: How does ChatGPT reason for or against using a dLM for a specific group of learners, characterized by their age and special learning needs, when provided with a text-based description of said dLM?

RQ2: How does ChatGPT's reasoning compare to pre-service teachers who evaluated the same dLM?

By answering these research questions, we want to understand the longevity of the assessment items and gain insights on how to refine the items in the future so that ChatGPT or similar tools cannot undermine an assessment of pre-service teachers using the items.

Theoretical framework: TPACK

To assess the skill of selecting dLM, we rely on the TPACK framework by Mishra & Koehler (2006).

This framework is frequently used to describe and assess the digital competence of pre-/in-service teachers (Gonscherowski & Rott, 2023). The framework describes the content knowledge (CK), pedagogical knowledge (PK), and technological knowledge (TK) that teachers require for successfully integrating technology in teaching. In the framework, pedagogical content knowledge (PCK), technological content knowledge (TCK), technological pedagogical (TPK), and technological pedagogical content knowledge (TPCK) detail the knowledge required because of the interplay of CK, PK, and TK. Mishra and Kohler (2006) define TK, TCK, TPK, and TPCK as follows:

TK is knowledge about standard technologies, such as books, chalk and blackboard, and more advanced technologies, such as the Internet and digital video. (ibid., pp. 1027–1028)

TCK is knowledge about the manner in which technology and content are reciprocally related. Although technology constrains the kinds of representations possible, newer technologies often afford newer and more varied representations and greater flexibility in navigating across these representations. (ibid., p. 1028)

TPK is knowledge of the existence, components, and capabilities of various technologies as they are used in teaching and learning settings, and conversely, knowing how teaching might change as the result of using particular technologies. (ibid., p. 1028)

TPCK [...] is the basis of good teaching with technology and requires an understanding of the representation of concepts using technologies; pedagogical techniques that use technologies in constructive ways to teach content. (ibid., pp. 1028-1029)

In this study, we refer to Mishra and Kohler (2006, pp. 1026-1027) for the definitions of CK, PK, and PCK. The descriptions of TCK, TPK, and TPCK entail reasons for using dLM in a teaching setting and can also be used to categorize reasons for or against using a dLM in a teaching situation.

Method

We use a qualitative case study method to answer the research questions. To do so, we compare ChatGPT 3.5's responses to items one to four (see Table 4) developed by Gonscherowski et al. (in review), with the responses by pre-service teachers when assessing the skill of selecting dLM using

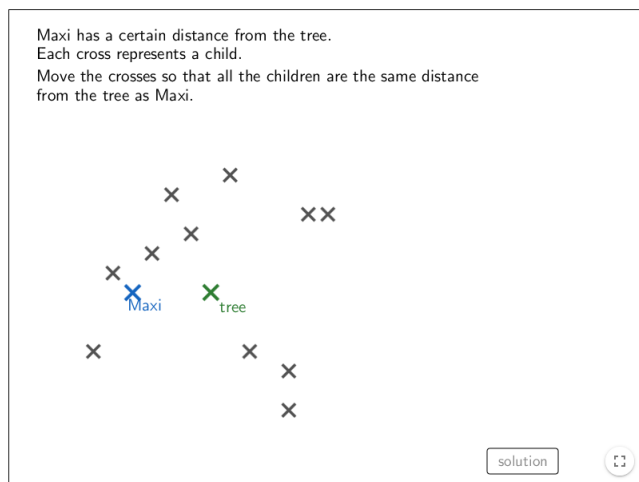
a specific dLM (see Figure 1). First, we outline the coding and scoring of the items, and then we compare ChatGPT and pre-service teachers' results.

Assessing the skill of selecting dLM

In Gonscherowski et al. (in review), four items were developed and validated to assess the skill of selecting dLMs (see Table 4). These items aim to evaluate pre-service teachers' understanding of the learning content that a particular dLM is intended to deliver (one open-ended item), the learner age group (grades 1-13, in two-year increments), and the special learner needs with which the dLM would be used (two closed items). A fourth open text item inquires about the reasons for or against using the dLM.

To evaluate the items' ability to assess the skill of selecting dLM reliable, valid, and objective, they were integrated into an online test using a specific dLM (see Figure 1). The online test was distributed among mathematics pre-service teachers, and participation was voluntary.

Maxi and the tree



Maxi and the tree

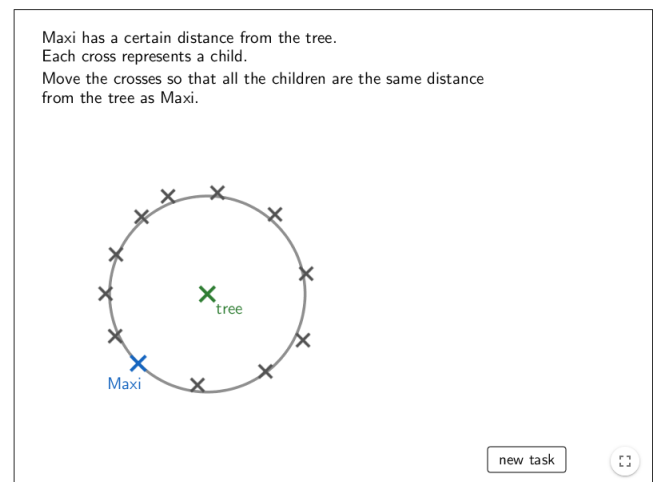


Figure 1: Starting and end point of the dLM

There were 379 participants each from one university in Germany ($n = 314$) and one in Austria ($n = 65$). The German pre-service teachers were distributed in programs on primary education ($n = 110$), special education ($n = 173$), and lower secondary education ($n = 31$); the Austrian pre-service teachers were enrolled in a combined program for lower and upper secondary levels. Furthermore, the participants covered all relevant semesters of their mathematics education program: first year ($n = 57$), second year ($n = 149$), third year ($n = 93$), and year seven or higher ($n = 80$).

The responses to the items were coded following qualitative content analysis and summarized in scores: items one to three were analyzed with scores from zero to three, and item four on a scale from zero to four. Table 1 shows example responses for items one to three for a generic and detailed description of the learning content appropriate for the learner age group and the special learner needs; the former scored two, and the latter scored three. Inadequate descriptions of the learning content and generic or detailed descriptions of the learning content but that were inappropriate for the selected learner age group or the selected special needs were scored zero and one, respectively.

Table 1: Example responses by pre-service teachers (items one to three)

Definition of the score (item 1 to 3)	Example response
A generic description of the learning content appropriate for learner age/special educational needs	“Introduction to the definition circle’ Learner age: 3-4 and hearing and communication learning needs” (WS2022/2023_ANJA07B_Pre)
A detailed description of the learning content and appropriate for learner age/special educational needs	“To derive the definition of a circle line. All points on the circumcircle have equal distance from the center.’ Learner age: 5-6 no special educational needs” (WS2022/2023_EDJO06B_Pre)

The responses to item four, “Justify why or why not you would use the digital material,” were coded and scored using the codebook developed based on the TPACK framework (Mishra & Koehler, 2006) and other validated and generally accepted studies. The reasons for or against the use of dLM were categorized as TCK, TPK, and TPCK-based arguments. Coding of the TCK reasoning entails arguments such as different or dynamic representations of the learning content and reducing or increasing extraneous cognitive load because of the dLM. The coding of the TPK reasoning encompasses arguments such as self-directed learning, differentiation, learner motivation, exploring, teacher efficiency, and learner distraction.

Table 2: Single TCK and TPK example reasoning by pre-service teachers (item four)

Example response 1	Example response 2
TCK: “My students (with special needs) would be overwhelmed with the dLM...” (WS2022/2023_MAHA10L_Pre)	TPK: “Because by moving the crosses around, the children can discover for themselves what properties all the points of a circle have.” (WS2022/2023_DOMA15S_Pre)

The example responses shown in Table 2 were scored as two, containing a single TCK or TPK reason for or against using the dLM. Responses with either two TCK or TPK reasons were scored three, and responses with both TCK and TPK, thus TPCK reasoning, were scored four, see Table 3.

Table 3: TCK and TPK, thus TPCK example reasoning by pre-service teachers (item four)

Example response 1	Example response 2
“...provides a different way of practicing and illustrating the properties of a circle. However, it is not necessarily suitable for all learners with special needs. The tree and the learning atmosphere outside the classroom (not in the usual environment) can lead to too much distraction.” (WS2022/2023_CHGU24M_Pre)	“...Changing environments (digital/non-digital) with students and exploring other learning environments encourages interest. Experiencing on their own how these mathematical relationships are connected makes understanding easier.” (WS2022/2023_SAHE17B_Pre)

Example one in Table 3 entails “...a different way of practicing and illustrating...” a TCK argument and “...lead to too much distraction...” a TPK and thus is TPKC reasoning. No reasoning was scored zero, and generic arguments were scored one.

The coding and the scoring of the suitability of the dLM for a particular learning age, learning needs, and learning content were derived from the local curriculum (items one to three). The arguments for or against the dLM (item four) were coded based on the codebook developed based on the framework by Mishra & Koehler (2006) and other validated and generally accepted studies. For the skill of selecting dLM, the scores were combined on a scale from zero to seven.

The development process of a textual description of the dLM for ChatGPT

The dLM presented in Figure 1 encompasses the mathematical topic of a circle. Learners should discover the concept of the circle based on its defining property: a figure consisting of all the points in the plane that are a specific distance (radius) from a certain point (the center). Since ChatGPT applies a language-processing AI model and only accepts text input, the three authors crafted a textual description of the dLM using collaborative editing and multiple review cycles. In the processes of crafting the description, the authors applied the following guiding rules: a) explicitly state that the description is one of dynamic learning material, b) describe the activity the learners need to perform, c) not use the term circle or properties of the circle, and d) use as much of the wording of the task description included in the dLM as possible. The process resulted in the following description of the dLM and its dynamic functionality, including an introduction. “The dynamic learning material we want to evaluate presents a task in a dynamic digital applet showing a tree, a child named Maxi, and further children represented by crosses. Maxi is at a pre-set distance from the tree. The learners should move the crosses representing the children so they all have the same distance from the tree as Maxi. Finally, learners can press a button named “solution,” and the solution to the task is revealed.” ChatGPT was provided with the description and the four assessment items.

In Table 4, abbreviated ChatGPT responses are contrasted with example responses of pre-service teachers. The full transcript of the ChatGPT session and the text description of the dLM were recorded (OpenAI, 12/25/2023). ChatGPT responses are coded and scored as outlined in the previous section (see Tables 1, 2, and 3).

Results

ChatGPT’s answer to item one, the learning content the dLM is intended to provide, is incorrect as the dLM is not intended for spatial reasoning in geometry or possibly proportional reasoning (see Table 4). The correct learning content contains the characteristics of a circle as a shape consisting of all points of equal distance from a given point, the center. The assigned grade level by ChatGPT (early to middle stage of primary education) is also incorrect. Although the circle is covered in the middle stages of primary school, the concept of equal distance from the center to the circle’s circumference is only covered in lower secondary school. ChatGPT’s responses to item three, the special needs of learners, contain multiple impairments. Each impairment is described with a sentence, of which “fine motor skills,” “cognitive challenges,” and “attention challenges” are appropriate. However, ChatGPT suggested that the dLM could be used with auditory learners if the dLM would be auditive narrated, which was not indicated in the description of the dLM, also correct,

but is conjecture and inappropriate. The scoring as outlined previously resulted in a score of zero pt. for items one to three. Within the text responses, ChatGPT references elements of the description of the dLM, e.g., “Maxi and other children represented by the crosses” or “solution button suggests [...] immediate feedback,” providing specificity and context to its responses, however not all are suitable in response to the items.

The justifications for using or not using the dLM, required in item four provided by ChatGPT, are appropriate but predominantly correspond to TPK reasoning, namely “engaging and interactive” (motivation), “immediate feedback” (self-directed learning), the potential for “inclusivity” or “sensory overload” (both differentiation), and “motor skill challenges” (inclusion). Other reasons provided are coded as overly general arguments (e.g., “overreliance on technology” and “lack of flexibility”) and incorrect argumentation, such as “concept reinforcement,” as the dLM is for the discovery of the learning content and not for its reinforcement. ChatGPT references again elements of the text description of the dLM, for example, “arranging crosses around the tree...” and “...button providing immediate feedback...” to support its arguments.

Table 4: Responses to items assessing the skill “selecting dLM” by pre-service teachers and ChatGPT

#	Items	Example responses of pre-service teachers	pt	ChatGPT 3.5 responses	pt
1	Describe the learning content for which you think the digital material is intended.	“To derive the definition of a circle line. All points on the circumcircle have equal distance from the center.” (WS2022/2023_EDJO06B_Pre)	3	“...it seems that the dynamic learning material is designed for teaching and practicing concepts related to spatial reasoning, geometry, and possibly proportional reasoning in a mathematics education context....”	0
2	For which learner age do you think the digital material is suitable?	“5-6 grade” (WS2022/2023_EDJO06B_Pre)		“...the digital material seems to be designed for learners in the early to middle stages of primary education...”	
3	In your opinion, is the digital material suitable for learners with special educational needs? If so, which ones?	“No special educational needs.” (WS2022/2023_EDJO06B_Pre)		“...the dynamic learning material seems to have characteristics that could make it inclusive and adaptable for a range of learners, including those with certain special educational needs: visual learners, auditory learners, fine motor skills, cognitive challenges, customization and adaptability, attention challenges...”	

4	Justify why or why not you would use the digital material.	<p>“...provides a different way of practicing and illustrating the properties of a circle. However, it is not necessarily suitable for all learners with special needs. The tree and the learning atmosphere outside the classroom (not in the usual environment) can lead to too much distraction.”</p> <p>(WS2022/2023_CHGU24M_Pre)</p>	4	<p>“Certainly, the decision to use the digital material depends on the specific educational context, learning objectives, and the needs of the learners. Here are justifications both for and against using the digital material: Reasons to use the dLM: Engaging and interactive, visual learning, immediate feedback, concept reinforcement, potential for inclusivity; Reasons not to use the dLM: motor skill challenges, overreliance on technology, sensory overload, lack of flexibility.”</p>	3
---	--	---	---	--	---

Without making an actual decision on the use of the dLM, ChatGPT’s reasoning for and against using the dLM corresponds to a score of three points, as previously outlined.

Regarding RQ1, “How does ChatGPT reason for or against using a dLM,” our analysis reveals that ChatGPT does not adequately capture the learning content and the learner age range of the learning content. However, it does provide generic and predominant correct TPK reasoning to justify using or not using the dLM.

Regarding RQ2, “How does the reasoning of ChatGPT compare to pre-service teachers who evaluated the same dLM?” Table 4 reveals that, for item four, ChatGPT’s responses are comparable to those shown in Tables 2 and 3. However, ChatGPT does not really decide and only provides reasons for and against using the dLM. The combined score for the entire sample of pre-service teachers ($n = 379$), as presented in Table 5, shows a total mean score of 2.18 ($SD = 1.59$), which underscores the good results of ChatGPT (score of 3) and echoes its reported capabilities and impact on assessments in higher (teacher) education (Livberber & Ayvaz, 2023).

Table 5: Scores of pre-service teachers and ChatGPT

mean; SD; max score	pre-service teachers ($n = 379$)	ChatGPT
Items one-three (learning content)	0.93; 1.06; 3.00	0.00
Item four (reasoning)	1.25; 1.03; 4.00	3.00
Σ representing the score for “selecting dLM”	2.18; 1.59; 6.00	3.00

Particular to note is ChatGPT’s capability of providing subject-unspecific (TPK) reasoning and its lack of TCK reasoning, which needs to be leveraged when using the four items and a dLM for assessing the skill “selecting dLM.” One should use a dLM that is complex enough to provide various arguments, and we recommend dLMs for discovering (mathematical) concepts, as such a dLM enables potentially more argumentation, specifically more TCK arguments, as, for example, a dLM for drill and practice.

Outlook and Limitations

As ChatGPT 3.5 cannot examine dynamic dLM, an inherent limitation of the case study lies in the authors' description of the dLM. Higher-quality responses by ChatGPT could have potentially been achieved with a more elaborate description of the dLM and by further refinement of the items to cater to ChatGPT. However, we did not want to optimize the responses of ChatGPT, but rather the opposite, as one can conjecture if pre-service teachers can precisely describe the functionality of a dLM and inquire ChatGPT in a way to achieve high-quality responses, they may also possess the skill of selecting dLMs we want to assess with the items. The use of dLM for evaluating the skill of selecting dLM increases the technology requirements of the assessment. In environments with limited online access, potentially the risk posed by using ChatGPT may also be lower. For the future, further testing with other and newer AI models (ChatGPT 5.0) and dLMs is required to fully understand the impact of assessing the skill with the developed items. In addition, the responses of ChatGPT should be compared with responses by in-service teachers whose responses potentially exceed those of pre-service teachers, as hypothesized in Gonscherowski and Rott (2022).

References

- Gonscherowski, P., Lindenbauer, E., Kaspar, K., & Rott, B. (in review). Selecting digital learning material as an approach to assess pre-service teachers' digital competence.
- Gonscherowski, P., & Rott, B. (2023). Selecting digital technology: A review of TPACK instruments. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel, & M. Tabach (Eds.), *Proceedings of the 46th conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 378–386). PME.
- Gonscherowski, P., & Rott, B. (2022). How Do Pre-/In-Service Mathematics Teachers Reason for or against the Use of Digital Technology in Teaching? *Mathematics*, 10(13), 2345. <https://doi.org/10.3390/math10132345>
- König, J., Heine, S., Jäger-Biela, D., & Rothland, M. (2022). ICT integration in teachers' lesson plans: A scoping review of empirical studies. *European Journal of Teacher Education*, 1–29. <https://doi.org/10.1080/02619768.2022.2138323>
- Livberber, T., & Ayvaz, S. (2023). The impact of Artificial Intelligence in academia. *Heliyon*, 9(9). <https://doi.org/10.1016/j.heliyon.2023.e19688>
- Mishra, P. & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017–1054. <https://doi.org/10.1111/j.1467-9620.2006.00684.x>
- OpenAI. (12/25/2023, 4:29:14 PM). ChatGPT [Large language model] <https://chat.openai.com>. <https://chat.openai.com/share/fce6ba96-b89a-4802-9093-04fee07bee5b>
- Redecker, C., & Punie, Y. (2017). European framework for the digital competence of educators – DigCompEdu. *Publications Office of the European Union*. <https://doi.org/10.2760/159770>