

How Knowledgeable is ChatGPT 4o? Assessing the Pedagogical Content Knowledge of a Generative Artificial Intelligence Tool

Scott Cameron

University of Melbourne

scameron@unimelb.edu.au

Carmel Mesiti

University of Melbourne

cmesiti@unimelb.edu.au

Paul Fijn

University of Melbourne

paul.fijn@unimelb.edu.au

Generative Artificial Intelligence (GenAI) chatbots provide teachers with opportunities to enhance their expertise by exploring topics relevant to their practice. However, studies have raised concerns about ChatGPT's mathematical accuracy. This pilot study evaluated ChatGPT's Pedagogical Content Knowledge (PCK) and Content Knowledge (CK), revealing that while it generates varied and detailed responses indicating competence in PCK and CK, some responses are flawed or incorrect. While ChatGPT 4o can serve as a valuable professional learning resource in certain areas of mathematics education, traditional teaching methods and human insight remain essential.

The emergence of ChatGPT and other Generative Artificial Intelligence (GenAI) tools has transformed professional learning opportunities for teachers, offering innovative ways to develop their professional expertise. Where teachers may have traditionally relied on expert presenters and curated online resources, GenAI tools now enable teachers to engage in personalised dialogues with AI chatbots, exploring questions and topics directly related to their practice. However, the integration of GenAI tools presents challenges. Ethical concerns have been raised when used in education, particularly regarding the trustworthiness and reliability of AI-generated outputs. These tools are known to include "hallucinations" or inaccurate information (e.g., Williams, 2024), potentially exposing teachers to misleading or incorrect ideas. Given these considerations, it is crucial to critically examine the potential of GenAI tools to support the professional learning of teachers. This paper reports a pilot study investigating the Pedagogical Content Knowledge (PCK) and Content Knowledge (CK) capabilities of one prominent GenAI tool, ChatGPT 4o. By assessing its strengths and limitations, we aim to evaluate its potential as a supplementary resource for enhancing the professional learning of teachers, whilst also addressing the need for discernment and verification in its use.

Background

Pedagogy and PCK

Alexander (2008) defines pedagogy as the act of teaching, connecting it with educational theories, values, and evidence: "It is what one needs to know, and the skills one needs to command, to make and justify the many different kinds of decisions of which teaching is constituted" (p. 47). Shulman (1987) agrees that expert teaching involves managing both students and ideas in classroom discourse. These descriptions highlight the complexity of classroom practices where educational objectives are mitigated by specific skills and knowledge required for teaching. Alexander (2000) employs "pedagogical" as an inclusive term for classroom-related events, actions, and practices. Pedagogy is integral to education, linking good practice with educational quality (Alexander, 2015; Hardman, 2015; Shulman, 1987). Alexander's metaphorical description of pedagogy as a "deep pool" (2015, p. 253) recognises the challenges of defining these practices.

Shulman (1987) characterises experiential knowledge as the "wisdom of practice" (p. 11), representing the accumulated wisdom of professional experiences. Educational theories of (2025). In S. M. Patahuddin, L. Gaunt, D. Harris & K. Tripet (Eds.), *Unlocking minds in mathematics education. Proceedings of the 47th annual conference of the Mathematics Education Research Group of Australasia* (pp. 93–100). Canberra: MERGA.

learning and instruction are encoded in the histories of conventionalised educational practice and are implicit in the key pedagogical terms used by each community. Schoenfeld (1999) argues that teachers' knowledge resources inform choices regarding instructional practices. "Pedagogical content knowledge" (Shulman, 1986, p. 8), a specific form of teaching knowledge conceptualised as the "transformation of subject matter knowledge in the context of facilitating student understanding" (van Driel et al., 2001, p. 979) is key in informing these choices. The development of this knowledge is facilitated by various enablers of professional learning, including access to resources, ongoing support, and collaboration. In recent years, GenAI chatbots such as OpenAI's ChatGPT, Google Gemini and Microsoft Copilot have emerged as a potential tool for teachers to use to support their professional growth. We expect this trend to continue as education systems further invest in GenAI tools to support teaching and learning.

GenAI and Teacher Professional Learning

Teachers engage in professional learning in a range of ways, including classroom-based learning, face-to-face learning and online learning (AITSL, 2020), with teachers valuing professional learning that focusses on student learning and contributes to improved teaching (Beswick et al., 2017). Recently, GenAI tools have emerged as a tool to support teacher professional learning. In their scoping survey of ChatGPT in mathematics education, Pepin and colleagues (2025) report that ChatGPT can aid teachers' understanding and ability to clearly convey mathematical concepts. For example, studies involving pre-service teachers found that using ChatGPT to analyse solutions to problems can support the critical thinking skills of pre-service mathematics teachers (Drushlyak et al., 2025). Buchholtz and Huget (2024) report similar findings regarding pre-service teachers using ChatGPT as a dialogic assistant for lesson planning and task modification. In contrast, Bagno and colleagues (2023) aimed to ascertain whether ChatGPT could function effectively as a teacher or teaching assistant for linear algebra content. While they reported promising results, ChatGPT was not endorsed for this role, citing solutions that were mathematically incorrect, contained contradictions, or included inaccurate information. More broadly, Ming and Mansor (2023) explored the potential opportunities and challenges associated with using ChatGPT in teacher professional development. The opportunities identified include increased access to resources and scholarly articles, efficiency gains, personalisation of professional learning, and support for collaboration and networking. However, challenges in using ChatGPT for teacher professional development include equity of access, ethical considerations such as academic integrity, and the reliability of outputs. Overall, while GenAI tools such as ChatGPT appear to have the potential to aid teacher professional learning, further research is required to explore the affordances and constraints of using these tools to effectively support the professional learning of mathematics teachers.

Research Design

Building on the need for further investigation into the potential of GenAI tools for teacher professional learning, this pilot study addresses the following research questions: (i) To what extent can the PCK and CK of ChatGPT 4o be assessed in the context of mathematics education? and (ii) How reliable and accurate is ChatGPT 4o's mathematics teaching advice, provided in response to PCK and CK assessment items, when compared to anticipated responses to items generated by experts in mathematics education?

The COACTIV Instrument

An existing instrument, the COACTIV Test of Mathematics Teachers' Professional Knowledge (PCK – CK) was selected for use in this study (Krauss et al., 2008). This instrument

was developed as part of the COACTIV project which aimed to assess secondary mathematics teachers' PCK and CK in Germany. The COACTIV project conceptualised PCK as "knowledge of explanations and representations, knowledge of students' thinking, and knowledge of multiple solutions to mathematical tasks" (p. 888) and CK as "deep background knowledge of school-level mathematics" (p. 888). These definitions have been adopted for this study.

The COACTIV instrument has been statistically validated and used to assess PCK and CK of mathematics teachers in a range of contexts (e.g., Krauss et al., 2008; Krauss, Brunner, et al., 2008). The 22 PCK items assess three subscales (i) knowledge of mathematical tasks, (ii) knowledge of student misconceptions and difficulties, and (iii) knowledge of mathematics-specific instructional strategies. The 13 CK items form a single subscale.

Data Collection

As this research is in its pilot phase, four items, one from each subscale, were carefully selected for inclusion (*Square*, *PCK – Task*; *Parallelogram*, *PCK – Student*; *Minus 1 times minus 1*, *PCK – Instruction*, *Recurring decimal - CK*). The selection was based on their availability (e.g., Krauss et al., 2008) in contrast to the full instrument which was provided by a COACTIV team member. The items and protocol are presented in Figure 1.

Figure 1

Assessment Protocol Including Tasks

Clear your memory of prior conversations

ChatGPT responds to prompt

<Item Square>

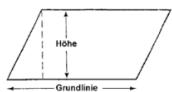
How does the surface area of a square change when the side length is tripled? Show your reasoning. Please note down as many different ways of solving this problem (and different reasonings) as possible (and number each method).

ChatGPT responds to task

<Item Parallelogram>

The area of a parallelogram can be calculated by multiplying the length of its base by its height as shown in the attached image where Höhe: Height and Grundlinie: Base. Please sketch an example of a parallelogram to which students might not be able to apply this formula.

<Image shown below is attached>



ChatGPT responds to task

<Item Minus 1 times minus 1>

A student says: I don't understand why $(-1) \cdot (-1) = 1$. Please outline as many different ways as possible of explaining this concept to your student (and number each explanation).

ChatGPT responds to task

<Item Recurring decimal>

Is it true that $0.999999 \dots = 1$? Please explain your answer.

ChatGPT responds to task

Note: <text> indicates annotations for this paper that were not provided to ChatGPT.

As the COACTIV tests were originally designed as written tests for human subjects, they required little adaptation for use with ChatGPT due to its conversational nature. All items, except *Parallelogram*, were able to be used verbatim. The *Parallelogram* item required

modification due to the inclusion of an image in the prompt, which, for ChatGPT, needed to be uploaded separately rather than embedded within the chat.

To identify potential variations in responses, data were collected from ChatGPT 4.0 on five separate occasions using the protocol described in Figure 1. Each chat session began with an instruction for ChatGPT to clear its memory of prior interactions, minimising the impact of existing chats on data outputs. Items were presented to ChatGPT sequentially, allowing sufficient time for ChatGPT to respond to each item before introducing the next prompt.

Data Analysis

Each chat session underwent independent analysis using the COACTIV Codebook Test of Mathematical Teachers' Professional Knowledge (PCK – CK) as the coding framework. The analysis process began with an assessment of correctness, where responses were evaluated against codes in the COACTIV codebook. Responses aligning with a correct code in the codebook were classified as *fully correct* (see Figure 2). Responses not aligning with a correct code were either *partially correct* or *incorrect*. A *partially correct* response was defined as one that aligns with a correct code in the codebook but is either incomplete or contains an easily rectifiable error that most teachers could identify and correct. An *incorrect* response describes a response that is incorrect due to errors in the mathematics, is didactically insufficient (i.e., appeals to rules without explanation), or does not meet the requirements for a correct answer as outlined in the codebook (e.g., reasoning is not provided, variables not defined).

Figure 2

A Correct, Partially Correct and Incorrect Response to Item Square (Chat_1)

<p>Correct response.</p> <p>This response aligns with the 'covariate/functional' code provided in the COACTIV codebook.</p>	<p>Method 5: Logical reasoning via Length Doubling</p> <ol style="list-style-type: none"> 1. We know that doubling the side length ($2s$) results in quadrupling the area: $(2s)^2 = 4s^2$ 2. Now, tripling the side length is an even larger increase. 3. Instead of $2^2 = 4$, we compute $3^2 = 9$. 4. So, the area must increase 9 times.
<p>Partially correct response.</p> <p>The codebook indicated that the correct use of a visual representation must include a diagram, so this response is considered partially correct.</p>	<p>Method 4: Visual Representation</p> <ol style="list-style-type: none"> 1. Draw a square with side s, and label the area as s^2. 2. Now triple the side length to $3s$, forming a new, larger square. 3. This larger square can be divided into 9 identical original squares (each of area s^2). 4. Now we have 9 times as many original-sized squares, the area increased by a factor of 9.
<p>Incorrect response.</p> <p>This response is incorrect as it is incomplete. The original length is not defined, the original area is not defined, and we must assume that the increase of side length by 3 corresponds to $3s$.</p>	<p>Method 6: Algebraic Expansion</p> <ol style="list-style-type: none"> 1. Express the new area in expanded form: $(3s)^2 = (3s) \times (3s)$ 2. Using the distributive property: $3s \times 3s = (3 \times 3) \times (s \times s) = 9s^2$ 3. The new area is 9 times the original.

Note: Emphasis in original.

Following the correctness coding, a score was calculated for each item. In accordance with the codebook guidelines, a score of '1' was assigned for each unique correct response to an item (i.e., duplicate responses did not contribute to the score). Consequently, a higher score for an item indicates a greater number of unique correct responses provided by ChatGPT. This

scoring method allows for a nuanced evaluation of ChatGPT's performance, considering not only the correctness of responses but also their uniqueness and variety. This approach provides insights into the accuracy and breadth of AI's knowledge. By emphasising unique correct responses, the scoring captures ChatGPT's ability to generate accurate and diverse solutions to these four tasks, offering a more comprehensive assessment of its capacity in this domain.

Findings and Discussion

Figure 3 provides a comprehensive summary of ChatGPT's responses for each item across all five test sessions. The 'number of responses' column indicates a total of 46 different responses to the *Minus 1 times minus 1* item across all five chats. ChatGPT provided ten responses to this item in the first chat, compared to eight in the fourth chat. Of the ten responses to the *Minus 1 times minus 1* task in Chat 1, six of these responses were correct and of these correct responses, four were unique (i.e., responses employing different logic or reasoning). No responses were categorised as partially correct, and four were deemed incorrect. Overall, this indicates variability in how ChatGPT responds to an identical task in different chats.

Figure 3

Summary of the Correctness of Responses by Task and Chat

Item	Chat	Number of responses	Correct Responses			Incorrect Responses
			Fully correct	Unique (Score)	Partially Correct	
<i>Square</i>	1	8	4	3	1	3
	2	9	7	4	0	2
	3	7	6	4	0	1
	4	6	5	4	1	0
	5	6	4	2	1	1
	Total	36	26 (72%)	17 (65%)	3 (8%)	7 (19%)
<i>Parallelogram</i>	1	1	0	0	1	0
	2	1	0	0	0	1
	3	1	0	0	0	1
	4	1	0	0	0	1
	5	1	0	0	0	1
	Total	5	0 (0%)	0 (0%)	1 (20%)	4 (80%)
<i>Minus 1 times minus 1</i>	1	10	6	4	0	4
	2	9	3	3	2	4
	3	10	7	4	0	3
	4	8	5	5	0	3
	5	9	6	3	2	1
	Total	46	27 (59%)	19 (70%)	4 (9%)	15 (33%)
<i>Recurring decimal</i>	1	6	4	4	2	0
	2	5	3	3	0	2
	3	5	4	4	1	0
	4	5	3	3	2	0
	5	7	4	4	1	2
	Total	28	18 (64%)	18 (100%)	6 (21%)	4 (14%)
Overall		115	71 (61.7%)	54 (76.1%)	14 (12.2%)	30 (26.1%)

Overall, ChatGPT provided fully correct responses approximately 60% of the time, with an additional 10% being partially correct, and 25% incorrect. However, the correctness of the responses appears to be task-dependent, as evidenced by the variation in the percentage of correct responses across different items. For example, a third of responses to *Minus 1 times minus 1* were incorrect, indicating that some mathematical concepts may pose greater challenges for ChatGPT. The following section delves into the patterns identified within each correctness category, offering insights into ChatGPT's strengths and limitations.

Fully Correct Responses

A comparative analysis of the scores and the number of correct responses reveals patterns in ChatGPT's performance across different items. Notably, only two items elicited unique correct responses: the *Minus 1 times minus 1* task in Chats 2 and 4; and the *Recurring decimal* task across all chats. Notably, the *Recurring decimal* task assesses CK rather than PCK, which is the focus of the other items. These findings suggest that ChatGPT may be more adept at generating diverse, unique responses when dealing with tasks related to pure mathematical content rather than mathematics pedagogy.

The ability to explain concepts in multiple ways is a hallmark of effective teaching (e.g., State of Victoria, 2017). Therefore, ChatGPT's capacity to generate various correct responses, particularly in content-focussed tasks, presents an opportunity for teachers to develop this expertise. By reviewing the range of responses provided by ChatGPT, teachers could potentially expand their repertoire of explanations and approaches. The potential to use ChatGPT in this manner contrasts with the use of ChatGPT to create resources, such as lesson plans, where ChatGPT generated lesson plans promote a narrow view of mathematics teaching as "telling" where the teacher explains or demonstrates key procedures (Cameron & Mesiti, 2024). We suggest that prompting ChatGPT to complete tasks and explain common misconceptions and teaching strategies to remedy these may be a useful tool to develop teacher knowledge and to inform mathematics teaching and learning.

Partially Correct Responses

The *Parallelogram* item emerged as an outlier in this study, with ChatGPT failing to provide any correct responses. The item required the respondent to draw a diagram of a parallelogram where students might not be able to use the formula $\text{Area} = \text{Base} \times \text{Height}$. Across all chat sessions, ChatGPT generated images that failed to depict a parallelogram (see Figure 4), and only in one chat did it generate a description of such a parallelogram (see Figure 5). The provision of a correct written response and an incorrect image meant this was considered a partially correct response as we anticipate that a mathematics teacher would be able to understand the written response and sketch the corresponding diagram.

Figure 4

Sample ChatGPT Image for Parallelogram and Anticipated Correct Responses to Parallelogram.

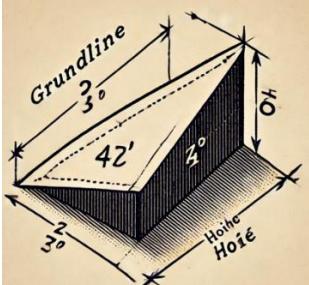
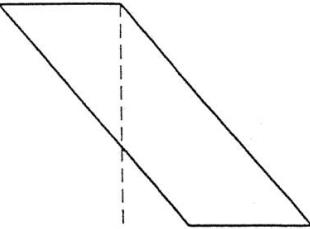
 <p>Incorrect diagram (Chat 4).</p>	 <p>Anticipated correct responses.</p>
--	---

Figure 5

Chat GPT's Written Response to Parallelogram (Chat_1).

A possible parallelogram that might confuse students when applying the area formula is one where the height lies **outside** the shape rather than within it. This often happens with **obtuse parallelograms** where the perpendicular height is drawn from a vertex to an extended base.

This result, while noteworthy, is not entirely unexpected. Although ChatGPT integrates Dall-E for image generation, it is important to recognise that Dall-E creates images from text (OpenAI, 2021) in contrast to ChatGPT, a Large Language Model that creates text from text and images (OpenAI, 2022). In this case, the absence of a detailed text description, specifying characteristics of the required parallelogram, likely contributed to ChatGPT's inability to generate a suitable image. This limitation highlights a significant constraint in supporting professional learning in areas that rely heavily on visual representations, such as measurement and geometry, and emphasises the need for careful consideration when employing AI tools in mathematics. While ChatGPT can be a valuable resource in some areas of mathematics, there are others where human activities, such as drawing diagrams, cannot be replaced.

Score on PCK Tasks

Studies using the COACTIV instrument typically report PCK of cohorts using average scores for tasks within three subscales (e.g., Krauss et al., 2008). As this pilot study did not administer the entire test, we cannot directly compare ChatGPT's performance to existing studies. Krauss and colleagues (2008) report the performance of several cohorts, including 198 mathematics teachers that were “fairly representative of German 10th grade mathematics teachers” (p. 878). These teachers averaged a mean score of 18.6 points across the 22 items on the for the PCK test, indicating a mean score of 0.85 points per item. The maximum score was 37 with a mean score of 1.68 points per item. While we have not yet administered the full test to ChatGPT 4o, its mean score for these four items was 2.4 per item. However, it is crucial to contextualise these results. The human test was administered under timed conditions; participants may not have had the time or stamina to describe all possible responses to a single item. In contrast, ChatGPT is not constrained by time or fatigue, presenting an unfair advantage in this testing format. This disparity in testing conditions raises important questions about the comparability of PCK or CK between GenAI tools and humans when using written instruments administered under timed conditions. While ChatGPT's performance is impressive, it may not accurately reflect the real-world application of PCK, where teachers must make rapid decisions in dynamic classroom environments (e.g., Barendsen & Henze, 2019).

Conclusions and Future Directions

We set out to assess the PCK and CK of ChatGPT, a GenAI tool. Using the COACTIV PCK and CK test instrument and codebook, we identified correct, partially correct, and incorrect responses to both PCK and CK items. The codebook's clear delineation of anticipated responses facilitated reliable judgements. Collaboration with a mathematician supported the identification of duplicate responses despite ChatGPT labelling them as different approaches.

We were impressed by ChatGPT's ability to provide a large range of unique responses to both PCK and CK items, noting the limitation of the *Parallelogram* item for a Large Language Model. ChatGPT's diverse range of responses demonstrates its potential as a valuable resource, offering a variety of approaches to enhance their teaching strategies and support student learning. However, only approximately 60% of responses were fully correct, indicating that ChatGPT may not be sufficiently accurate to reliably develop mathematics teachers' PCK, especially for novices who may struggle to identify incorrect and partially correct responses.

This pilot study tested a limited range of tasks on one GenAI tool. Future research should expand the assessment of PCK and CK across different task and content areas. Comparing multiple GenAI tools could determine which best supports the professional learning of mathematics teachers. As these tools evolve, and new tools are released to the market, monitoring of their PCK and CK is crucial. We posit that as the reliability of a GenAI tool on a PCK and CK test increases, they may become powerful resources for tailored professional learning, potentially leading to positive outcomes for teaching and learning.

References

Alexander, R. (2000). *Culture and pedagogy: International comparisons in primary education*. Blackwell.

Alexander, R. (2008). *Essays on pedagogy*. Routledge. <https://doi.org/10.4324/9780203609309>

Alexander, R. (2015). Teaching and learning for all? The quality imperative revisited. *International Journal of Educational Development*, 40, 250–258. <https://doi.org/10.1016/j.ijedudev.2014.11.012>

Australian Institute for Teaching and School Leadership [AITSL] (2020). *Selecting and undertaking professional learning*. <https://www.aitsl.edu.au/teach/improve-practice/practical-guides/selecting-and-undertaking>

Bagno, E., Dana-Picard, T., & Reches, S. (2023). *ChatGPT may excel in states medical licensing examination but falters in basic linear algebra*. arXiv. <https://doi.org/10.48550/arXiv.2306.16282>

Barendsen, E., Henze, I. (2019). Relating PCK and teacher practice using classroom observation. *Research in Science Education*, 49, 1141–1175. <https://doi.org/10.1007/s11165-017-9637-z>

Beswick, K., Fraser, S., & Crowley, S. (2017). "The best ever": Mathematics teachers' perceptions of quality professional learning. In M. Graven, H. Venkat, A. Essien, & P. Vale (Eds.), *Proceedings of the 43rd Conference of the International Group for the Psychology of Mathematics Education* (pp. 105–112). PME.

Buchholtz, N., & Huget, J. (2024). ChatGPT as a reflection tool to promote the lesson planning competencies of pre-service teachers. In E. Faggiano, A. Clark-Wilson, M. Tabach & H.-G. Weigand (Eds.), *Proceedings of the 17th ERME Topic Conference MEDA 4* (pp. 129–136). University of Bari Aldo Moro.

Cameron, S., & Mesiti, C. (2024). What kind of mathematics teacher is ChatGPT? In J. Višňovská, E. Ross, & S. Getenet (Eds.), *Surfing the waves of mathematics education. Proceedings of the 46th annual conference of the Mathematics Education Research Group of Australasia* (pp. 135–142). MERGA.

Brushlyak, M., Lukashova, T., Shamonia, V., & Semenikhina, O. (2025). ChatGPT-based simulation helps to develop the pre-service mathematics teachers' critical thinking. *International Journal of Instruction*, 18(1), Article 1. <https://doi.org/10.29333/iji.2025.1819a>

Hardman, F. C. (2015). *Making pedagogical practices visible in discussions of educational quality. Background paper prepared for the Education for All Global Monitoring Report 2015*. UNESCO.

Krauss, S., Baumert, J., & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. *ZDM*, 40(5), 873–892. <https://doi.org/10.1007/s11858-008-0141-9>

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology*, 100(3), 716–725. <https://doi.org/10.1037/0022-0663.100.3.716>

Ming, G. K., & Mansor, M. (2023). Exploring the impact of ChatGPT on teacher professional development: Opportunities, challenges and implications. *Asian Journal of Research in Education and Social Sciences*, 5(4).

OpenAI (2021). *DALL-E: Creating images from text*. <https://openai.com/index/dall-e/>

OpenAI (2022). *Introducing ChatGPT*. <https://openai.com/index/dall-e/>

Pépin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education*. <https://doi.org/10.1007/s40751-025-00172-1>

Schoenfeld, A.H. (1999). Models of the teaching process. *The Journal of Mathematical Behavior*, 18(3), 243–261. [https://doi.org/10.1016/s0732-3123\(99\)00031-0](https://doi.org/10.1016/s0732-3123(99)00031-0)

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4–14. <https://doi.org/10.3102/0013189x015002004>

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1–23. <https://doi.org/10.17763/haer.57.1.j463w79r56455411>

State of Victoria (2017). *High impact teaching strategies; Excellence in teaching and learning*. Department of Education and Training, Victoria.

van Driel, J.H., Veal, W.R., & Janssen, F.J.J.M. (2001). Pedagogical content knowledge: an integrative component within the knowledge base for teaching. *Teaching and Teacher Education*, 17(8), 979–986.

Williams, R. T. (2024). The ethical implications of using generative chatbots in higher education. *Frontiers in Education*, 8. <https://doi.org/10.3389/feduc.2023.1331607>