

Secondary mathematics teachers' experiences of using ChatGPT to design probability and statistics assessment items

Minsung Kwon¹ and Inah Ko²

¹California State University Northridge, USA; minsung.kwon@csun.edu

²University of Michigan, USA; inahko@umich.edu

The purpose of this study was to explore secondary mathematics teachers' experiences of using ChatGPT to design probability and statistics assessment items. For this purpose, we analyzed 22 secondary mathematics teachers' conversations with ChatGPT and their survey responses in terms of their overall experiences with ChatGPT, their intentions to use ChatGPT-generated assessment items, and affordances and challenges of using ChatGPT to design assessment items. The results showed that most teachers did not specify the purpose of assessment and only one teacher identified mathematical errors in the ChatGPT's responses. The teachers employed a wide range of follow-up questions in responding to the ChatGPT's suggestions. The survey results showed that their intentions to use ChatGPT were polarized. The teachers perceived that ChatGPT provided affordances such as creativity and efficiency but shared their concerns about mathematical errors, inaccuracy, ethical issues, and its security.

Keywords: Summative assessment, probability and statistics, Artificial Intelligence (AI), ChatGPT, secondary mathematics teachers

Introduction

Since Shulman's seminal work on the pedagogical content knowledge (PCK) as a special amalgam of content and pedagogy needed for teaching (Shulman, 1986), the scholarship in teacher education has made a major shift from identifying teacher characteristics toward conceptualizing PCK, developing instruments to measure PCK, and designing teacher education programs to develop PCK. Besides the efforts made in the subject-matter specific PCK, researchers have identified the knowledge needed for teaching with technology in response to the emergence of new digital technology and the importance of technology competence. This specialized knowledge that teachers need to teach with technology is conceptualized as Technological, Pedagogical, and Content Knowledge (TPACK) and it is further specified into sub-domains (Misha & Koehler, 2006). After the publication of the *Second Handbook of Technological Pedagogical Content Knowledge (TPACK) for Educators* in 2016, we have experienced a rapid change and social demands of incorporating technology in teaching and learning over the past few years. Especially, the emergence of generative Artificial Intelligence (AI) and the public release of ChatGPT on November 30, 2022 have attracted more attention from educators, both for the benefits of using AI and concerns about using AI in teaching and learning.

Given these rapid changes using technology, we aimed to explore the experiences of secondary mathematics teachers to design probability and statistics assessment items using ChatGPT in Korea.

Because of the nature of grading on a curve and its importance for the college admission in Korea, many secondary mathematics teachers have experienced challenges in writing assessment items that can result in a desired item difficulty and item discrimination but can be differentiated from commercially available workbooks. The advancement of ChatGPT made us to wonder how secondary mathematics teachers might use ChatGPT to design assessment items, what experiences they have with ChatGPT, and whether ChatGPT might resolve their persistent challenges or create new issues to write assessment items. More specifically, this paper examines the following research questions:

1. What prompts do secondary mathematics teachers use to design probability and statistics assessment items in ChatGPT?
2. What probability and statistics assessment items do ChatGPT generate? To what extent do teachers have intentions to use these ChatGPT-generated assessment items?
3. What affordances and challenges do secondary mathematics teachers perceive in using ChatGPT to design probability and statistics assessment items?

Methods

Using a convenience sampling method (Pattern, 1990), we collected the data from 22 secondary mathematics teachers who enrolled in a three-credit graduate course in mathematics education in Korea. After the instructor's short introduction of ChatGPT, the teachers were instructed to have conversations with ChatGPT 3.5 approximately for 10-15 minutes to design probability and statistics assessment items for Grade 11. We asked teachers to design an assessment item for probability and statistics, a topic that tends to require less use of advanced mathematical expressions or graphical representations. After conversations with ChatGPT, teachers were asked to complete a short survey including a URL link for their ChatGPT conversations, one final ChatGPT-generated probability and statistics assessment item, their overall experiences with ChatGPT, and whether ChatGPT understood their questions, intentions, and feedback. Additionally, we asked whether they had any mathematical or pedagogical issues with ChatGPT, the benefits and challenges of using ChatGPT for assessment, and whether they had intentions to use ChatGPT-generated assessment items using a five-point Likert scale (1: strongly disagree and 5: strongly agree). Lastly, the teachers were asked to explain their perceived item difficulty generated by ChatGPT and to evaluate its appropriateness for their students.

First, we analyzed the prompts that teachers used in ChatGPT to design probability and statistics assessment items¹. In analyzing the teachers' ChatGPT conversations, it became clear that the strategy and skill of providing effective prompts are crucial for obtaining desirable responses. This is because ChatGPT generates responses based on the user's inputs or prompts. Some reports suggest that ChatGPT has often provided unvalidated and incorrect information (Einarsson et al., 2023). Considering that ChatGPT's mathematical skills are not as strong as in other areas (Frieder et al., 2023), it is important for the users, in this case secondary mathematics teachers, to employ the strategy of using effective prompts to get the desirable outcomes and assess the validity, accuracy, and credibility of ChatGPT's responses. In this paper, we categorized the teachers' prompts into four categories: 1) Specificity of the prompts to design an assessment item; 2) Identification of errors in ChatGPT's responses; 3) Quality of follow-up questions; 4) Evaluation of ChatGPT's responses. Table 1 illustrates our coding rubric to analyze teachers' prompts to ChatGPT.

¹ We asked the teachers to share a URL link for their ChatGPT conversations. At the time of analyzing the data, three links were invalid which resulted in 19 links for the analysis.

Table 1. Coding rubric for teachers' prompts to ChatGPT in designing assessment items

Category	Description	Scoring Rubric
1. Specificity of prompts	Does a teacher's prompt identify... <ul style="list-style-type: none"> • a topic? • a difficulty level? • a target grade-level? • a purpose? 	<ul style="list-style-type: none"> • 1: Address one of the sub-categories • 2: Address two of the sub-categories • 3: Address three of the sub-categories • 4: Address four of the sub-categories
2. Identifying an error in ChatGPT's responses	Does a teacher identify... <ul style="list-style-type: none"> • a contextual error? • a mathematical error? 	<ul style="list-style-type: none"> • 0: neither identified contextual nor mathematical error • 1: identified either contextual or mathematical error • 2: identified both contextual and mathematical error
3. Quality of follow-up questions	<ul style="list-style-type: none"> • Is a teacher's follow-up question based on interpretation or evaluation of the ChatGPT's responses? • Does a teacher's follow-up question further specify or challenge the ChatGPT-generated item? 	<ul style="list-style-type: none"> • 0: accept ChatGPT's responses without posing any follow-up questions or asking unrelated questions to ChatGPT's responses • 1: ask a follow-up question to clarify ChatGPT's responses or request simple modification (e.g., different format) • 2: ask a follow-up question to further specify ChatGPT's responses by adding conditions or related concepts • 3: ask a follow-up question to challenge the ChatGPT's responses or to provide a specific feedback to ChatGPT
4. Evaluating ChatGPT's responses	<ul style="list-style-type: none"> • Does a teacher evaluate ChatGPT's responses? 	<ul style="list-style-type: none"> • 0: no • 1: yes

To analyze the teachers' intentions to use ChatGPT-generated assessment items, we recoded disagree responses (1:strongly disagree and 2:somewhat disagree) to negative and recoded agree responses (4:somewhat agree and 5:strongly agree) to positive. For the affordances and challenges of using ChatGPT to design assessment items, we repeatedly read the teachers' responses and found themes emerged from their open responses using an inductive coding.

Results

RQ1. Teachers' Prompts Used in ChatGPT

The first analysis focused on determining whether the prompts included key details such as a topic, difficulty level, target grade-level, and purpose (Category 1). The analysis revealed that 14 out of 19 teachers (74%) received a score of 2 for this category. This score indicates that they provided two components of their desired assessment items, such as a broad topic and target grade-level, without

specifying the purpose for using the assessment item (e.g., "Please provide a problem about probability and statistics for high school students"). Only one of the 19 teachers provided all the four components, including the goal of an assessment item ("Please rephrase the problem conditions to assess if a student can derive the solution by using a normalization.")

Second, the analysis of identifying a contextual or mathematical error (Category 2) suggested that only one out of 19 teachers explicitly mentioned an error in ChatGPT's responses. For example, the teacher identified a mathematical error in ChatGPT's explanation of the suggested problem. Specifically, the teacher stated, "You mentioned that this problem will help students understand the concept of conditional probability, but your suggested problem is not related to the concept."

Third, an analysis of the quality of follow-up questions (Category 3) indicated that five teachers (26%) either accepted the ChatGPT's suggestions (e.g., "Good. Thank you!") or asked unrelated questions to the ChatGPT's responses (e.g., "Please give me another question"). Additionally, five teachers (26%) asked a simple parallel question (e.g., "Please give me a more difficult question"); five teachers (26%) further specified the ChatGPT's responses (e.g., "Please add a given that uses the concept of conditional probability"); while four teachers (21%) either provided specific revisions (e.g., "Please use a lower value for the number of plays to simplify the calculation") or challenge the ChatGPT's responses (e.g., "Isn't it possible for people with assigned seats to move among them?"). Figure 1 illustrates an example of one teacher's ChatGPT conversation. As shown in Figure 1, the teacher provides specific feedback on ChatGPT-generated assessment item. In response to this feedback, ChatGPT revised the initial assessment item, using a smaller number for matches and different probability values. However, solving the problem would still require a calculator.

<p>Teacher: Please create a probability problem related to soccer.</p> <p>ChatGPT: Of course, I'll create a probability problem related to soccer.</p> <p>Problem: A certain soccer team has played 20 matches in the league. The probability of this team winning a match is 40%, and the probability of a draw is 25%. The remainder is considered a loss. Answer the following questions about this. Q. What is the probability that this soccer team will win exactly 10 out of 20 matches? (Provide the calculation up to the second decimal place.)</p> <p>Teacher: The problem requires a formula that is too complex to solve without a calculator. please use a lower value for the number of plays to simplify the calculation </p> <p>ChatGPT: Great! Let's simplify the problem. Let's create a probability problem related to soccer with the number of matches reduced to 10.</p> <p>Problem: A certain soccer team has played 10 matches in the league. The probability of this team winning a match is 50%, and the probability of a draw is 20%. The rest is considered a loss. Please answer the following questions about this. Q. What is the probability that this soccer team will win exactly 5 out of 10 matches?</p>

Figure 1. An example of a teacher's ChatGPT conversation to design a probability item

This teacher also received high scores in other categories. Specifically, the teacher clarified the purpose of an assessment item and evaluated ChatGPT-generated assessment item in terms of its applicability in real-life situations and item difficulty.

Finally, we investigated if there was clear evidence that teachers evaluated responses from ChatGPT. The analysis indicated that seven out of 19 teachers (37%) assessed the correctness and validity of ChatGPT-generated assessment item. For example, one of the prompts stated that "Your problem necessitates overly complex calculations to determine the standard deviation."

RQ2. ChatGPT-generated assessment items and teachers' intentions to use them

For the second research question, we analyzed ChatGPT-generated assessment items that teachers came up with and their intentions to use these assessment items. As discussed in the previous section, teachers employed a wide range of prompts to design an assessment item which resulted in a wide range of assessment items. Table 1 illustrates four teachers' ChatGPT-generated probability and statistics assessment items with their intentions to use these items for the assessment (1: strongly disagree and 5: strongly agree).

Table 1. Examples of ChatGPT-generated assessment items and teachers' intention to use them

Assessment Item	Intention
<p>Teacher A's ChatGPT-generated assessment item: A class of 10 students sits at a round table with 12 chairs. Two specific students, A and B, must sit next to each other, and two other students, C and D, must sit facing each other. The remaining six students can sit at random. Find the number of cases in which the students are seated at the round table.</p>	4
<p>Teacher B's ChatGPT-generated assessment item: Jimin and Junho are playing a coin flip game. The coin is assumed to be a fair coin, and the probability of getting heads (H) and tails (T) is 1/2 each. The rules of the game are as follows:</p> <ol style="list-style-type: none"> 1. Jimin and Junho each flip a coin. 2. If it lands on heads (H), they win; if it lands on tails (T), they lose. 3. When both friends start flipping at the same time, Jimin flips first. <p>Answer the following questions:</p> <ol style="list-style-type: none"> 1. What is the probability that Jimin wins and what is the probability that Junho wins? 2. What is the probability that Jimin and Junho both win the first two coin tosses? 3. What is the probability that Jimin and Junho flip a total of 5 coins and win exactly 3 of them? 	2
<p>Teacher C's ChatGPT-generated assessment item: A soccer team has played 15 games in a league. The team has a 60% chance of winning a game and a 10% chance of drawing a game. The rest of the games are considered losses. Answer the following questions:</p> <ol style="list-style-type: none"> 1. What is the probability that the soccer team will win exactly 10 out of 15 games? 2. What is the probability that this soccer team will win at least 12 out of 15 games? 3. What is the probability that this soccer team will lose or tie at least 5 out of 15 games? 	4
<p>Teacher D's ChatGPT-generated assessment item: You have a store that sells goods and you have five different types of goods. You need to display these items in a row, but the store's shelves are circular, so the first and last items are next to each other, i.e., they are arranged in a circle. Find the number of cases in which the store displays the five products in a circle.</p>	2

Teacher A had an intention to use the ChatGPT-generated assessment item (rated 4 in a five-point Likert scale) but would like to revise the item by changing the number of chairs from 12 chairs to 10 chairs. However, Teacher A did not further explain the justification for this revision. Teacher B did not have an intention to use the ChatGPT-generated assessment item (rated 2 in a five-point Likert scale) because the ChatGPT-generated assessment item is simple, easy, and different from the expected item difficulty. Teacher B also pointed out that the assessment item has some inaccurate statements. Teacher C had an intention to use the ChatGPT-generated assessment item (rated 4 in a five-point Likert scale) but would like to revise the assessment item for the security reason. Teacher D did not have an intention to use the ChatGPT-generated assessment item (rated 2 in a five-point Likert scale) because it is quite similar to examples provided in textbooks, so it is more efficient for teachers to write their own assessment items.

In the survey, teachers have polarized responses about their intentions to use the ChatGPT-generated assessment items. Among 22 teachers, 10 teachers (45.5%) indicated that they had intentions to use the ChatGPT-generated assessment items, whereas the same number of teachers (45.5%) indicated that they did not have intentions to use the ChatGPT-generated assessment items. Two teachers (9%) responded neutrally about their intentions to use the ChatGPT-generated assessment items. Teachers who did not have intentions to use these assessment items explained that the ChatGPT-generated assessment items are too simple and easy, are not aligned well with their instruction, are not aligned with the curriculum, have inaccurate expressions, have incorrect answers, do not produce items with the intended item difficulties, are very similar to the textbooks, and are not quite different from items they can make. Two teachers wrote:

I have very little intention of using it; the difficulty level is very low; the solutions it presents may not be what the curriculum intends; and I don't think ChatGPT is up to the task of developing items that accurately assess the competencies that the curriculum wants students to develop.

I don't intend to use it yet, because the problems that ChatGPT creates are very similar to the textbooks. Even if I kept asking for new problems, they would just repeat the first problem with different numbers. In other words, I don't think I would use ChatGPT specifically because it only gives me typical problems from the textbooks or problem sets I have.

RQ3. Affordances and Challenges of using ChatGPT to design assessment items

In analyzing the teachers' survey responses, we found that the teachers perceived the affordances of using ChatGPT in terms of creativity (rich ideas, new types of problems, various contexts, extending teacher's limited thinking, and reducing the pains of creating items), efficiency (cost, time, and speed), specific difficult levels (easy or medium-level difficulty), specific type of assessment (performance assessment or formative assessment), specific type of items (multiple-choice items), providing solutions to the problems, diagnosis of errors, creating scoring rubrics, convenience, neutral (excluding teachers' own biases or preferences), producing anticipated solutions, and using ChatGPT for students' learning. Among these affordances, nine teachers (41%) identified creativity and five teachers (23%) identified efficiency in cost, time, and speed. However, the teachers identified its challenges as potential mathematical errors (e.g., incorrect answers, inaccurate solutions, and vague expressions), misalignment with curriculum, ethical issues (e.g., copyright issues, lack of information about sources), possibility of teacher's heavy reliance on ChatGPT, and security issues. In addition to these potential limitations, the teachers also identified that ChatGPT does not reflect their interactions with their students during their lessons, insufficient database (not quite different from commercially available workbooks or textbooks), teacher's intention for the assessment, the validity

of assessment created by ChatGPT, inappropriate item difficulty, and insufficient item discrimination. Among these limitations, eight teachers (36.4%) identified the potential mathematical errors and seven teachers (31.8%) concerned about the security of assessment because of its nature of open source.

Discussions

The purpose of this study was to explore secondary mathematics teachers' experiences of using ChatGPT to design probability and statistics assessment items. For this purpose, we analyzed secondary mathematics teachers' conversations with ChatGPT and their survey responses about overall experiences with ChatGPT, their intentions to use ChatGPT-generated assessment items, and affordances and challenges of using ChatGPT in designing assessment items in Korea. The first research question examined the prompts that the secondary mathematics teachers used to design probability and statistics assessment items. The results show that most teachers specified some of the key details of the assessment items and did not identify mathematical errors or issues in ChatGPT's responses. On the other hands, the quality of follow-up questions is widely ranged (five teachers received score 0, five teachers received score 1, five teachers received score 2, and four teachers received score 3) and seven teachers evaluated ChatGPT's responses. The second research question examined the ChatGPT-generated probability and statistics assessment items after the teachers' conversations with ChatGPT and their intentions to use these assessment items. The survey results showed that the teachers had polarized responses to their intentions to use. The third research question examined what secondary mathematics teachers perceived its affordances and challenges of using ChatGPT in designing probability and statistics assessment items. Teachers perceived that ChatGPT provides affordances such as creativity and efficiency but identified limitations of potential mathematical errors, ethical issues, misalignment with curriculum, misalignment with instruction, and security of assessment because of its nature of open source.

The results of this study provide implications as follows. First, the analysis of teachers' prompts to design assessment items using ChatGPT provides implications that teachers need to be exposed to different types of prompts they can use in ChatGPT and they need to explore that the prompts they employed would determine whether they could get the desired outcomes. Without specific prompts, challenges, or evaluation of ChatGPTs' responses, some teachers simply accepted the ChatGPT's responses or repeated the same prompts to ChatGPT. Especially, many teachers mentioned that ChatGPT generated too easy items but did not produce the items with the intended item difficulty and sufficient item discrimination. As an exception, one teacher, who rated the ChatGPT-generated item as difficult, mentioned that ChatGPT was able to produce a more difficult item once the teacher added more conditions to the initial ChatGPT-generated item. In this study, we did not provide specific prompts that teachers can use in ChatGPT because we aimed to explore the quality of prompts that teachers use. However, we might offer examples of different types of prompts that teachers can use and then analyze the frequency of using specific prompts or explore how teachers employ different types of prompts to the same ChatGPT's responses.

Second, the teachers perceived that ChatGPT was creative, efficient, and convenient as affordances but addressed mathematical errors, inaccuracy, incorrectness, and vagueness as its major challenges. It is interesting to observe that many teachers identified these mathematical issues in the survey but few of them actually addressed these concerns in their conversations with ChatGPT. Facing such issues in ChatGPT, teachers should be able to use their mathematical knowledge to critically examine the mathematical accuracy, correctness, and performance of ChatGPT and address them to ChatGPT. Another challenges of using ChatGPT in designing assessment items are the misalignment between ChatGPT-generated items and curriculum in Korea and misalignment between ChatGPT-generated

items and their own instruction. Because of its importance to use the specific grade-level mathematical vocabulary, concepts, or ideas outlined in the curriculum, it needs to be further examined whether ChatGPT understands or has an access to the specific curriculum materials and grade-level expectations in each country.

Lastly, writing assessment items is very stressful for teachers in Korea because students and parents are very sensitive to the assessment items and often complain if there are any vagueness, errors, or issues in the assessment and if there are any similarities between assessment items and commercially workbook items. Because the security of assessment items is one of the most important issues for teachers in Korea, they might not use the ChatGPT-generated assessment items without any major revisions or modifications. However, as three teachers commented in the survey, it would be a great learning opportunity for students to design or discuss a mathematical problem using ChatGPT. Because teachers experienced that ChatGPT often produced inaccurate, incorrect, and vague mathematical ideas, they would like to use ChatGPT-generated items as formative assessment for their students to explore whether students can detect any mathematical errors or issues in ChatGPT's responses.

References

- Einarsson, H., Lund, S. H., & Jónsdóttir, A. H. (2023). Application of ChatGPT for automated problem reframing across academic domains. *Computers and Education: Artificial Intelligence*, 100194.
- Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). Mathematical capabilities of ChatGPT. arXiv preprint, arXiv :2301 .13867.
- Koehler, M. J., Shin, T. S., & Mishra, P. (2012). How do we measure TPACK? Let me count the ways. In R. N. Ronau, C. R. Rakes, & M. L. Niess (Eds.), *Educational technology, teacher knowledge, and classroom impact: A research handbook on frameworks and approaches* (pp. 16-31). IGI Global.
- Mishra, P., & Koehler, M.J. (2006). Technological pedagogical content knowledge: A framework for integrating technology in teacher knowledge. *Teachers College Record*, 108(6), 1017-1054.
- Patton, M.Q. (1990). *Qualitative Evaluation and Research Methods*, (2nd ed.), Newbury Park, CA: Sage
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15, 4–14.
- Scherer, R., Siddiq, F., & Tondeur, J. (2019). The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Computers & Education*, 128, 13-35.