When the Tutor Becomes the Student: Design and Evaluation of Efficient Scenario-based Lessons for Tutors

Danielle R. Thomas
Carnegie Mellon University
drthomas@cmu.edu

Xinyu Yang
Carnegie Mellon University
scinkoyang@gmail.com

Shivang Gupta
Carnegie Mellon University
shivangg@andrew.cmu.edu

Adetunji Adeniran
Carnegie Mellon University
adetunja@andrew.cmu.edu

Elizabeth A. Mclaughlin
Carnegie Mellon University
mimim@andrew.cmu.edu

Kenneth R. Koedinger
Carnegie Mellon University
koedinger@cmu.edu

ABSTRACT

Tutoring is among the most impactful educational influences on student achievement, with perhaps the greatest promise of combating student learning loss. Due to its high impact, organizations are rapidly developing tutoring programs and discovering a common problem- a shortage of qualified, experienced tutors. This mixed methods investigation focuses on the impact of short (~15 min.), online lessons in which tutors participate in situational judgment tests based on everyday tutoring scenarios. We developed three lessons on strategies for supporting student self-efficacy and motivation and tested them with 80 tutors from a national, online tutoring organization. Using a mixed-effects logistic regression model, we found a statistically significant learning effect indicating tutors performed about 20% higher post-instruction than pre-instruction (β = 0.811, $p < 0.01$). Tutors scored ~30% better on selected compared to constructed responses at posttest with evidence that tutors are learning from selected-response questions alone. Learning analytics and qualitative feedback suggest future design modifications for larger scale deployment, such as creating more authentically challenging selected-response options, capturing common misconceptions using learnersourced data, and varying modalities of scenario delivery with the aim of maintaining learning gains while reducing time and effort for tutor participants and trainers.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI); • **Applied computing** → Education; Computer-managed instruction.

KEYWORDS

Tutoring, Learnersourcing, Scenario-based learning, Design-based research

ACM Reference Format:
Danielle R. Thomas, Xinyu Yang, Shivang Gupta, Adetunji Adeniran, Elizabeth A. Mclaughlin, and Kenneth R. Koedinger. 2023. When the Tutor Becomes the Student: Design and Evaluation of Efficient Scenario-based Lessons for Tutors. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3576050.3576089

LAK 2023, March 13–17, 2023, Arlington, TX, USA

ACM ISBN 978-1-4503-9865-7/23/03.
https://doi.org/10.1145/3576050.3576089

1 INTRODUCTION

Multiple studies have shown that tutoring can improve student achievement and demonstrates considerable promise in narrowing opportunity gaps [27, 28]. Due to its known impact, personalized instruction via tutoring or mentoring for providing additional support to students is in increasing demand [20]. Although individualized instruction has proven to be effective, few students have access due to high cost and lack of resources, particularly marginalized students (i.e., Black, Hispanic, students from low socioeconomic backgrounds) further contributing to educational inequities [4, 20, 29]. In response to the unequal access to tutoring across student populations, many for-profit and nonprofit tutoring organizations are springing up across the country shining light on one problem—the shortage of trained and qualified tutors ready to support the nation's students. Trained tutors engage in more behaviors aligned with student learning, such as focusing on knowledge-building activities and attending to motivation [13]. Online tutors are particularly effective when trained on building relationships and rapport with students [23]. The pool of qualified tutors is small with a disproportionate number of available tutors lacking both experience and skills to be successful. With schools desperate to fill the increasing need for tutoring support, they are reaching out to non-certified teachers, undergraduate students, and volunteers [14]. In addition, tutor training often focuses on content-specific academic support and does not tend to other important research-based principles, such as social-emotional learning, tutor-student relationship building, and effective tutoring strategies related to pedagogy. This work focuses on the impact of brief, scenario-based lessons that introduce tutors to situational judgment tests related to student's motivational challenges. We investigate the changes in tutors' perceived learning and actual competence caused by lesson participation.

1.1 Tutoring Competence & Learning by Doing

There is considerable research evidence demonstrating that the most successful tutoring organizations attend to both student's academic and socio-motivational needs allowing for relationship building and active feedback [4, 16]. Our lessons focus on improving tutor-student interaction by instructing human tutors to effectively apply instructional strategies to increase student engagement and self-efficacy.

The consensus among theorists and devoted Dewey enthusiasts is that active learning is preferred over passive forms of learning, such as reading text and watching lectures (e.g., [8, 26]). Learning by doing requires the application of skills which model the needs of the real world [21] for the purpose of transferring new knowledge to similar experiences that trigger recall [30]. Authentic scenarios, coming from real-life tutoring situations, are used for tutors to practice learning by doing with the intention that the learning transfers to real-life tutoring environments.

1.2 Related Work

Scenario-based learning is an instructional approach which emphasizes fostering meaningful learning by contextualizing the learning activities into authentic contexts [35]. In addition, scenario-based models can expedite expertise that is often acquired through situational-based activities [6] and support active learning with numerous opportunities for feedback [2]. These learning experiences have been reported to be effective in multiple domains for different learners: nursing students' and patient safety skills [36]; high school students development of prospective thinking skills [1]; pre-service teacher professional development on learning strategies [15]. Scenario-based models in an online setting, such as digital simulations show promise as a method providing novice teachers practice in a low-risk educational setting [31]. In addition, scenario-based learning transfers to tutor learning environments by providing situational experiences to inexperienced tutors [3, 5].

The perceived larger impact of open-ended activities, such as constructed-response questions, compared to closed-ended activities, such as selected-response or multiple-choice questions, has been debated among researchers [33]. Recent studies have challenged the long-held belief that multiple-choice questions cannot foster the same level of rigor prevalent during open-response questioning. [33] suggests theoretical situations where multiple-choice questions can match or even beat the instructional impact of open-response questions while supporting learning at scale without sacrificing learning by doing. Our work plans to examine tutor performance on selected-response and constructed-response questions to assess learning quality for future lesson design iterations and scale.

One method of strengthening question authenticity and quality while simultaneously improving tutor learning is through learnersourcing, an emergent strategy of engaging learners with instructional content while concurrently using their contributions and inputs to iteratively improve the content and learner experience [18, 19]. An example of learnersourcing in context of this present work is the use of tutor responses to open-response questions to create more authentically challenging multiple-choice options in later iterations of the same questions, enabling rapid scaling and instant feedback (modeled from [34]). Explanations of the use of learnersourcing, particularly experiences learned from design and development within adaptive educational systems, are underrepresented within the learning analytics community [18]. Active learning through learnersourcing of multiple-choice questions [7, 18] is of particular interest to this work.

2 SCENARIO-BASED LESSON DEVELOPMENT

Our brief scenario-based lessons are strategically designed using the learning by doing approach that provides actionable feedback and has tutors apply what they learned to a similar situation. There are two goals regarding our pilot lessons: 1) to develop and evaluate training for tutors supporting student self-efficacy and 2) to demonstrate and refine an interactive design process by evaluating tutor learning gains. The latter goal has tutors responding to a training scenario (left oval in Figure 1) (i.e., a student struggling to stay motivated) by asking them to *predict* and *explain* how to best respond (steps 1 & 2), observe the expert-based approach with feedback (step 3), and, lastly, *explain* their thoughts and agreements with the research-recommended approach (step 4).

Learners apply what they learned in the training scenario (pretest) to the following transfer scenario (posttest) (right oval in Figure 1) using the same cyclic learning process (steps 5-8) to gauge learning transference and, ultimately, the tutor's learning gain [5]. This modified predict-observe-explain (POE) approach is theoretically connected to Gibbs' Reflective Cycle, a cyclical instructional model providing structure for learning by doing to individual learning experiences [11]. Aside from determining learning gain, an additional purpose of this research is to collect learnersourced data from constructed responses to create more authentic and effective selected-response options for both less- and more-desired responses (modeled from [34]).

2.1 Research Questions

The following research questions were developed to determine the effectiveness of the lessons on tutor learning and provide insight into tutor's individual learning perceptions.

RQ1: Are scenario-based lessons effective in teaching tutors new strategies and skills?

RQ2: How do tutors' perceptions of their own learning vary with different demographics (i.e., race, gender, age) and self-reported experience levels?

RQ3: Are tutor's self-evaluations of experience level consistent with their actual performance on the lessons?

RQ4: How can lesson design be iteratively improved based on participant feedback and learning analytics?

RQ5: How does tutor performance compare between selected-response questions that feature learnersourced options compared to those without learnersourced options?

3 METHODS

Three scenario-based lessons were created and designed to align with the expressed needs of the tutoring organization.

- *Giving Effective Praise*- Tutors practice responding to students to increase their engagement and motivation to learn. Tutors apply strategies by responding to students by giving effective feedback and praise.
- *Learning What Students Know*- Tutors practice meeting students where they are in their learning by determining what they know and what they need to learn. Tutors apply strategies to assess student's prior knowledge.
- *Responding to Students Errors*- Tutors practice responding to a student making an error to increase motivation and engagement. Tutors apply strategies by effectively responding to a student making an error.

Within each of the lessons, self-reported demographic information was collected regarding tutor race, gender, age, and tutoring

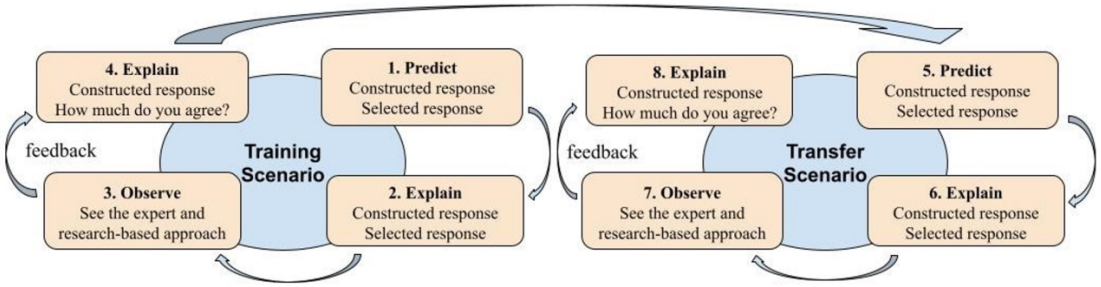


Figure 1: The modified POE cycle for the training and transfer scenarios.

experience level. The latter was determined using a 5-point Likert scale with a rating of 1 indicating little to no experience, or novice tutor, and 5 indicating an expert tutor. In addition, measures of the tutor's self-perceptions of their learning related to the lesson objectives were reported using a 5-point Likert scale ranging from 1-strongly disagree to 5-strongly agree. Using this same Likert scale, four self-assessment questions were asked upon completion of each lesson: *This module was valuable.; I can apply what I have learned from this module to my tutoring.* The other two self-assessment questions were specific to the lesson objective. For example, for the *Responding to Errors* lesson the two lesson-specific questions were: *I know how to best respond to a student making an error; I can apply strategies to increase student's motivation when they make an error.*

3.1 Lesson Delivery & Validity

The lessons were delivered through Google Forms and initial use testing of the lessons among 22 participants showed that the average time taken to complete was around 14 minutes. There were a total of 80 participants, unpaid volunteers within the tutoring organization, who completed between one and three of the lessons. The number of participants per lesson were 65, 71, and 57, for *Giving Effective Praise*, *Learning What Students Know*, and *Responding to Student Errors*, respectively. All three lessons were co-created with researchers and the director of the tutoring organization working in collaboration to ensure the lesson operationalization is an accurate translation of the construct being taught and assessed. Strong face validity was achieved by using tutoring scenarios handpicked by trainers at the tutoring organization based on frequent, real-life experiences. Both parties also iteratively co-designed lessons.

3.2 Constructed Response Coding

The constructed-response questions were open coded with most-desired, or "correct" tutor responses coded as "1" and less-desired, or "incorrect" responses receiving a "0". At least two experienced researchers coded participant responses to determine inter-rater reliability for all three lessons. Responses were determined to be correct if understanding of lesson objectives and the research-recommended approach was evident from the response. The constructed-response coding schema for each of the three lessons is described below.

In the *Giving Effective Praise* lesson, the scenario presented featured a student struggling to persevere on an assignment. The tutor's responses had to align with the following research-based elements of effective praise. As stated, praise should be: 1) immediate, earned, and truthful, 2) specific by giving details on what the student did well, 3) genuine and not repeated often, such as just saying "great job.", and 4) focused on the learning process, not student ability. Correct responses had to be encouraging, positive, and indicate acknowledgement of the student's focus on the learning process. Table 1 below highlights some sample tutor responses to a struggling student in both scenarios with an explanation of the reasoning behind the coding of "correct" and "incorrect" responses.

In the *Learning What Students Know* lesson, a tutor had to respond to the given scenario in which a student is given a math problem they do not know how to solve. The tutor needs to determine the student's prior knowledge, so they can give appropriate instructions building on what the student already knows. The tutor's responses had to align with the following research-shown elements of assessing knowledge: 1) asking the student to explain what they know or have already done, 2) guiding the conversation to catch the student's misconceptions, and 3) supporting productive struggling, namely guiding the student to find the answer themselves. If the tutor asks a question, it should be open and not content-specific to avoid making assumptions about the student's prior knowledge. Table 2 below highlights some sample tutor responses.

In the *Responding to Student's Error* lesson, a tutor responds to a given scenario of a student making a mistake in the problem solution. The tutor's responses had to align with the following research-shown elements of appropriate tutor reactions: 1) praising for the attempt or effort, 2) indirectly drawing the student's attention to the mistake, and 3) guiding the student to self-correct. Any response explicitly pointing out the student's error or telling the student what to do is considered incorrect. Table 3 below highlights some sample tutor responses to a struggling student with coding rationale.

Coding of the *explanation* constructed responses was performed similarly using corresponding rationale and ensuring alignment with the research-recommended approach. For participants to get the constructed-response questions prompting for *explanation* correct, tutors had to get the corresponding selected-response question correct and explain their reasoning using the rationale of the research-recommended approach.

4 RESULTS

The constructed-response questions were coded manually by two experienced researchers using the coding scheme described previously. The inter-rater reliability using Cohen's Kappa across all

Table 1: Sample tutor responses for prediction of the best approach for *Giving Effective Praise* with coding rationale.

Tutor Response	Rationale
You did a great job on that problem. Well done.	Incorrect: The response is positive and sincere; however, the response does not focus on the process or give praise for specific student actions.
It was difficult but you persevered and succeeded. Such grit is an important life skill and I'm proud of what you accomplished, and you should be proud, too.	Correct: The response is positive, sincere, and praises the student for persevering. Acknowledgement of the student for working hard and the process of learning is evident.
You have done a good job so far. It is normal to feel challenged at this stage. Keep going.	Correct: This response is slightly nuanced in that there is little emphasis on positivity and uses the generic "good job" phrase. However, it focuses on persevering despite being challenged and for this reason is deemed a more-desired response.

Table 2: Sample tutor responses for prediction of the best approach for *Learning What Students Know* with coding rationale.

Tutor Response	Rationale
Cindy, are you familiar with the types of triangles and the relationships of the triangle's sides and angles?	Incorrect: This is a content-specific question, and it is a "yes or no" question. For both of these reasons, it is not a good example of how to determine the student's knowledge.
No problem, let's try to do this together.	Incorrect: Although encouraging the response does not assist with determining a student's prior knowledge.
What do you know about the triangle?	Correct: This response is an "open" question asking a student what they know and is not specific to a certain knowledge component.

Table 3: Sample tutor responses for prediction of the best approach for *Responding to Student's Errors* with coding rationale.

Tutor Response	Rationale
Good effort! You got most of the addition right but there is a small problem at the start. What was your first step when you solved this problem?	Incorrect: Despite the response praising the student's effort and asking them to walk through the steps, it still highlights the student's "problem," or mistake.
Thank you for writing the problem and your answer so I can see your work. Let me show you some examples that show how to add numbers that add up to a number greater than 10.	Incorrect: Though encouraging, giving the student more examples without clearing up the misconception may not be helpful.
Great effort so far, can you explain to me how you approached this problem?	Correct: Praising the student's effort and asking them to explain how they attempted to solve the problem will aid the student in finding their mistake.

three lessons demonstrated substantial agreement: *Giving Effective Praise*- 0.85, *Learning What Students Know*- 0.72, and *Responding to Errors*- 0.88, respectively. Each lesson contained two scenarios transposing training and transfer scenarios to create Form A and Form B. Scenarios were individually identified by the name of a tutor or student in the scenario. The lessons with corresponding scenarios are identified as follows: *Giving Effective Praise* (i.e., Carla, Kevin), *Learning What Students Know* (i.e., Cindy, Roberto), *Responding to Student's Errors* (i.e., Lucy, Kanye).

4.1 Q1: Are scenario-based lessons effective in teaching tutors new strategies and skills?

The maximum score for the pre- and posttest was four points each with the following descriptive statistics (by percentage) shown in Table 4. The mean pre- to posttest differences suggestive of learning were *Giving Effective Praise*-17.6%, *Learning What Students Know*-*13.4%*, and *Responding to Student's Errors*-17.6%. *Giving Effective Praise* was the easiest lesson with average performance of 88.8% while demonstrating the lowest variation from the mean (SD = 18.2%).

The pre- to posttest scores for all lessons and each lesson by scenario are shown in Figures 2 and 3, respectively, illustrating the differences between selected- and constructed-response question performance. The high scores for many selected response scenarios, particularly at pretest (i.e., Lucy Scenario in *Responding to Student's Errors*), make it difficult to measure tutor learning gains due to ceiling effects. We hypothesized the selected-response questions were "too easy" in that the incorrect, or less desired, selected-response options were obviously incorrect. This is evidenced by the overall high performance on many selected-response questions. Figure 2 illustrates tutor performance on each lesson from pre- to posttest

Table 4: Descriptive statistics for each lesson shown as percentages.

Lesson	Pretest Mean (SD)	Posttest Mean (SD)
Giving Effective Praise	71.2 (23.5)	88.8 (18.2)
Learning What Students Know	58.1 (33.2)	71.5 (27.8)
Responding to Student's Errors	61.8 (26.7)	79.4 (23.7)

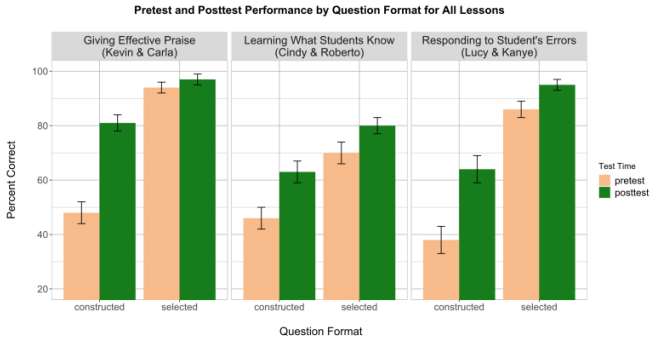


Figure 2: Average percent correct at pre- and posttest for constructed and selected responses. Tutors scored higher on selected responses but demonstrated larger learning gains on constructed-response questions.

for constructed and selected response problem formats. Notice the higher scores on the selected responses in relation to constructed responses. Larger standard deviation at pretest (see Table 4) for all lessons indicate a larger dispersion of tutor scores suggesting a wider variability of tutor ability at pretest. Breaking down lessons by scenario (see Figure 3a-c) shows differences in difficulty by scenario, however, learning gain was apparent for either question format.

4.1.1 Multilevel Logistic Regression. We performed a multilevel logistic regression to determine possible predictors of posttest score (*Outcome Score*). The *Outcome Score* for the logistic model was 1 or 0 indicating whether or not the individual tutor, or trainee, got the question correct, respectively. The generalized mixed model fit by maximum likelihood (Laplace Approximation), or 'glmerMod,' in R was used with *Scenario* nested into *Lesson* in a hierarchical structure with *Participant* indicating individual trainee (see Equation 1). The predictor variables designated as fixed effects were *Test time* (pretest = 0, posttest = 1), *Question type* (*explain* = 0, *predict* = 1), and *Question format*, (constructed response = 0, selected response = 1). *Lesson*, *Scenario*, and *Participant* were identified as random effects. Main effects were found for *Test time* ($p<0.001$) and *Question format* ($p<0.001$) and an interaction was found for test time and question type. Table 5 shows the results (number of observations = 1544; groups = 80; scenarios = 6; lessons = 3).

$$OutcomeScore \sim (1|Participant) + (1|`Lesson`/`Scenario`) + Test-time * Question-type * Question-format \quad (1)$$

There was a significant effect of test time ($\beta = 0.811$, $p < 0.001$) indicating that posttest scores are significantly higher than pretest scores. A follow-up ANOVA test indicates the model with the *Test time* factor is significantly better than one without, $F(1, 79) = 38.51$, $p < .01$. The $\beta = 0.811$ for *Test time* indicates that scores on the posttest are 0.811 log odds higher than scores on the pretest. For example, when the pretest score is 50% (0.00 log odds), this value corresponds to a predicted posttest score of ~70% which is computed by converting log odds of 0.811. In addition, tutors scored significantly better on selected- compared to constructed-response questions (see Figure 4a) performing ~30% better on selected responses ($\beta = 2.42$, $p < 2e\text{-}16$). Upon posttest, there were larger gains on tutors' ability to *predict* than to *explain* the best response to a given scenario ($\beta = 0.847$, $p < 0.012$). Constructed response questions prompting tutors

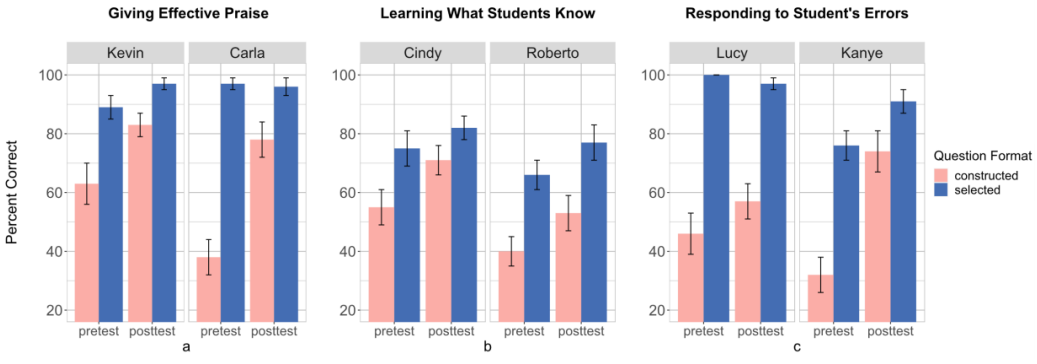


Figure 3: Average tutor scores on pre- and posttests comparing constructed and selected response performance for both scenarios within (a) *Giving Effective Praise*, (b) *Learning What Students Know*, (c) *Responding to Student's Errors*. High scores for certain selected responses for (a) and (c), particularly at pretest, suggest of possible ceiling effects.

Table 5: Factors in the logistic regression model predicting posttest score.

Predictors	Estimate	SE	Pr(>z)
(intercept)	-0.0454	0.3185	0.886656
Test time (posttest = 1)	0.8109	0.2338	0.000522 ***
Question format (selected = 1)	2.4187	0.2863	< 2e-16 ***
Question type (*predict* = 1)	-0.3715	0.2305	0.106963
Test time: Question type	0.8468	0.3352	0.011538 *
Test time: Question format	-0.2148	-0.496	0.619778
Question type: Question format	-0.3622	0.3749	0.334005
Test time: Question type: Question format	-0.8199	0.5755	0.154271

*** p < 0.001, ** p < 0.01, * p < 0.05

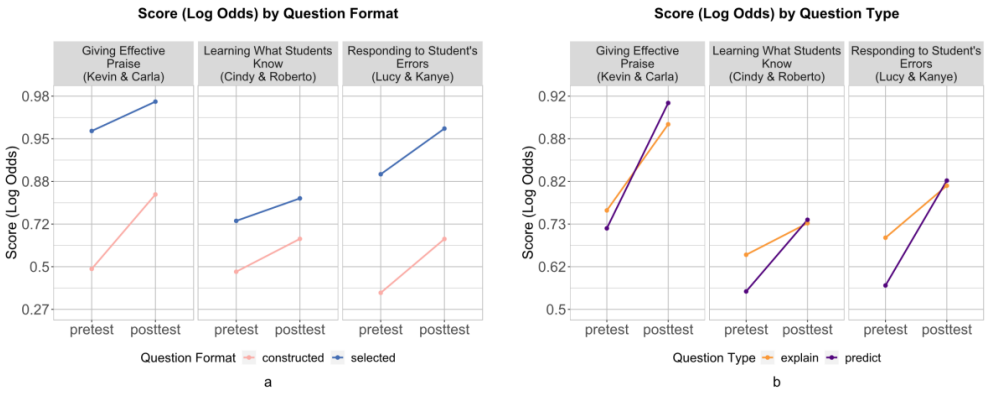


Figure 4: Score probability in log odds at pre- and posttest for responses by question format (a) and by question type (b). Across all lessons tutors had lower baseline knowledge on how to *predict* the best response, particularly for constructed responses, however, demonstrating substantial learning gains at posttest for both question types.

to *predict* demonstrated the largest pre- to posttest gains (see Figure 4b) suggestive of two interpretations: tutors' baseline knowledge of determining or *predicting* how to best respond is lower indicative of exposure to a novel situation and, possibly, we need to do a better job "*explaining*" the research-recommended approach. There was no statistically significant interaction between 1) test time and question format, 2) question type and question format, and 3) test time, question type, and question format.

The addition of lesson type as a fixed effect to determine possible interaction of tutors performing differently across lessons did not yield statistically significant results for any predictors except for within *Learning What Students Know* and selected responses (β = -2.38, p = .003) and a three-way interaction between tutor performance with tutor's ability to *predict* on selected-response questions within the *Giving Effective Praise* lesson (β = 2.69, p = .002).

4.2 Q2: How do tutors' perceptions of their own learning vary with different demographics (i.e., race, gender, age) and self-reported experience levels?

Tutor's self-reported perceptions of learning were determined by adding the scores among the four self-reported perceptions of learning items (Likert scaled 1-5) to create the tutor's Perceptions of Learning Score (PLS). The maximum possible PLS was 20. Then the PLS was used to determine correlation between tutor demographics (i.e., self-reported age, gender, and race) using nominal scales. Experience level was reported on a 1-5 Likert-type scale with beginning tutor-1 and expert tutor-5. Participant demographics by observation are shown in Table 6. Tutor participants completed anywhere from one to all three lessons with demographics self-reported each lesson. Tutors reported consistent demographics and experience levels across lessons with minor discrepancies. Slightly more men participated (51%) than women and other genders. The majority of tutors self-identified as White (52%) though there were a substantial number of Asian (21%) and some Black tutors (5%).

Using Pearson's correlation, we found a statistically significant correlation between PLS and gender, with women predicted to perceive their learning higher than other genders, β = 0.66, $t(187)$ = 2.07, p < .05. Associations between PLS and race, age, or experience level were not found to be statistically significant. However, the predicted perceptions of learning for participants ranging 25-34 and 51-64 years of age was greatly positive at β = 0.75 and β = 0.31, $t(186)$, while participants ranging 35-50 years of age were predicted to perceive their learning grossly less than the total participants, β = -0.65, $t(186)$. Although this is an interesting finding, it was not statistically significant.

Table 6: Self-reported age and experience level by observation.

Group	Demographics	Frequency (n)	Group	Demographics	Frequency (n)
Age	18-24	27 (14%)	Experience level	1- beginning tutor	11 (6%)
	25-34	26 (14%)		2	39 (20%)
	35-50	30 (16%)		3	64 (33%)
	51-64	68 (35%)		4	65 (34%)
	65 and up	40 (21%)		5- expert tutor	14 (7%)

In general, tutors' PLS scores indicated overall satisfaction with the lessons, and more than half of the participants provided optional comments at the conclusion of the lesson. Tutors reported the lesson was "too easy," particularly among selected-response questions; tutors found value in the lesson; and tutors expressed an increase in learning and confidence, as evidenced by the following tutor comments, respectively: "*The direction seems obvious, so not much new learning"* (*Giving Effective Praise)*; "*Thank[s] for this training! I feel more confident that I can help the students that I am tutoring" (Giving Effective Praise)*; "*Most valuable of the three lessons as it is not obvious to refrain from saying anything about the student being wrong" (Responding to Students Errors*).

Several comments and feedback from participants were noted and used as qualitative data for purposes of design modification and optimization, such as "*The triangle question was too complex for me"* (*Learning What Students Know*) informing us that some tutors may not have the content-level knowledge for certain scenarios. One particularly astute comment helped researchers see the need for using learnersourcing to generate more genuine phrases: "*It would help to provide some other word choices to praise student effort. . ."* (*Responding to Students Errors*).

4.3 Q3: Are tutor's self-evaluations of experience level consistent with their actual performance on the scenario-based lessons?

To determine if tutors reporting a high level of experience perform better, and conversely, if tutors reporting a low level of experience perform below similar tutors, we correlated participant performance on all three lessons with performance metrics. We found the correlation to be low (and even trending negative) with a Pearson correlation coefficient between tutor experience level and pretest, r = -0.05, and correlation between experience level and posttest, r = -0.17, $p < 0.05$.

4.4 Q4: How can lesson design be iteratively improved based on participant feedback and learning analytics?

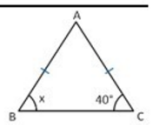
We identified several design modifications based on tutor feedback and learning performance which will be considered in future lesson iterations (see *Implications for Future Work*). A common theme among feedback was that the lesson was "important" and "valuable" indicated by tutors stating, *"Again this is very important and I hope most tutors already do this,"* (*Giving Effective Praise*). In addition, a tutor expressed the following: "*I have been using similar strategies during my tutoring. This training module has helped me tremendously. . ."* (*Learning What Students Know*). This feedback reinforced the utility value of our lesson content and assured us the particular lessons used in the pilot were of importance to tutors. However, there was conflicting opinions on new learning expressed by tutors with some stating they already apply the research-recommended approach to their tutoring: *"I feel that I already knew these lessons, but I will now keep them in mind more"* (*Giving Effective Praise*). Other tutors shared the opposite experience suggesting new learning: "*I now know some of the pitfalls that I have fallen into during previous tutoring sessions. I will quickly recognize these snares in the future"* (*Learning What Students Know*).

In general, selected-response questions and corresponding answer options were too easy as indicated by participant feedback and tutor performance metrics. Some participants provided their thoughts on the selected-response options within the subsequent open-response explanation prompt. Such feedback will be used to optimize selected-response options by creating more authentic responses for both less-desired (incorrect) or most-desired (correct) response options: "*It feels the most sincere, but I'm not sure any of the options are all that great. This student sounds like they need a quick break and some water.* (*Giving Effective Praise*); "*Option[s] 1 and 2 are false, and students shouldn't be misled; they should receive praise. Option 3 is chastising Kevin about quitting. Option 4 is true praise and encourages Kevin to keep trying in the future, rather than [saying] don't quit." (Giving Effective Praise).* Last, some tutors chose the wrong, or "less-desired" option in the *predict* selected response (i.e., *"To begin, you need to use the order of operations. First, you. . . What should you do next. . ."*) but their explanation or rationale was correct. For example, in *Learning What Students Know*: *"This not only praises the student, but this also finds the baseline knowledge where the student is familiar with and also recommends the student to try on their own"; "Because I explained [to] him the concept and I emphasized what he did, and I want to check what he knows about the next step."*

4.5 Q5: How does tutor performance compare between selected-response questions that feature learnersourced options compared to those without learnersourced options?

Both *Learning What Students Know* and *Responding to Student's Errors* each contained one scenario with learnersourced, "incorrect"

You're matched with a student named Cindy. She is having trouble solving a geometry problem dealing with triangles. She draws the following diagram displaying a triangle (shown right). You greet Cindy by saying, "Hi Cindy! What can I help you with?" She replies she needs help determining the measure of angle x.



1. Which of the following tutor's responses below do you think effectively begins Cindy's session?

Option 1: Cindy, this is an isosceles triangle which has two congruent sides and angles. Given this hint, can you solve the problem now?

Option 2: To begin, what type of triangle is this? This information will be very helpful in solving the problem.

Option 3: Let's talk about how to begin, Cindy. What do you know about the triangle?

Option 4: So, you start. What do you think is the answer? How many degrees is angle x?

Figure 5: The Cindy Scenario within *Learning What Students Know* displaying the correct answer, Option 3 (shown in bold). The remaining three, "incorrect" response options were learnersourced from chat log data.

Table 7: Tutor error rates (as percentages) for selected-response questions containing learnersourced and non-learnersourced answer options. Notice questions containing learnersourced answer options are considerably easier than non-learnersourced suggesting we did not capture common misconceptions or distractors, but merely blatantly "incorrect" responses.

Lesson with Scenario	Are the less-desired options learnersourced?	Error rate		
		Pretest	Posttest	Total
Learning What Students Know				
Cindy	Yes	26.7%	14.6%	19.7%
Roberto	No	46.3%	36.7%	42.2%
Responding to Student's Errors				
Lucy	Yes	0%	2.9%	1.8%
Kanye	No	38.2%	17.4%	29.8%

options (i.e., Cindy and Lucy scenarios, respectively) and the other corresponding scenario containing all researcher-created options for selected responses prompting tutors to *predict* the best response. In *Learning What Students Know*, all three "incorrect" selected-response options within the Cindy Scenario were learnersourced with tutors' less-desired responses coming from chat log data provided by the online tutoring organization. Figure 5 below displays the correct answer (Option 3, in bold) and the learnersourced, incorrect responses. Similarly in *Responding to Student's Errors*, all three "incorrect," selected-response options within the Lucy Scenario were created from the same chat log data.

Comparing tutor error rates for learnersourced and non-learnersourced selected responses for *Learning What Students Know* and *Responding to Student's Errors* (see Table 7), we conclude that the learnersourced selected responses were chosen at a much lower frequency (19.7% and 1.8%, respectively) than the non-learnersourced options (42.2% and 29.8%, respectively). We conclude the learnersourced options were considerably easier than the non-learnersourced options — contradictory to our initial hypothesis. The tutor organization provided chat log data of exemplar, "bad" responses which were used as "incorrect" response options. The "incorrect," or less-desired options were not deemed as common misconceptions, frequent errors, or even distractors but examples of blatantly wrong responses. We posit that having examples of real-life tutoring situations with obviously wrong responses aided researchers with scenario development, however, did not assist with capturing common misconceptions or distractors. An example of common misconceptions determined from this pilot analysis is described within the *Discussion*.

5 DISCUSSION

5.1 Tutors are learning from brief, scenario-based lessons

Overall, there was a statistically significant learning gain in comparing participant performance across all three lessons suggestive that tutors are learning from our brief, scenario-based lessons scoring 20% higher on posttest compared to pretest. A key question for us is whether we can get as effective assessment and learning from just the automatically-graded, selected-response questions. The results indicate tutors may be learning from just the selected responses, $F(1, 79) = 7.24$, $p > .01$, with tutors predicted to demonstrate learning gains on close-ended question types alone. In particular, among selected responses tutors demonstrated moderate and statistically significant gains in their ability to *predict* the best approach ($\beta = 1.99$, $p < 0.001$). We posit we need to do a better job of explaining the research-recommended approach to improve tutors' performance on selected responses prompting tutors to *explain*.

Nevertheless, there is room for improvement. For example, the selected-response questions in the *Giving Effective Praise* were too easy for many tutors with tutors achieving an average of 94% at pretest and 97% at posttest. The majority of tutors scoring high at pretest hinders tutor learning due to ceiling effects and suggests more challenging scenarios are warranted. Increasing scenario difficulty with better "incorrect," selected-response options will increase tutor learning gain even more. Tutors scored approximately 30% better on selected compared to constructed responses at posttest. Tutors found *predicting* the best response easier than *explaining* the preferred approach aligning with research citing generating

a high-quality explanation for why a response is correct is more difficult than stating the best approach [12].

5.2 Women perceive their learning higher than others

The biggest takeaway was the statistically significant and thought-provoking association between gender and learning perceptions with women perceiving their learning higher than other genders, $\beta = 0.66$, $t(187) = 2.07$, $p < .05$. Some hypotheses for why this occurred include a tendency for women to refrain from providing undesirable feedback due to higher social desirability [22] or simply an effect of selection bias. Further research is needed to uncover this finding. Regarding tutors' reflections about their learning, provided as survey feedback at the end of each lesson, there was little to no correlation among tutors' perceptions of learning and race. As more tutors complete our lessons, further analysis can be conducted with a larger sample size.

5.3 Tutors may not be great at assessing their own experience level

The lack of correlation between tutor's experience level and lesson performance indicates that for our future plans of identifying tutor strengths and assigning tutors to students appropriately, we may not be able to rely on their self-reported experience— tutors may not be great at assessing their own level of expertise. Further scaffolding of self-evaluation through detailed questions could also result in more accurate self-evaluation.

5.4 Tutor feedback and performance analytics suggest several design modifications

We identified several design modifications based on tutor feedback and learning performance which will be considered in future lesson iterations. Examples of possible design modifications include: increasing the difficulty of selected responses by capturing common misconceptions and errors from matched constructed responses; ensuring the math content knowledge necessary to successfully complete the lesson is universally understood by tutors; and possibly removing time-consuming problem types (i.e., *predict* constructed responses during the pretest) which did not demonstrate proportional learning gain given the amount of time to complete (see *Implications for Future Work*).

5.5 Learnersourced selected responses were too "easy," but common misconceptions were identified

Some of the selected-response questions were too easy for many tutors, particularly the questions with options capturing blatantly "wrong" responses and not necessarily common misconceptions or distractors (e.g., *"Great job, Kevin! You are so smart!"* within *Giving Effective Praise*). The matched constructed response questions were more difficult. We propose that selected responses can be just as pedagogically valuable as constructed responses and provide equally robust learning experiences- a controversial idea disputed by [17] and supported by [34]. Methods of improving the instructional impact of selected responses include labeling common errors and distractors from learnersourced constructed responses and using them as future selected-response options. Using learnersourced data for "incorrect" and "correct " constructed responses will help create better selected-response options- in other words, we will create better "wrong" answers capturing misconceptions and common errors in tutor understanding.

An example of how we are able to identify common tutor misconceptions within constructed responses was found in the *Responding to Student's Errors* lesson. Several tutors attempted to use the research-recommended approach of avoiding calling attention to a student's error, but instead tutors tried to minimize the severity of the error signifying it as "minor" or "small," indicated by: *"This is almost correct, but you have a small error. Can you tell me how you started this problem?"* and *"Great! Can you check your math. There is a minor error and it's a common mistake."* Tutors tried to use softer language (underlined), but the response is still pointing out the error. This was a common mistake among tutors in the constructed responses but this mistake was not an option in the selected response options. Our next iteration will include one of these common mistake responses in the selected response options. Below Figure 6 shows another example (from the *Giving Effective Praise* lesson) of learnersourcing to improve a selected-response question based on prior incorrect constructed responses. Tutors were 92% correct on this original item and the most frequently selected incorrect response was *"Kevin, great job on working through that problem. Next time, don't quit when you don't get it correct the first time."* In contrast, many of the constructed responses were wrong in a more nuanced way and these are illustrated in Figure 6 as options 1-3 (with option 4 being the recommended, or "correct" response). Future research will analyze whether this redesigned question is a more effective assessment that more closely approximates the difficulty of the matched constructed response question.

5.6 Limitations

There were several limitations affecting the generalizability of this investigation. First, demographics and experience level of tutors varies greatly across organizations suggesting the findings of this work may not generalize to a broader population. Second, through internal reliability estimates we know our lessons are brief, however lesson delivery via Google Forms inhibited data collection of completion time for the entire lesson and by problem type/format. In addition, we are not sure how much tutoring experience plays a role in performance as self-reporting does not seem to be a good measure of tutor ability, or our operationalization is not a valid construct- meaning our lessons may not be teaching and assessing a competent tutoring skill. Tutors were unpaid volunteers providing support through an online chatroom suggesting possible selection bias of tutors. Last, and perhaps most importantly, the transfer of learning evidenced by tutors applying the learned tutoring strategies in real-life situations was not validated within this investigation. Future work involves assessing the transfer of learning through chat log analysis within the tutoring platform.

6 IMPLICATIONS FOR FUTURE WORK

6.1 Design Modifications

Upon analysis of participant performance reflecting upon question format (i.e., constructed or selected response), type (i.e., *predict*

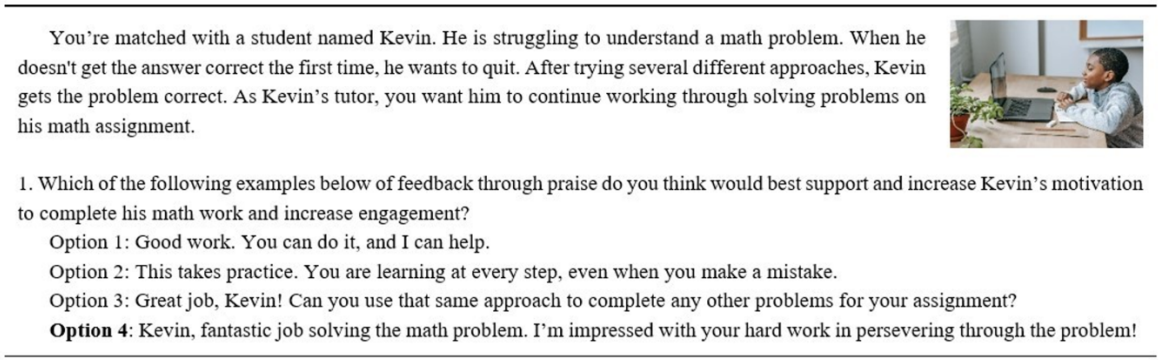


Figure 6: The Kevin Scenario within *Giving Effective Praise* displays the correct answer, Option 4 (in bold), and optimized, "incorrect" options using learnersourced data from this pilot study (Options 1-3). Future research will analyze whether this redesigned question is a better assessment that more closely approximates the difficulty of the matched constructed response question.

or *explain*), and participant feedback, there are several identified areas for improvement to increase tutor learning while maintaining brevity. We propose several design modifications described below:

- providing *corrective* feedback explaining why the "correct" option is most desired and aligns with the expert-based approach and the "incorrect" options are less desired or do not align with the expert-based approach
- increasing the difficulty of the selected-response questions making "correct" response options less obvious and capturing common misconceptions and tutoring errors within "incorrect" response options.
- ensuring scenarios deal with math content knowledge that all tutors have the ability of understanding.
- removing specific lesson objectives at the beginning of the lesson which hint at the specific assessed strategy.
- exploring removal of some problem types not necessarily connected to participant learning but take substantial time to complete (i.e., *predict* constructed responses during the training scenario).
- lessening the number of selected-response options from four to three by removing "incorrect" options not often chosen, suggesting they are obviously "wrong."
- varying modalities of scenario delivery (i.e., video- or audio-delivered scenarios).

Is there any loss in tutor learning upon removing the constructed-response questions within the pretest, or training scenario? Removing questions shown by our research findings to not have a substantial impact on learning gain, would conserve time, and allow for more impactful learning by doing activities. In addition, adding a third scenario with only selected-response questions between the training and transfer scenarios may give participants more deliberate practice, shown to increase learning [9] prior to completing the posttest, or transfer scenario. Additional research questions involve exploring the influence of explicitly telling tutors within the introductory text that the constructed responses will be graded manually by humans. Does knowledge of humans reading and assessing your open responses impact the quality of answers received? Additional questions include: Does assessed tutor competency predict performance in the field (transfer to real-life scenarios)? In other words, do tutors scoring higher on motivation support strategies get higher return rates in practice?

6.2 Machine Learning Models & Automated Short Answer Scoring

Machine learning models using natural language processing are being created from pilot data for future comparison of accuracy and speed in autograding versus manual coding of responses. To preserve the necessary number of open-ended questions, but at the same time reduce the labor required for manually scoring the constructed responses, an auto scoring tool using natural language processing was developed for *Responding to Student's Errors* with the user interface giving a predicted score for each response and highlighting the high-weight keywords in the sentences to assist manual score adjustment. We hope to speed up the scoring process by prioritizing the responses with relatively uncertain prediction results, meaning these responses are more in need of human evaluation [10]. Future studies can then focus on the improvement of model accuracy and highlight quality, and the evaluation of instructor's benefits and satisfaction with the autograding tool. In addition, receiving explanatory feedback instead of merely a holistic score has been found crucial to learning in an enormous number of studies related to different domains [32]. Expanding on the work of [24], we are exploring using natural language processing to recognize response details, such as common misconceptions and errors [25] and give pre-edited targeted feedback. Learnersourcing can be leveraged to provide future tutors (learners) with pre-edited targeted feedback and explanations, which is labor intensive when manually performed [34]. Further studies will evaluate the effectiveness of auto feedback on tutor learning.

7 CONCLUSION

This work investigated the effectiveness of tutor training on instructing tutors on how to respond to situations dealing with

students' motivational challenges. In addition, we sought to assess tutors' perceptions of their learning. We used both learning analytics and tutor feedback to guide our lesson design modifications. Results indicate tutors are learning from our brief, scenario-based lessons with tutors performing an average of 20% better on posttest compared to pretest. Women are shown to perceive their learning higher than other genders— a finding warranting further investigation and of particular interest among tutoring organizations. Tutors are not good self-evaluators suggesting self-reporting expertise is not a reliable indicator of tutor performance. From this mixed methods analysis we suggest several design modifications (i.e., creating more authentic selected-response options using learnersourced data, adding audio- or video- delivery of scenarios). Learnersourced selected responses were too "easy," but common misconceptions and distractors were readily identified from this pilot data for future lesson iterations. Next steps involve improving lesson design for larger scale deployment using design modifications discussed to ensure continued brevity while maximizing learning gain.

ACKNOWLEDGMENTS

We thank Hui Cheng for her dedicated and considerable contribution to the data analysis. This work is supported with funding from the Chan Zuckerberg Initiative (Grant #2018-193694), Richard King Mellon Foundation (Grant #10851), Bill and Melinda Gates Foundation, and the Heinz Endowments (E6291). Any opinions, findings, and conclusions expressed in this material are those of the authors.

REFERENCES

[1] Reem H. Al-Attar. 2019. The effectiveness of using scenario-based learning strategy in developing EFL eleventh graders' speaking and prospective thinking skills. *The Islamic University of Gaza, Palestine.* https://library.iugaza.edu.ps/thesis/126913.pdf.

[2] Lisa Bardach, Robert M. Klassen, Tracy L. Durksen, Jade V. Rushby, Keiko C. P. Bostwick, and Lynn Sheridan. 2021. The power of feedback and reflection: Testing an online scenario-based learning intervention for student teachers. *Computers & Education, 169*, 104194.

[3] Pallavi Chhabra, Danielle R. Chine, Adetunji Adeniran, Shivang Gupta, and Kenneth R. Koedinger. June 2022. An Evaluation of Perceptions Regarding Mentor Competencies for Technology-based Personalized Learning. *In E. Langran (Ed.), Proceedings of Society for Information Technology & Teacher Education International Conference.* 1812-1817. San Diego, CA: Association for the Advancement of Computing in Education (AACE).

[4] Danielle R. Chine, Cassandra Brentley, Carmen Thomas-Browne, J. Elizabeth Richey, Abdulmenaf Gul, Paulo F. Carvalho, Lee Branstetter, and Kenneth R. Koedinger. 2022. Educational Equity Through Combined Human-AI Personalization: A Propensity Matching Evaluation. In *International Conference on Artificial Intelligence in Education.* 366-377. Springer, Cham.

[5] Danielle R. Chine, Pallavi Chhabra, Adetunji Adeniran, Shivang Gupta, and Kenneth R. Koedinger. 2022. Development of Scenario-based Mentor Lessons: An Iterative Design Process for Training at Scale. *In Proceedings of the Ninth ACM Conference on Learning@Scale.*

[6] Ruth Clark. 2009. Accelerating expertise with scenario-based learning. *Learning Blueprint. Merrifield, VA: American Society for Teaching and Development, 10.*

[7] Paul Denny, John Hamer, Andrew Luxton-Reilly, and Helen Purchase. September 2008. PeerWise: students sharing their multiple choice questions. In *Proceedings of the fourth international workshop on computing education research.* 51-58.

[8] John Dewey. 1916. (2007 edition). Democracy and Education, Teddington: Echo Library.

[9] Angela Lee Duckworth, Teri A. Kirby, Eli Tsukayama, Heather Berstein, and K. Anders Ericsson. 2011. Deliberate practice spells success: Why grittier competitors triumph at the National Spelling Bee. *Social psychological and personality science, 2*(2), 174-181.

[10] Hiroaki Funayama, Tasuku Sato, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2022. Balancing Cost and Quality: An Exploration of Human-in-the-Loop Frameworks for Automated Short Answer Scoring. In *International Conference on Artificial Intelligence in Education.* 465-476. Springer, Cham.

[11] Graham Gibbs. 1988. Learning by doing: A guide to teaching and learning methods. *Further Education Unit.*

[12] Arthur C. Graesser, Patrick Chipman, Brian C. Haynes, and Andrew Olney. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. IEEE Transactions on Education, 48(4), 612-618.

[13] Martin Hänze, Marion Müller, and Roland Berger. 2018. Cross-age tutoring: How to promote tutees' active knowledge-building. *Educational Psychology*, 38(7), 915–926.

[14] Elizabeth Heubeck. October 21, 2021. *Schools are in desperate need of tutors. But qualified ones are hard to find.* EducationWeek. https://www.edweek.org/leadership/schools-are-in-desperate-need-of-tutors-but-qualified-ones-are-hard-to-find/2021/10.

[15] Cigdem Hursen and & Funda Gezer Fasli. 2017. Investigating the Efficiency of Scenario Based Learning and Reflective Learning Approaches in Teacher Education. *European Journal of Contemporary Education, 6*(2), 264-279. https://doi.org/10.13187/ejced.2017.2.264.

[16] Anne Jelfs, John T. E. Richardson, and Linda Price. 2009. Student and tutor perceptions of effective tutoring in distance education. *Distance Education, 30*(3), 419-441.

[17] Sean H. K. Kang, Kathleen B. McDermott, and Henry L. Roediger III. 2007. Test format and corrective feedback modify the effect of testing on long-term retention. *European journal of cognitive psychology, 19*(4-5), 528-558.

[18] Hassan Khosravi, Gianluca Demartini, Shazia Sadiq, and Dragan Gasevic. April 2021. Charting the design and analytics agenda of learnersourcing systems. In *LAK21: 11th International Learning Analytics and Knowledge Conference.* 32-42.

[19] Juho Kim. 2015. *Learnersourcing: improving learning with collective learner activity* (Doctoral dissertation, Massachusetts Institute of Technology).

[20] Matthew A. Kraft and Grace T. Falken. 2021. A Blueprint for Scaling Tutoring and Mentoring Across Public Schools. *AERA Open.* 7(1). 1-21.

[21] Kenneth R. Koedinger, Jihee Kim, Julianna Zhuxin Jia, Elizabeth A. McLaughlin, and Norman L. Bier. March 2015. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the second (2015) ACM conference on learning@ scale.* 111-120.

[22] Ronald B. Larson. 2019. Controlling social desirability bias. *International Journal of Market Research*, *61*(5), 534-547.

[23] Lydia Marshall, Jonah Bury, Robert Wishart, Rebekka Hammelsbeck, and Emily Roberts. 2021. *The national online tuition pilot.* Education Endowment Foundation. https://d2tic4wvo1iusb.cloudfront.net/documents/projects/National_Online_Tuition_Pilot.pdf?v=1630925212

[24] Joshua J. Michalenko, Andrew S. Lan, and Richard G. Baraniuk. April 2017. Data-mining textual responses to uncover misconception patterns. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale.* 245-248.

[25] Rafael Ferreira Mello, Rodrigues Neto, Giuseppe Fiorentino, Gabriel Alves, Verenna Arêdes, Joao Victor, Galdino Ferreira Silva, Taciana Pontual Falcao, and Dragan Gašević. 2022. Enhancing Instructors' Capability to Assess Open-Response Using Natural Language Processing and Learning Analytics. In *European Conference on Technology Enhanced Learning.* 102-115. Springer, Cham.

[26] Pardjono Pardjono. 2016. Active learning: The Dewey, Piaget, Vygotsky, and constructivist theory perspectives. *Jurnal Ilmu Pendidikan Universitas Negeri Malang*, *9*(3), 105376.

[27] Marta Pellegrini, Cynthia Lake, Amanda Inns, and Robert E. Slavin. 2018. Effective Programs in Elementary Mathematics: A Best-Evidence Synthesis. Best Evidence Encyclopedia (BEE). *Center for Research and Reform in Education.*

[28] Carly D. Robinson, Matthew A. Kraft, Susanna Loeb, and Beth E. Schueler. 2021. Accelerating Student Learning with High-Dosage Tutoring. EdResearch for Recovery Design Principles Series. *EdResearch for Recovery Project.* https://annenberg.brown.edu/sites/default/files/EdResearch_for_Recovery_Design_Principles_1.pdf

[29] Peter Schaldenbrand, Nikki E. Lobczowski, J. Elizabeth Richey, Shivang Gupta, Elizabeth A. McLaughlin, Adetunji Adeniran, and Kenneth R. Koedinger. June 2021. Computer-Supported Human Mentoring for Personalized and Equitable Math Learning. In *International Conference on Artificial Intelligence in Education.* 308-313. Springer, Cham.

[30] Roger C. Shank, Tamara R. Berman, and Kimberli A. Macpherson. 1999. Learning by doing. *Instructional-design theories and models: A new paradigm of instructional theory, 2*(2), 161-181.

[31] Meredith Thompson, Kesiena Owho-Ovuakporie, Kevin Robinson, Yoon Jeon Kim, Rachel Slama, & Justin Reich. 2019. Teacher Moments: A digital simulation for preservice teachers to approximate parent–teacher conversations. *Journal of Digital Learning in Teacher Education*, *35*(3), 144-164.

[32] Daniela Torres, Surya Pulukuri, and Binyomin Abrams. 2022. Embedded Questions and Targeted Feedback Transform Passive Educational Videos into Effective Active Learning Tools. *Journal of Chemical Education, 99*(7), 2738-2742.

[33] Xu Wang, Carolyn Rose, and Kenneth R. Koedinger. May 2021. Seeing Beyond Expert Blind Spots: Online Learning Design for Scale and Quality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1-14.

[34] Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth R. Koedinger. June 2019. UpGrade: Sourcing student open-ended solutions to create scalable learning opportunities. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*. 1-10.

[35] Serap Samsa Yetik, Halil Ibrahim Akyuz, and Hafize Keser. 2012. Preservice teachers' perceptions about their problem solving skills in the scenario based blended learning environment. *Turkish Online Journal of Distance Education, 13*(2), 158-168.

[36] Derya Uzelli Yilmaz, Esra Akin Palandoken, Burcu Ceylan, and Ayse Akbiyik .2020. The effectiveness of scenario-based learning to develop patient safety behavior in first year nursing students. *International Journal of Nursing Education Scholarship, 17*(1).