Contents lists available at ScienceDirect

The Internet and Higher Education

journal homepage: www.elsevier.com/locate/iheduc

The benefit of reflection prompts for encouraging learning with hints in an online programming course

Heeryung Choi [a,*], Jelena Jovanovic [b], Oleksandra Poquet [c], Christopher Brooks [d], Srećko Joksimović [e], Joseph Jay Williams [f]

[a] *Massachusetts Institute of Technology, Center for Transportation and Logistics, E40-369, 1 Amherst St, Cambridge, MA 02142, USA*
[b] *University of Belgrade, Faculty of Organizational Sciences, Jove Ilica 154, 11000 Belgrade, Serbia*
[c] *Technical University of Munich, School of Social Sciences and Technology, Marsstraβe 20, 80335 München, Germany*
[d] *University of Michigan, School of Information, North Quadrangle 4439, 105 S State St, Ann Arbor, MI 48109, USA*
[e] *University of South Australia, Education Futures, 101 Currie St, Adelaide, SA 5001, Australia*
[f] *University of Toronto, Department of Computer Science, 27 King's College Circle, Toronto, Ontario M5S 1A1, Canada*

ARTICLE INFO

Keywords:
Reflection prompt
Learning with hint
Performance
Perceived learning

ABSTRACT

While giving learners hints is a commonly used scaffolding practice to facilitate learning, previous work questioned the effectiveness of hints. In this study, we examined if prompting learners to reflect along with receiving hints could improve learning outcomes, including immediate and delayed performance, perceived learning, and enjoyment. A field experiment was conducted in a four-week long online master's degree course on data science where we compared two conditions: a condition with hints and a condition providing reflection prompts along with hints. Results showed that using hints with reflection prompts increased learner performance in delayed knowledge transfer tasks while also increasing learners' perception of learning. The combination of reflection prompts and hints did not reduce learners' enjoyment of the tasks, suggesting that the use of hints with reflection prompts is not only an intervention which can improve learning outcomes but is one which will be naturally adopted by learners.

1. Introduction

1.1. Hints and reflection prompts

Online learning environments often support self-study processes by offering hints to learners. Hints are designed to trigger a particular cognitive process that will help learners understand a given task or a set of concepts better (Nokes, Hausmann, VanLehn, & Gershman, 2011; Roll, Aleven, McLaren, & Koedinger, 2011). Often, hints provide key concepts or approaches which are critical for solving task problems. Hints are mostly provided on demand, but a few studies also proposed algorithms detecting learners' struggles that could be used for offering hints without waiting for learners' requests (Pan & Liu, 2022; Aleven, Mclaren, Roll, & Koedinger, 2006; Roll et al., 2011; Dong, Marwan, Shabrina, Price, & Barnes, 2021; Messer, 2022). Despite the broad use of hints, multiple pieces of evidence have led researchers to question the effectiveness of hints. Some pointed out the so-called 'hint abuse' behavior, where learners mindlessly click through the hints to find and copy the answer (Aleven et al., 2006; Muñoz-Merino, Valiente, & Kloos, 2013). Moreover, the support designed to reduce such undesirable interactions with hints alone has been shown insufficient to increase domain knowledge (Roll et al., 2011; Baker et al., 2006). To increase the effectiveness of hints for learning, additional pedagogical support may be required.

Theoretical and empirical work suggests that reflection can be an essential complementary intervention that can help learners meaningfully interact with hints. Reflection is generally defined as a process of expanding and deepening one's understanding by critically analyzing what has been learned and how (Dewey, 1997; Moon, 2013; Boud, Keogh, & Walker, 1985; Lew & Schmidt, 2011). Through reflection, learners evaluate their current strategy and aim to determine if there are better strategies to complete a given task, a crucial component of self-regulated learning (Boekaerts, 1999; van den Boom, Paas, van Merriënboer, & van Gog, 2004). Reflection also leads learners to deliberately

[*] Corresponding author.
E-mail addresses: heeryung@mit.edu (H. Choi), jelena.jovanovic@fon.bg.ac.rs (J. Jovanovic), sasha.poquet@tum.de (O. Poquet), brooksch@umich.edu (C. Brooks), Srecko.Joksimovic@unisa.edu.au (S. Joksimović), williams@cs.toronto.edu (J.J. Williams).

https://doi.org/10.1016/j.iheduc.2023.100903
Received 4 August 2022; Received in revised form 13 February 2023; Accepted 17 February 2023
Available online 8 March 2023
1096-7516/

review their learning experiences and learn from them. Practicing reflection has been found to increase academic performance (Lew & Schmidt, 2011; O'Rourke, 1998; Dewey, 1997; Moon, 2013; Boud et al., 1985). It should be noted that most learners do not deliberately activate meaningful learning processes such as reflections without intervention (Lin & Lehman, 1999; Berardi-Coletta, Buyer, Dominowski, & Rellinger, 1995; Bransford, Brown, & Cocking, 1999), suggesting that a reflection process needs to be explicitly triggered within the pedagogical design. When such a process is triggered, it is often positively associated with the outcome. Shih, Koedinger, and Scheines (2011) showed that time spent between receiving hints and writing up the insight obtained from the hint was positively associated with learning outcomes, suggesting that those who engaged in a reflective process upon the use of the hint, have learned more.

Many pedagogical methods have been developed to encourage productive reflection. For instance, Richardson and Maltby (1995) had learners write reflection diaries and found these writings helped students remember and analyze what they had learned and improved their practice. Bye, Smith, and Rallis (2009) used discussion forums to set up an environment where learners shared their reflection essays and engaged in peer review and discussion. Learners who interacted with their peers in the discussion forum reported a significantly higher sense of mastery of course objectives than learners who submitted a hard-copy reflection essay to instructors without integrating any peer responses. Several studies have looked at integrating prompts into the reflection process, and findings have shown that such prompts successfully triggered productive reflection during learning (Chen, Kinshuk, Wei, & Liu, 2011; Lin & Lehman, 1999; Hung, Yang, Fang, Hwang, & Chen, 2014). Prompts are often categorized into directed and generic prompts. Directed prompts offer specific instructions such as 'stop and think about what you might have misunderstood before seeing hints,' while generic prompts do not give specific instructions (e.g., 'What are you thinking now?'). The distinction between direct and generic prompts was proposed by Davis (2003), yet there is no agreement within the community on which of the two is more effective for learning (Kramarski & Kohen, 2017; Ifenthaler, 2012; Davis, 2003; Kramarski, Weiss, & Sharon, 2013). Nonetheless, directed prompts have been studied broadly (Lin & Lehman, 1999; Chen et al., 2011; Coulson & Harvey, 2013; Bannert, Sonnenberg, Mengelkamp, & Pieger, 2015). While each has its own strengths, prompts have been frequently adopted in online courses as they can be easily embedded directly into the course material instead of requiring a second learning context (e.g. a discussion forum). Furthermore, unlike diaries, prompts can evoke reflection with minimal delay after each episode of learning, a behavior that may be beneficial in promoting reflection after each portion of the course content has been reviewed.

While many studies have demonstrated the benefits of reflection on learning, it has often been outside of the scope of those studies to investigate the effect of reflection when accompanied by hints. When cognitive and metacognitive prompts have been combined, cognitive prompts are often different from what would generally be considered hints. For example, Berthold, Nückles, and Renkl (2007) designed cognitive prompts to activate organization and elaboration strategies rather than providing a hint which targeted missing or misconstrued domain knowledge. The work of Marwan, Williams, and Price (2019) showed that hints significantly increased immediate programming performance only when they were accompanied by self-explanation prompts. However, in that context, self-explanation prompts did not seem to elicit reflection processes and instead were (1) designed primarily for general critical thinking rather than reflection and (2) offered to learners before they could apply hints to solve a problem. Furthermore, it is hard to assume the effect of reflection on learning with hints since most prior studies used metacognitive prompts that were aimed at eliciting several metacognitive processes beyond the reflection instead of reflection only. For example, in Lin and Lehman (1999)'s study, prompts were used to promote self-monitoring as well as reflection, making it hard to disentangle the direct effect of reflection alone. In sum, further investigations are necessary to better understand how learning outcomes are affected by the reflection prompts integrated with hints.

This study addresses the limitations of previous work by examining how the combination of prompt-based reflection and problem-specific hints impacts learners. To this end, we conducted a randomized field experiment examining two conditions: (1) hints alone and (2) hints and reflection prompts combined. We measured how these conditions affected both immediate and delayed performance on transfer tasks, perceived learning, and enjoyment of learning. Our findings show that hints can benefit immediate and delayed performance on transfer tasks, as well as perceived learning, but only when reflection prompts accompany the use of hints. These findings provide practical implications for the design of reflection prompts that are effective, easily deployable, and likable.

1.2. *Design factors for reflection prompts*

Several important factors need to be considered for the evaluation of prompts triggering reflection, namely (1) task complexity, (2) inclusion of tasks that are immediately after the intervention or delayed, and (3) features that balance likeability and effectiveness (Bransford et al., 1999; Day & Goldstone, 2012; Kim & Kendeou, 2021; Lin & Lehman, 1999; Lew & Schmidt, 2011; Schworm & Renkl, 2002; Bannert, 2006; Bannert & Reimann, 2012; Bannert et al., 2015; Krause & Stark, 2010). First, types of tasks and their complexity have been noted as factors that are related to learner reflection processes. The benefits of reflection prompts seem to appear more clearly when learners are working on knowledge transfer tasks. Transfer tasks are those which require that learners apply adopted knowledge to new contexts (Bransford et al., 1999; Barnett & Ceci, 2002; Day & Goldstone, 2012; Kim & Kendeou, 2021). Earlier studies reported that learners within the same intervention condition did not benefit from reflection prompts when working on simpler tasks such as those measuring recall or knowledge comprehension (Lin & Lehman, 1999; Lew & Schmidt, 2011; Schworm & Renkl, 2002; Bannert, 2006; Bannert & Reimann, 2012; Bannert et al., 2015). For instance, three studies by Bannert and colleagues, which examined the effects of reflection prompt interventions, consistently presented a significant increase in transfer task performance with medium effect sizes ($d = 0.55$, $d = 0.58$, $d = 0.44$) (Bannert, 2006; Bannert & Reimann, 2012; Bannert et al., 2015). Schworm and Renkl (2002) also reported increase in learners' performance on transfer tasks in their experiments where reflection prompts were accompanied by self-explanation tasks. Similar positive effects of reflection prompts have been observed with tasks that were not necessarily transfer tasks but were generally tasks of higher complexity. For instance, although Krause and Stark (2010) did not find any impact of reflection prompts on task performance, they showed that learners made significantly more progress when working on more complex problems with a large effect size ($d = 0.8$). Similarly, Lin and Lehman (1999) observed positive effects of reflection prompts on performance only for complex transfer tasks. As these studies collectively show, reflection prompts are highly likely to improve performance on transfer tasks or other tasks with higher complexity, whereas they do not affect performance on less complex tasks that only involve recall or knowledge comprehension.

In addition to task complexity, the inclusion of both immediate and delayed tasks has shown to be a meaningful approach to evaluating the effects of reflection prompts. Multiple previous studies examined the effects of reflective prompts both immediately after the reflection process and after a delay, and reported mixed results. Some researchers found no immediate effect on learning outcomes (Krause & Stark, 2010; van den Boom et al., 2004). In contrast, Bannert and colleagues (Bannert, 2006; Bannert & Reimann, 2012) found significant positive effects of reflection prompts on immediate post-test scores. Delayed effects of reflection prompts on learning have been far less explored and also with mixed results. Bannert et al. (2015) found that learners who received reflective prompts performed significantly better on their delayed transfer tasks, whereas Jeong et al. (2008) found no effect of prompts evoking reflections along with other metacognitive processes on delayed task performance. These mixed results suggest that the duration of the effect requires more investigation. In particular, to clearly account for when the potential effect of reflection prompts appears, it becomes necessary to measure both immediate and delayed effects while controlling task complexity to transfer tasks.

Finally, it is important to design reflection prompts in a way that balances their effectiveness with likeability. Perceived learning, which is related to likeability, has been shown to play a relevant role in metacognitive monitoring and evaluation, potentially affecting learners' willingness to engage with a certain intervention (Barzilai & Blau, 2014; Veenman, 2011; Efklides, 2008). Enjoyment of the interaction with the intervention may play a similar role, and previous studies show that low enjoyment and high annoyance are factors that can lead learners to low compliance with prompts (Berthold et al., 2007; Bannert & Reimann, 2012). Taken together, if learners do not perceive there is learning or enjoyment value in a reflection prompt activity, they may easily lose interest in the intervention and cease interacting with and benefiting from the activity. A lack of interaction with reflection prompts would be a substantial problem in this study which aims to measure the effect of the reflection prompts and hints. Thus, a pilot study was designed and conducted prior to the main study to investigate effective and interesting intervention designs that the particular target learners may accept.

2. Research questions

We aimed to understand the effect of the combined use of reflection prompts and hints on the following learning outcomes: task performance, the number of task submissions, perceived learning, and enjoyment of learning; for the former two learning outcomes, we were interested in both immediate and delayed effects. Accordingly, each of our research questions is focused on one of these learning outcomes:

RQ1. What is the immediate effect of a combined hint and reflective prompt intervention on transfer task performance?
RQ2. What is the delayed effect of a combined hint and reflective prompt intervention on transfer task performance?
RQ3. What is the immediate effect of a combined hint and reflective prompt intervention on the number of task submissions?
RQ4. What is the delayed effect of a combined hint and reflective prompt intervention on the number of task submissions?
RQ5. What is the immediate effect of a combined hint and reflective prompt intervention on learners' perceived learning?
RQ6. What is the immediate effect of a combined hint and reflective prompt intervention on learners' enjoyment of learning?

RQ1 and RQ2 focused on investigating the difference between the immediate (RQ1) and delayed effects (RQ2) of the interventions on task performance. RQ3 and RQ4 aimed to understand immediate (RQ3) and delayed changes (RQ4) in the way learners interacted with the given tasks. Related to RQ3 and RQ4, it should be noted that learners were allowed to submit their tasks as many times as they wanted and could check their scores for each submission. If a learner submitted tasks significantly more and achieved the same or lower scores compared to other learners, it would be evidence of the learner's less meaningful engagement with tasks, and would provide evidence of gaming the system by learning carelessly, guessing answers, and checking if their guesses were correct. Finally, RQ5 and RQ6 were posed to understand the effect of the interventions on perceived learning (RQ5) and enjoyment of learning (RQ6) after using reflection prompts.

We approached these questions through a two-condition randomized field study which included (1) the *hint condition*, where learners were offered an option to use hints, and (2) the *hint-reflection condition* where learners were presented with both hints and reflection prompts. We opted for a randomized field study in order to preserve an authentic learning environment, where learners are more genuinely motivated stakeholders compared to a controlled lab study. Since the study aimed to answer the impact of reflection prompts on learning with hints, the hint condition was designed as a control condition, whereas the hint-reflection condition was designed as the primary treatment condition to measure the effect of combined hints and reflection prompts.

3. Pilot study

3.1. Study context

Prior to the main experiment, a pilot study was conducted to develop a design of hints and reflection prompts. The primary design decisions to be made were (1) when to present reflection prompts - every time before or after a hint was presented or only after a learner completed the task, (2) which concepts or code elements to explain through hints, (3) what types of reflection prompt questions were effective in eliciting reflection, and (4) if learners could clearly understand tasks, reflection prompts, and hints. The pilot study was necessary to understand the specific learner population we were engaging with. These largely part-time learners pursuing an online degree, with various levels of prior knowledge on the subject may show different patterns of interactions with hints and reflection prompts compared to previous studies in the area. Previous studies applied various designs of hints and reflection prompts, primarily in lab studies as well as with more typical undergraduate populations. Therefore, we first engaged with a pilot study to ensure our design was suitable for the target population. The pilot study was conducted in an online course offered by the University of Michigan, which focused on learning programming, and 19 learners who had already taken the course were recruited for the pilot study.

3.2. Study procedure

The authors and the course instructor reviewed a pool of hints and discussed if the hints clearly explained challenging concepts to learners. A set of reflection prompts were designed based on the literature (Ifenthaler, 2012; Devolder, van Braak, & Tondeur, 2012; Davis, 2003; Zepeda, Richey, Ronevich, & Nokes-Malach, 2015). During the pilot study, the first author explained the benefits of reflection prompts to the participants to reduce potential annoyance and unwillingness to interact with the prompts (Bannert & Reimann, 2012). Learners were then asked to think aloud while working on tasks with both hint and reflection prompt interventions. Learners were provided with a reflection prompt before and after every hint request. After they completed tasks and think-aloud, the first author interviewed them.

3.3. Instruments

Follow-up interviews. The first author asked common questions and follow-up questions to clarify some of the statements made during each participant's think-aloud session. Common questions were as follows:

- Think about how you interacted with the reflection prompts. Did the prompts help you check your understanding of hints?
- Was it easy to understand what prompts asked you to do?
- Think about how you interacted with hints while working on the Jupyter Notebook programming tasks. Did hints address what you wanted to know to solve your problem?
- Was it easy to understand the suggestions that hints gave?

Follow-up questions were varied and determined by specific statements the participants made while self-reporting their work on the task.

3.4. Pilot study results

Overall, many pilot study participants did not comply with reflection prompts regardless of when the prompts were presented (right before or after a hint). The reflection prompt given before each hint was designed to encourage them to reflect on what they already knew and what they were not sure of. However, 13 out of 19 participants ignored these reflection prompts and proceeded to see hints. During the follow-up interview, one participant (P2) said that they did not have enough motivation to engage with a reflection prompt when they were frustrated and eager to see a hint to resolve the frustration. Another participant (P12) also said that using reflection prompts before seeing hints did not add anything to their understanding. One participant (P15) even described the prompt as 'an extra obstacle to the selection of hints.' Similarly, 9 out of 19 participants considered reflection prompts given after every hint annoying and did not comply with prompts asking them to activate reflection.

The participants' annoyance due to the high frequency of prompting along with the low compliance with prompts is not surprising considering previous studies (Berthold et al., 2007; Bannert & Reimann, 2012). It has been frequently shown that learners often prefer to invest less effort in learning and wrongly believe that they choose an efficient way to increase learning gain (Bjork, Dunlosky, & Kornell, 2013; Clark, 1982). Yet, it was concerning that most participants did not engage with reflection prompts even though the benefits of engaging with reflection prompts were explained to them, to motivate them for such engagement (Bannert & Reimann, 2012). To prevent learners' annoyance and to draw their attention to reflection prompts, for the main study, we decreased the frequency of prompting to only once when learners completed their tasks. In the modified design, reflective prompts aimed to activate learners' reflection process over the entire task, including but not limited to their use of hints. Marwan et al. (2019) have also suggested such a design to encourage overall task reflections instead of eliciting them after each hint.

Based on learners' feedback, we also included additional hints to address a few more challenging concepts in the assignment. Furthermore, participants showed a clear preference for directed prompts over generic prompts during the interviews. Otherwise, learners agreed that the assignment and reflection prompts were clear.

4. Method

4.1. Participants

Data were collected within a four-week long introductory data science course in a fully online Master's degree program of the University of Michigan. In total, 165 learners enrolled in the 2021 Winter iteration of the course (Female $= 47$, Male $= 118$, age $M = 31.96$, age $SD = 6.79$). The course was self-paced with fixed dates of assignment deadlines. Learners were randomly assigned to either the hint-reflection condition (treatment) or the hint condition (control).

Considering that learners could make multiple submissions of the same assignment, we used mixed-effect models for RQ1, RQ2, RQ3, and RQ4 as further explained in the Section 4.3. To determine a sample size to detect the effect of the interventions on performance (RQ1, RQ2), we ran a simulation using simr R package (Green & MacLeod, 2016) with the parameter that learners would make 5 submissions per assignment on average. The simulation showed that there should be at least 55 learners per condition to achieve an alpha of 0.05 and a power of 0.80 (Fig. 1 (a)). Another simulation was run to determine the sample size required for detecting the immediate and delayed effects of the interventions on assignment submission numbers (RQ3, RQ4). The simulation indicated that the total number of learners needed to be about 100 to achieve the same level of alpha and power (Fig. 1 (b)). Finally, G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) was used to calculate a sample size with a medium effect size, an alpha of 0.05, and a power of 0.80 for detecting a significant difference in perceived learning and enjoyment between conditions (RQ5, RQ6). G*Power showed that the number of learners per condition should be equal to or larger than 51.

To meet all the sample size requirements above, we aimed for a minimum of 55 learners per condition, randomly split into two groups and respectively assigned to the treatment condition and the control condition. The total number of learners in the dataset used for analysis after data cleaning, which will be described in the Section 4.3, was 132 with the following breakdown across the conditions: 70 participants in the hint condition and 62 participants in the hint-reflection condition. Overall, the number of participants ensures power larger than.80, medium effect size, and an alpha of 5%.

4.2. Instruments

The instruments were embedded into the course as shown on Fig. 2.

Questionnaire on metacognitive skills. To confirm that learners in different conditions did not significantly differ in metacognitive skills, including reflection, before engaging with the interventions, a survey on metacognitive skills was conducted at the beginning of the course. Six question items were adopted from the Metacognitive Self-Regulation section of the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich, 1991) and slightly re-worded to follow the context of the course.

Tasks and assignments. To measure learners' performance on immediate and delayed knowledge transfer tasks, four weekly assignments were provided through Jupyter Notebook, a web-based programming environment (Kluyver et al., 2016). All assignments were composed of

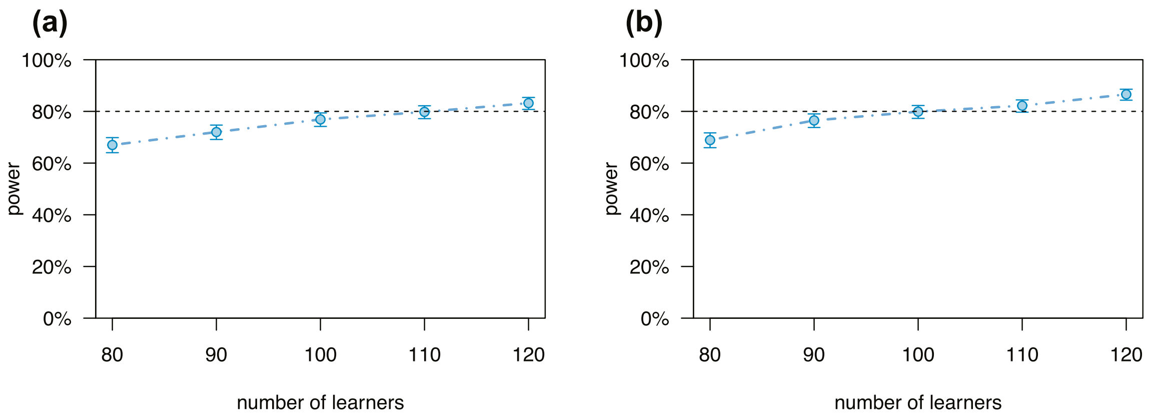


Fig. 1. (a) To answer RQ1 and RQ2, the power of 0.80 could be achieved with 110 learners (i.e., more than 55 learners per conditions). (b) For RQ3 and RQ4, the same level of power could be reached with total 100 learners.

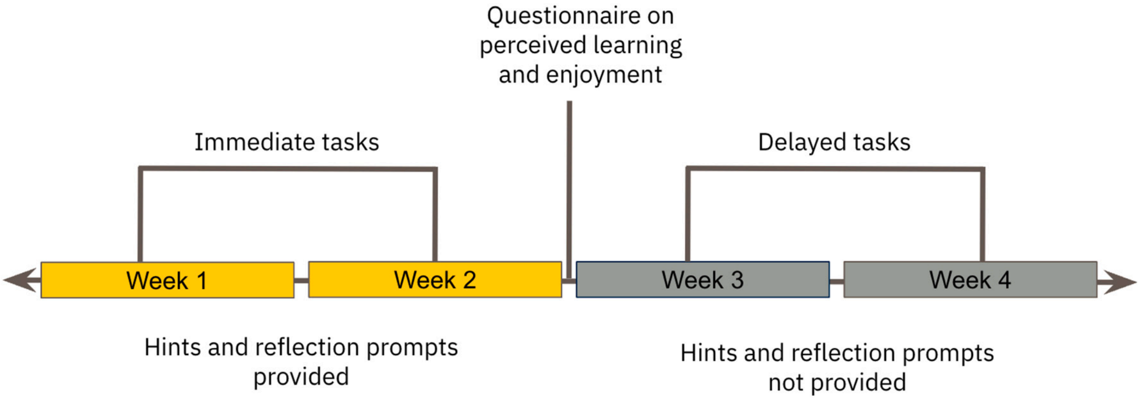

Questionnaire on perceived learning and enjoyment
Immediate tasks
Delayed tasks
Week 1
Week 2
Week 3
Week 4
Hints and reflection prompts provided
Hints and reflection prompts not provided

Fig. 2. (a) A graphic showing the study procedure which took place in the course.

Python programming tasks that required learners to apply their newly gained knowledge and skills to solve the task problems. The goals of these assignments were to help learners develop the knowledge and skills related to manipulating and analyzing data to derive insights. For example, in week 1, learners were asked to apply their regular expression skills to extract certain information from the given log data and retrieve it as a list of dictionaries. Each of the first two assignments, assignments 1 and 2, consisted of three separate tasks and included an intervention. These tasks were used to measure the immediate effect of the intervention on learners' academic performance and submission behaviors (RQ1 and RQ3). Since each of these tasks was provided separately, learners who were in the hint-reflection condition could activate their reflection per task. That is, learners were presented with reflection prompts at least three times per assignment since they had to submit three different tasks for each of assignments 1 and 2. The last two assignments, assignments 3 and 4, included more than one task per assignment and did not offer any interventions to learners in any of the conditions (i.e., neither hints nor prompts). These assignments were used to measure the delayed effect of interventions on learners' academic performance (RQ2 and RQ4).

Learners were allowed to submit assignments as many times as they wanted prior to a weekly deadline, and their submissions were graded through an automated code grading (autograder) system 'nbgrader' (Blank et al., 2019) within 20 min after submission. This system evaluated student submissions using a series of unit tests, a common software design for determining whether a given program produces expected outputs. All unit tests were written by the instructor, and the number of unit tests passed determined the students' grades. After reviewing their grade on a submission, learners could revise their code and repeat the submission process.

All tasks were designed and reviewed by the course staff and the first author to confirm that they measured what learners were supposed to learn each week. Conceptually, each assignment was built upon the previous assignments. Therefore, it was expected that learners who did not perform well on earlier assignments would not perform well on the subsequent assignments. Learners had to reach 80% of the full credit per assignment to pass the course.

Hints. Learners could engage with the hint intervention by clicking the 'Show Hint' button while working on a task. When a learner clicked the button, a pop-up with a list of summaries of available hints was displayed (Fig. 3 (a), (b)). When the learner chose a hint and clicked the 'Next' button, they could see the full text of the chosen hint on the next pop-up. This full hint was inscribed below the associated task cell (Fig. 3 (c)), so that learners could easily use hints while working on a task after closing the hint pop-up. Assignments with hints are available on the Open Science Foundation here[1].

Reflection prompts. Learners assigned to the hint-reflection condition were given a reflection prompt when they clicked the button to submit their task and receive auto-graded credit, as shown in Fig. 3 (d). Regarding the design of reflection prompts, we opted for the directed prompts based on the feedback of the pilot study participants who favored such prompts over generic prompts. Upon every task submission, a learner was provided with a reflection prompt question randomly chosen from the following list:

- What steps did you take when solving the problem? Why? Provide a short justification for each step.
- What did you find difficult or challenging in this task? Why? Provide a quick explanation.
- What was the main thing you learned by completing this task?

We designed more than one prompt question to avoid the negative effect of monotony due to the repetition of the same question. Considering that learners could make as many task submissions as they wanted, showing the same prompt throughout multiple submissions could discourage them from engaging with reflection prompts. Fig. 4 presents the submission and revision process for each assignment.

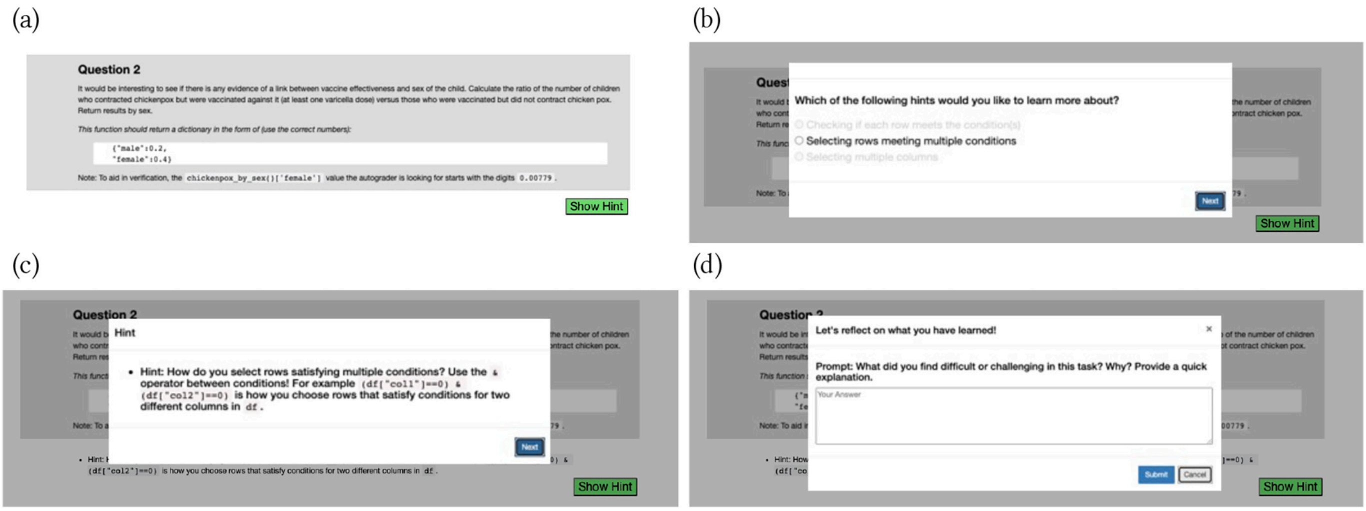
Questionnaire on perceived learning and enjoyment of learning. After the first two assignments where learners interacted with interventions, a questionnaire was distributed to learners to measure perceived learning and enjoyment of the learning experience with the given intervention. The questionnaire had four question statements related to perceived learning and three question statements on the enjoyment of learning, and both question sets were based on a 7-point Likert scale. Statements were adopted from Barzilai and Blau (2014) and adapted to the context of this study, as presented in Table 1.

4.3. Data analyses

For data cleaning purposes, we removed incomplete questionnaire submissions while including learner data as long as they made at least one complete submission of a task. In addition, there were learners who received a grace period for a portion of assignments due to their personal circumstances, and their tasks were manually graded with different criteria. In this case, we still retained their other submissions and questionnaire response data instead of dropping the entire data of the learner.

To examine the immediate (RQ1) and delayed (RQ2) effects of interventions on the transfer tasks, we conducted a mixed-effect model analysis in R using the lme4 package (Bates, Mächler, Bolker, & Walker,

[1] https://osf.io/ugaez.



(a)

Question 2

It would be interesting to see if there is any evidence of a link between vaccine effectiveness and sex of the child. Calculate the ratio of the number of children who contracted chickenpox but were vaccinated against it (at least one varicella dose) versus those who were vaccinated but did not contract chicken pox. Return results by sex.

This function should return a dictionary in the form of (use the correct numbers):

```
{"male":0.2,
"female":0.4}
```

Note: To aid in verification, the `chickenpox_by_sex()['female']` value the autograder is looking for starts with the digits `0.00779`.

(b)

Which of the following hints would you like to learn more about?

○ Checking if each row meets the condition(s)
○ Selecting rows meeting multiple conditions
○ Selecting multiple columns

(c)

Hint

- Hint: How do you select rows satisfying multiple conditions? Use the `&` operator between conditions! For example `(df["col1"]==0) &` `(df["col2"]==0)` is how you choose rows that satisfy conditions for two different columns in `df`.

`(df["col2"]==0)` is how you choose rows that satisfy conditions for two different columns in `df`.

(d)

Let's reflect on what you have learned!

Prompt: What did you find difficult or challenging in this task? Why? Provide a quick explanation.

Your Answer

Fig. 3. (a) Each task (labeled as "Question" in Fig. 3) has a 'Show Hint' button below the task text. (b) When a learner clicks the 'Show Hint' button, a pop-up appears and shows a list of available hints. Some hints are grayed out since they have already been chosen by the learner in previous interactions. (c) The full text of the chosen hint is shown on a pop-up and also inscribed below the associated task so that a learner can see hints after closing the pop-up. (d) When a learner clicks a button to submit the current task, a reflection prompt pop-up appears. Upon completing the prompt, the learner's submission is graded by an autograder. If the learner is not satisfied with their grade, they have the option to revise and re-submit their code as many times as desired.

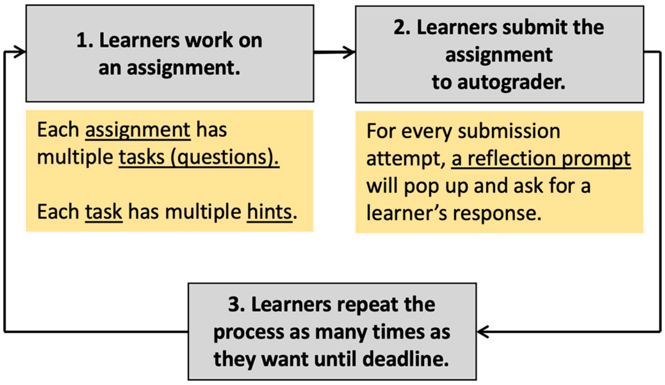

1. Learners work on an assignment.
Each assignment has multiple tasks (questions).
Each task has multiple hints.
2. Learners submit the assignment to autograder.
For every submission attempt, a reflection prompt will pop up and ask for a learner's response.
3. Learners repeat the process as many times as they want until deadline.

Fig. 4. A flowchart illustrating the steps in the submission and revision process for assignments.

Table 1
Question items for measuring perceived learning and enjoyment of learning.

Construct measured	Questionnaire items
Perceived learning	The weekly assignments helped me learn more about the topic (e.g., regular expression). I learned new things from the weekly assignments. The weekly assignments helped me remember the things I learned. The weekly assignments helped me apply the things I learned to other problems.
Enjoyment	I enjoyed the weekly assignments. I had fun working on weekly assignments. Working on the weekly assignments was pleasant.

2014). A linear mixed-effect model was adopted to account for multiple assignment submissions per learner. In this model, there were two fixed effects: condition and assignment label. The condition was a factor denoting if a learner was assigned to the control condition (i.e., the hint condition) or to the treatment condition (i.e., the hint-reflection condition). The assignment label was also a factor with possible values of 'immediate' (i.e., assignments given in week 1 and week 2) or 'delayed' (i.e., assignments given in week 3 and week 4). Learner ID was used as the random effect.

A generalized linear mixed-effect model analysis was run to examine the immediate (RQ3) and delayed (RQ4) effects of interventions on the count of task submissions. Since the number of submissions was count variables, it is assumed that the dependent variables would follow the Poisson distribution. Thus, a generalized linear mixed-effect model was chosen instead of a linear mixed-effect model which assumes a normal distribution of continuous variables. Fixed effects and the random effect in this model were the same as the ones in the mixed-effect model for RQs 1 and 2. For RQ5 and RQ6, we ran Welch's t-test over the questionnaire responses on perceived learning (RQ5) and enjoyment (RQ6).

For RQ1-4, Tukey's post hoc tests were run and effect size estimates were reported along with p-values at 5% significance level to provide the magnitude of the observed effects (Wasserstein & Lazar, 2016; Wasserstein, Schirm, & Lazar, 2019). All effect sizes were reported in Cohen's d values, where 0.20, 0.50, and 0.80 are considered, respectively, as a small, medium, and large effect sizes.

Since the assignment score data for RQ1 and RQ2 are continuous data, the normality of the residuals of the assignment score data was visually inspected and confirmed by a quantile–quantile (QQ) plot. On the other hand, the dataset of the number of submissions is composed of count data which do not assume the normality.

A few more analyses were conducted to consolidate the findings for RQ1 through RQ6 and to examine if there was any significant effect of confounding variables. Firstly, Welch's t-test was run on the pre-questionnaire on metacognitive skills to see if there was any significant difference in learners' initial metacognitive skills between conditions. If there was a significant difference, the impact of interventions could be questioned. Another Welch's t-test was conducted to reveal if there was any significant difference in the number of hints used between hint and hint-reflection conditions, since all RQs focused on the impacts of combined reflection prompts and hints. Furthermore, since one of three prompt questions was randomly given to learners each time they made a task submission, it was tested if any of these question was more effective than other prompts in evoking reflection. If any questions were more effective than others, it should be considered in analyzing the impacts of reflection prompts on learning outcome. To this end, learners' responses to reflection prompts were collected and categorized, by the first author, based on whether these responses were meaningful or not. If a learner typed strings irrelevant to the course materials (e.g., “asdfasf asdf asd”) or left their responses blank, such responses were considered ‘not meaningful.’ On the other hand, if a learner responded with interpretable and reasonable responses to reflection prompts (e.g., “I acknowledged the raw string formatting notation `r` as incorrect and commented out the original answer. I grouped upper case letters followed by lower case letters in one group but matching them consecutively. I needed a hint to remember that `finditer` would return groups of tuples and I was searching for a list of strings containing just the names.”), it was considered ‘meaningful.’ Then, to investigate if any particular prompt question was a stronger predictor of meaningful reflection responses, a mixed-effect logistic regression was built with binary codes as a dependent variable. There were two fixed effects in this model: prompt questions (See Section 4.2 for a list of prompt questions) and task label (e.g., task 1–1, task 1–2), since each task with different difficulty and topic could have affected contents of learners’ responses. Learner IDs were used as random effects.

5. Results

5.1. Data overview

The overall mean of the submission score on the tasks from the first two assignments (“immediate tasks”) was 56.34 out of 100 ($SD = 49.61$) and on the tasks from assignments 3, and 4 (“delayed tasks”) was 46.29 out of 100 ($SD = 35.66$). Means and standard deviations of the submission scores per condition are given in Table 2. It is important to note that assignments 1 and 2 were split into single tasks (i.e., tasks 1–1, 1–2, 1–3, 2–1, 2–2, and 2–3) and therefore scores of these tasks were either 0 or 100, which explains large standard deviations. On the other hand, assignments 3 and 4 were not split into individual tasks but were composed of multiple tasks (i.e., assignment 3 composed of 13 tasks).

In terms of submission counts which were assumed to follow Poisson distribution, the first quartile (Q1), median, and third quartile (Q3) along with means and standard deviations of submission counts per condition are presented in Table 3. The first quartile, median, and third quartile of the overall submission counts on the immediate tasks were 1, 1, and 2, while they were 5, 9, and 19.25 for the delayed tasks. The average submission counts of the immediate tasks and the delayed tasks were respectively 1.96 ($SD = 1.78$) and 13.88 ($SD = 12.69$). Learners’ average enjoyment score was 17.67 out of 21 ($SD = 3.56$) and the average score of perceived learning was 24.61 out of 28 ($SD = 3.52$). Means and standard deviations for these scores per condition are shown in Table 4.

There was no significant difference in learners’ metacognitive skills between conditions prior to the present study ($p = 0.19$). Furthermore, all 132 learners used hints, and there was no significant difference between conditions regarding how many hints each learner used per task. Table 5 shows descriptive statistics for the number of hints learners used and p-values from the statistical comparison between the two conditions. Finally, a mixed-effect logistic regression revealed that there was no significant difference in the effectiveness of each prompt question in evoking reflection; there was no significant difference between the number of meaningful reflection responses for each prompt question at 0.05 level. About 67% of responses were categorized as meaningful, and each prompt question respectively evoked 62.82%, 68.80%, and 69.32% of meaningful reflection responses.

Table 2
Mean (standard deviation) of assignment score per condition.

Assignment/task labels	Hint M (SD)	Hint-reflection M (SD)
Task 1–1	70.73 (45.68)	73.52 (44.33)
Task 1–2	70.58 (45.75)	88.15 (32.52)
Task 1–3	35.46 (47.95)	41.87 (49.49)
Task 2–1	59.23 (49.33)	70.83 (45.69)
Task 2–2	58.51 (49.45)	70.32 (45.93)
Task 2–3	42.04 (49.05)	43.62 (49.75)
Total (immediate task)	53.38 (49.91)	60.23 (48.97)
Assignment 3	42.94 (36.28)	51.69 (34.39)
Assignment 4	41.44 (35.37)	50.23 (34.82)
Total (delayed task)	42.59 (36.06)	51.34 (34.49)

Table 3
Median (25% quartile-75% quartile) and mean (standard deviation) of submission counts per condition.

Assignment/ task labels	Hint Median (Q1-Q3)	Hint-reflection Median (Q1-Q3)	Hint M (SD)	Hint-reflection M (SD)
Task 1–1	1.00 (1.00–2.00)	1.00 (1.00–2.00)	1.75 (1.06)	1.64 (1.34)
Task 1–2	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.70 (0.99)	1.22 (0.49)
Task 1–3	2.00 (1.00–3.00)	2.00 (1.00–3.00)	2.90 (2.76)	2.58 (2.17)
Task 2–1	1.00 (1.00–2.00)	1.00 (1.00–1.00)	1.85 (1.54)	1.73 (1.32)
Task 2–2	1.50 (1.00–2.00)	1.00 (1.00–2.00)	1.92 (1.27)	1.46 (1.01)
Task 2–3	2.00 (1.00–3.00)	2.00 (1.00–3.00)	2.51 (2.84)	2.40 (1.79)
Total (immediate task)	1.00 (1.00–2.00)	1.00 (1.00–2.00)	2.10 (1.97)	1.81 (1.53)
Assignment 3	16.50 (7.00–30.75)	16.00 (6.00–25.00)	20.24 (16.59)	16.59 (11.59)
Assignment 4	9.00 (4.50–11.00)	6.00 (3.00–8.00)	8.66 (6.81)	5.65 (5.29)
Total (delayed task)	11.00 (6.00–23.00)	8.00 (5.00–18.50)	15.36 (14.30)	12.27 (10.51)

Table 4
Means (standard deviations) of perceived learning and enjoyment per condition.

Assignment/task labels	Hint M (SD)	Hint-reflection M (SD)
Perceived learning	23.87 (4.06)	25.45 (2.57)
Enjoyment	17.14 (3.86)	18.28 (3.13)

5.2. Study results

RQ1 and RQ2 asked about the immediate effect (RQ1) and the delayed effect (RQ2) of a combined hint and reflective prompt intervention on transfer task performance. A linear mixed-effect model analysis revealed that there was a significant difference between conditions, at 0.05 level with medium effect size, in the task performance ($\chi^2 = 9.63, p < 0.01$, Cohen’s $d = 0.54$). There was also a significant difference between assignment labels (immediate vs delayed) in the task performance at.05 level with small effect size ($\chi^2 = 47.08, p < 0.001$, Cohen’s $d = 0.20$). However, there was no significant interaction between the conditions and assignment labels ($\chi^2 = 1.17$, $p = 0.27$, Cohen’s $d = 0.02$). Table 6 presents a summary of fixed effects. Tukey’s post hoc test confirmed that the immediate task scores of the hint-reflection group were significantly higher than that of the hint group at.05 level with less than a small effect size ($\widehat{\beta} = -6.49, SE = 3.07, p = 0.03$, Cohen’s $d = 0.16$). The delayed task scores of the hint-reflection group were significantly higher than those of the hint group at 0.05 level with a small effect size ($\widehat{\beta} = -9.19, SE = 2.80, p < 0.01$, Cohen’s $d = 0.23$).

RQ3 and RQ4 asked how the interventions affected the number of immediate and delayed task submissions. A generalized linear mixed-

Table 5
Means (standard deviations) of the number of hints used.

	Task 1–1 M (SD)	Task 1–2 M (SD)	Task 1–3 M (SD)	Task 2–2 M (SD)	Task 2–2 M (SD)	Task 2–3 M (SD)
Hint	2.42 (1.15)	1.88 (0.66)	3.23 (1.45)	2.68 (0.90)	2.56 (0.75)	1.92 (0.62)
Hint-reflection	2.21 (0.96)	2.00 (0.81)	3.06 (1.60)	2.40 (1.10)	2.73 (0.60)	1.91 (0.66)
Total	2.32 (1.06)	1.93 (0.72)	3.15 (1.51)	2.55 (0.99)	2.63 (0.69)	1.91 (0.64)
p-value	0.408	0.593	0.638	0.258	0.328	0.940

Table 6
Estimates, standard error, and t-value of fixed effects in the linear mixed-effect model for RQs 1 and 2.

Fixed effects	Estimate	Standard Error	t-value
(Intercept)	56.73	2.07	27.40
Hint-reflection condition	6.48	3.07	2.11
Delayed assignment	−9.65	1.64	−5.88
Hint-reflection:delayed assignment	2.70	2.49	1.08

effect model analysis reported no significant difference between conditions. There was a significant difference between submission counts of immediate assignments and delayed assignments ($p < 0.001$). Table 7 shows a summary of the model analysis.

RQ5 asked if the interventions affected perceived learning after immediate tasks. Welch's t-test revealed that the hint-reflection group reported significantly higher perceived learning than the hint group with a small effect size ($t = 2.63, p < 0.01$, Cohen's $d = 0.46$). Lastly, RQ6 asked how much learners enjoyed assignments 1 and 2. Welch's t-test showed a significant difference at 0.05 level between the hint and the hint-reflection condition, but the follow-up Tukey revealed that there was no statistically significant difference in enjoyment (at alpha = 0.05) between these conditions ($t = 1.84, p = 0.06$, Cohen's $d = 0.32$).

6. Discussion

6.1. General discussion

In this study, we investigated how reflection prompts support learning with hints in a field experiment. Specifically, we focused on the differences between the hint condition and the hint-reflection condition. To answer research questions, we compared immediate and delayed performance on knowledge transfer tasks (RQ1, RQ2), the number of immediate and delayed task submissions (RQ3, RQ4), perceived learning (RQ5), and enjoyment (RQ6) between conditions.

The results of this study indicate that the combination of hints and reflection prompts had a positive impact on both immediate and delayed transfer task performance compared to the hint-only condition (RQ1, RQ2), and did not significantly increase the number of submission attempts (RQ3, RQ4). The positive effects of the reflection prompt on the transfer task complexity are consistent with previous works (Krause & Stark, 2010; Lin & Lehman, 1999; Schworm & Renkl, 2002; Lew & Schmidt, 2011; Bannert, 2006; Bannert & Reimann, 2012; Bannert et al., 2015). Furthermore, enjoyment, a construct which has been rarely studied, was not negatively affected by the additional reflection prompt intervention, that is, there was no difference between the hint and hint-reflection conditions in terms of enjoyment (RQ6). Finally, learners' perceived learning was significantly higher in the hint-reflection condition (RQ5).

Table 7
Estimates, standard error, and z-value of fixed effects in the generalized linear mixed-effect model for RQs 3 and 4.

Fixed effects	Estimate	Standard Error	z-value
(Intercept)	0.63	0.06	9.44
Hint-reflection condition	−0.15	0.09	−1.56
Delayed assignment	1.96	0.04	48.08
Hint-reflection:delayed assignment	−0.04	0.06	−0.66

Compared to the hints-only group, learners in the hint-reflection condition achieved significantly higher performance with a similar number of submissions. That is, the learners who were given both hints and reflection prompts did not simply achieve better performance by submitting more assignments; in other words, their behavior could not be considered an attempt to game the system. Furthermore, there was no significant difference in the number of hints used between the conditions, which further supports our finding that reflection prompts helped learners understand and apply hints more effectively and efficiently to the tasks.

Furthermore, the benefits of reflection prompts lasted even after they were removed, as evidenced in the analysis for RQ2. Considering that each task was conceptually built on the previous tasks, the reflection prompts might have supported learners to acquire and maintain knowledge from immediate tasks better. The other compatible explanation might be that by being exposed to the reflection prompts, learners in the hint-reflect condition learned to reflect after a task and thus would do so even without reflection prompts. Learners with this new desirable learning skill might have had a better sense of what their current approach to the tasks was and how they could improve it if that failed. This should be investigated further in future work.

Additionally, our results suggest that there may be both a timing effect and a priming effect as learners progress through the assignments. Table 2 shows the mean and standard deviation of assignment scores for each condition. While the difference between the two conditions is minor in the first row (Task 1–1), the scores tend to diverge more as the assignments progress. A reasonable interpretation can be that reflective prompts may prime learners to engage in reflective thinking, and that the impact of these prompts becomes more pronounced with increased exposure and interaction over time. These findings suggest that reflection prompts influenced learners' use of hints, and that this effect cannot be attributed solely to the presence of reflection prompts without considering their impact on hint usage patterns. In summary, our findings support the idea that reflection prompts have a meaningful impact on learners' hint usage behavior during learning.

6.2. Practical implications

The current findings show implications on how to encourage learners to meaningfully interact with hints. The first implication is the importance of reflection. Previous studies have been consistent in showing the ineffectiveness of hints in increasing performance due to learners' mindless and passive approach to hints (Aleven et al., 2006; Roll et al., 2011). Our results showed that the reflection prompts effectively address the issue by encouraging learners to actively interact with hints, by reviewing how they have applied hints while solving the given problems and what they learned from the hints.

The study also provides a practical design implication on reflection prompts which could be easily adopted and used. Most previous studies designed a system which displayed reflection prompts more than once or even showed reflection prompts after every small learning events (Krause & Stark, 2010; Bannert, 2006; Bannert & Reimann, 2012;

Bannert et al., 2015; van den Boom et al., 2004). While the intention was to maximize the effects of reflections by evoking them as many times as possible, our pilot study showed that learners are highly likely not to comply with frequently presented reflection prompts due to annoyance. Our main study showed that presenting a reflection prompt only once at the end of each task still significantly increases task performance without losing the feeling of enjoyment or perceived learning. That is, these findings provide practical design implications for authentic learning environments.

Combining hints and reflection prompts offers a practical and scalable approach for instructors who lead large-scale online courses. These instructors often need more time or resources to provide individualized hints and monitor learners' engagement with them. By including a reflection prompt at the end of each activity that includes hints, instructors may facilitate deeper learning and engagement without requiring additional effort.

This study raises several questions for future research. First, given the importance of reflection in the context of hint use, it would be valuable to investigate using incentive structures to encourage the reflection process at the end of each activity. Furthermore, it is equally important to study whether repeated exposure to reflection prompts could lead learners to respond automatically without thinking. While the positive impact of reflection prompts was demonstrated in this study, it is not uncommon for repeated interventions to become ineffective or even numbing to learners. One potential solution to explore is training learners to engage in metacognitive reflection through practice, so that they would not need to see additional prompts and avoid feeling numb. Overall, this study sets the stage for further investigation into the limits and leverages of reflection prompts.

6.3. Limitations

We acknowledge that not all of our findings reached the level of medium effect size, while the study sample size was originally intended for measurement of medium effects (Cohen, 2013).

While our study involved a sufficiently large sample for our analysis, it is limited in its generalizability. The sample studied largely consisted of part-time students who were pursuing an online data science degree program. At the population level, such students differ in terms of time constraints, motivation for learning, and demographics from full-time undergraduate students who are typical participants of research experiments (a challenge for the field in and of itself (Henrich, Heine, & Norenzayan, 2010)). Therefore, it would be meaningful to replicate the study with different learner populations. In addition to demographic, modality, and disciplinary sensitivity, another particularity of this study is its reliance on an auto-grading system that provides learners with relatively immediate feedback on their submissions. This opens the question if the same design features would be applicable to open-ended online learning tasks, where such prompt feedback would not be feasible; a research direction worth exploring in future work. It should also be mentioned that the study used a pool of hints that was limited in size. This limited pool of hints might not have served every learner's needs, despite our efforts to identify learners' frequent questions through the pilot study and then generate hints that answer those questions.

In this study, we did not examine the effect that the complexity of transfer tasks might have on the effectiveness of reflection prompts. While the literature review shows that learners generally benefit more from reflection prompts when working on more complex tasks such as transfer tasks than on simpler tasks such as recall or retention, there could be a difference in the impact of reflection prompts on various types of transfer tasks, which is an interesting question to investigate in future research.

Finally, a qualitative analysis of the reflection writings of learners may shed light on both the quality of and the reasoning behind the learners' reflection processes. Future work examining this may uncover whether there were specific kinds of reflection stances or mindsets which lead to increased impact.

7. Conclusion

This work investigated the immediate and delayed effects of the combined use of reflection prompts and hints on performance in transfer tasks, perceived learning, and satisfaction in the domain of programming education. We have demonstrated that only when reflection prompts and hints were combined, immediate task performance, delayed task performance, and perceived learning increased. This study provides critical practical implications on how to design a programming learning environment that can lead to deeper learning from hints.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aleven, V., Mclaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education, 16*, 101–128.

Baker, R. S. J.d., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, I., Wagner, A. Z., et al. (2006). Adapting to when students game an intelligent tutoring system. In *Intelligent Tutoring Systems* (pp. 392–401). Springer Berlin Heidelberg.

Bannert, M. (2006). Effects of reflection prompts when learning with hypermedia. *Journal of Educational Computing Research, 35*, 359–375.

Bannert, M., & Reimann, P. (2012). Supporting self-regulated hypermedia learning through prompts. *Instructional Science, 40*, 193–211.

Bannert, M., Sonnenberg, C., Mengelkamp, C., & Pieger, E. (2015). Short-and long-term effects of students' self-directed metacognitive prompts on navigation behavior and learning performance. *Computers in Human Behavior, 52*, 293–306.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*, 612.

Barzilai, S., & Blau, I. (2014). Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences. *Computers & Education, 70*, 65–79.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L., & Rellinger, E. R. (1995). Metacognition and problem solving: A process-oriented approach. *The Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 205–223.

Berthold, K., Nückles, M., & Renkl, A. (2007). Do learning protocols support learning strategies and outcomes? the role of cognitive and metacognitive prompts. *Learning and Instruction, 17*, 564–577.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *The Annual Review of Psychology, 64*, 417–444.

Blank, D., Bourgin, D., Brown, A., Bussonnier, M., Frederic, J., Granger, B., et al. (2019). nbgrader: A tool for creating and grading assignments in the jupyter notebook. *The Journal of Open Source Education*.

Boekaerts, M. (1999). Self-regulated learning: Where we are today. *International Journal of Educational Research, 31*, 445–457.

van den Boom, G., Paas, F., van Merriënboer, J. J. G., & van Gog, T. (2004). Reflection prompts and tutor feedback in a web-based learning environment: effects on students' self-regulated learning competence. *Computers in Human Behavior, 20*, 551–567.

Boud, D., Keogh, R., & Walker, D. (1985). What is reflection in learning. *Reflection: Turning Experience into Learning*, 7–17.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school.* National Academy Press.

Bye, L., Smith, S., & Rallis, H. M. (2009). Reflection using an online discussion forum: Impact on student learning and satisfaction. *Social Work and Education, 28*, 841–855.

Chen, N. S., Kinshuk, Wei, C. W., & Liu, C. C. (2011). Effects of matching teaching strategy to thinking style on learner's quality of reflection in an online learning environment. *Computers & Education, 56*, 53–64.

Clark, R. E. (1982). Antagonism between achievement and enjoyment in ATI studies. *Educational Psychology, 17*, 92–101.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Coulson, D., & Harvey, M. (2013). Scaffolding student reflection for experience-based learning: a framework. *Teaching in Higher Education, 18*, 401–413.

Davis, E. A. (2003). Prompting middle school science students for productive reflection: Generic and directed prompts. *Journal of the Learning Sciences, 12*, 91–142.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist, 47*, 153–176.

Devolder, A., van Braak, J., & Tondeur, J. (2012). Supporting self-regulated learning in computer-based learning environments: systematic review of effects of scaffolding in the domain of science education. *Journal of Computer Assisted Learning, 28*, 557–573.

Dewey, J. (1997). *How We Think*. Courier Corporation.

Dong, Y., Marwan, S., Shabrina, P., Price, T., & Barnes, T. (2021). *Using student trace logs to determine meaningful progress and struggle during programming problem solving*. International Educational Data Mining Society.

Efklides, A. (2008). Metacognition. *European Psychologist, 13*, 277–287.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.

Green, P., & MacLeod, C. J. (2016). Simr: an r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*, 493–498.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–83. discussion 83–135.

Hung, I. C., Yang, X. J., Fang, W. C., Hwang, G. J., & Chen, N. S. (2014). A context-aware video prompt approach to improving students' in-field reflection levels. *Computers & Education, 70*, 80–91.

Ifenthaler, D. (2012). Determining the effectiveness of prompts for self-regulated learning in problem-solving scenarios. *Journal of Educational Technology & Society, 15*, 38–52.

Jeong, H., Gupta, A., Roscoe, R., Wagster, J., Biswas, G., & Schwartz, D. (2008). Using hidden markov models to characterize student behaviors in Learning-by-Teaching environments. In *Intelligent Tutoring Systems* (pp. 614–625). Berlin Heidelberg: Springer.

Kim, J., & Kendeou, P. (2021). Knowledge transfer in the context of refutation texts. *Contemporary Educational Psychology, 10*, Article 102002.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., et al. (2016). Jupyter notebooks-a publishing format for reproducible computational workflows. In *ELPUB* (pp. 87–90).

Kramarski, B., & Kohen, Z. (2017). Promoting preservice teachers' dual self-regulation roles as learners and as teachers: Effects of generic vs. specific prompts. *Metacognition and Learning, 12*, 157–191.

Kramarski, B., Weiss, I., & Sharon, S. (2013). Generic versus context-specific prompts for supporting self-regulation in mathematical problem solving among students with low or high prior knowledge. *Journal of Cognitive Education and Psychology, 12*, 197–214.

Krause, U. M., & Stark, R. (2010). Reflection in example- and problem-based learning: effects of reflection prompts, feedback and cooperative learning. *Educational Research and Evaluation, 23*, 255–272.

Lew, M. D. N., & Schmidt, H. G. (2011). Self-reflection and academic performance: is there a relationship? *Advances in Health Sciences Education: Theory and Practice, 16*, 529–545.

Lin, X., & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 36*, 837–858.

Marwan, S., Williams, J. J, & Price, T. (2019). An evaluation of the impact of automated programming hints on performance and learning. In *Proceedings of the 2019 ACM Conference on International Computing Education Research* (pp. 61–70). New York, NY, USA: Association for Computing Machinery.

Messer, M. (2022). Detecting when a learner requires assistance with programming and delivering a useful hint. In *Proceedings of the 15th International Conference on Educational Data Mining* (p. 778).

Moon, J. A. (2013). *A handbook of reflective and experiential learning: Theory and practice*. London, England: Routledge.

Muñoz-Merino, P. J., Valiente, J. A. R., & Kloos, C. D. (2013). Inferring higher level learning information from low level data for the khan academy platform. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 112–116). New York, NY, USA: Association for Computing Machinery.

Nokes, T. J., Hausmann, R. G. M., VanLehn, K., & Gershman, S. (2011). Testing the instructional fit hypothesis: the case of self-explanation prompts. *Instructional Science, 39*, 645–666.

O'Rourke, R. (1998). The learning journal: from chaos to coherence. *Assessment & Evaluation in Higher Education, 23*, 403–413.

Pan, Z., & Liu, M. (2022). The effects of learning analytics hint system in supporting students problem-solving. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 77–86). New York, NY, USA: Association for Computing Machinery.

Pintrich, P. R. (1991). A manual for the use of the motivated strategies for learning questionnaire (MSLQ).

Richardson, G., & Maltby, H. (1995). Reflection-on-practice: enhancing student learning. *The Journal of Advanced Nursing, 22*, 235–242.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*, 267–280.

Schworm, S., & Renkl, A. (2002). Learning by solved example problems: Instructional explanations reduce self-explanation activity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. escholarship.org.

Shih, B., Koedinger, K. R., & Scheines, R. (2011). A response time model for bottom-out hints as worked examples. *Handbook of Educational Data Mining*, 201–212.

Veenman, M. V. J. (2011). Learning to self-monitor and self-regulate. In R. E. Mayer, & P. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 197–218). New York: Routledge.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *American Statistician, 70*, 129–133.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p <$ 0.05". *American Statistician, 73*, 1–19.

Zepeda, C. D., Richey, J. E., Ronevich, P., & Nokes-Malach, T. J. (2015). Direct instruction of metacognition benefits adolescent science learning, transfer, and motivation: An in vivo study. *Journal of Educational Psychology, 107*, 954.