Does autonomy help Help? The impact of unsolicited hints on help avoidance and performance

Christa Cody*
North Carolina State University
cncody@ncsu.edu

Mehak Maniktala
North Carolina State University
mmanikt@ncsu.edu

David Warren
North Carolina State University
dswarre2@ncsu.edu

Min Chi
North Carolina State University
mchi@ncsu.edu

Tiffany Barnes
North Carolina State University
tmbarnes@ncsu.edu

ABSTRACT

Research has shown that autonomy can be beneficial to both learning and motivation; however, limited research has explored unsolicited hints impacts on students' autonomy. Furthermore, some research has shown that unsolicited hints can improve student learning while other research suggests that on-demand hints are more beneficial. In this study, we compare three types of student autonomy regarding hints: 1) *Control*, with on-demand hints, 2) *Choice*, with periodic popups asking whether the student would like a hint, and 3) *Assertions*, with periodic unsolicited hints. We found that the Control and Assertion groups performed similarly, and significantly better on the post-test than Choice. Further, the Assertions group had the fewest steps where help was needed but was not received, effectively solving the help avoidance problem. Overall, our results suggest that unsolicited hints can effectively ensure that more help is delivered when it is needed, reducing autonomy without reducing learning.

1 Introduction

Although research has shown that allowing students to have autonomy while learning a new domain can benefit learning[6, 18, 19, 17, 7], studies have shown that students many not have the required skills to self-regulate their learning to seek help appropriately [2, 13, 22, 12, 3, 2]. Further, research has shown that students often cannot make effective decisions regarding when they need a hint [22]. Students lacking help-seeking abilities often partake in help avoidance, where they do not use assistance available in a tutoring system [1, 15]. To address help avoidance, some ITSs employ proactive assistance[21]. While one paper found that on-demand assistance, where students have to request hints, produced better learning outcomes [16], other studies have shown that providing tutor-initiated, unsolicited hints at the appropriate time, i.e. with no student autonomy about hints, can augment students' learning experience and improve performance [5, 14, 4].

The goal of this work is to investigate whether unsolicited hints can solve the help avoidance problem. We compare three groups: 1) *Control*, with on-demand hints, 2) *Choice*, where students were periodically asked if they would like a hint, and 3) *Assertions*, where unsolicited hints were periodically added to the student's workspace without any element of student choice. The Assertions group provided students with the least autonomy regarding when to receive a hint, by adding unsolicited hints to the workspace. Students may ignore these hints, but as they are the most efficient next step, students avoiding them will have less efficient solutions. The Choice group is the middle ground for hint autonomy because students can choose not to receive a hint. Due to the need to make a help-seeking decision, we consider this group to have a medium level of hint autonomy. The Control group is considered the most autonomous because they control the entire interaction surrounding hints. Overall, we hypothesize that the benefit of receiving help when it is needed outweighs the negative impact of removing student autonomy about when and whether to receive a hint.

We constructed the following hypotheses based on prior work in Deep Thought, a logic tutor, and research in students' self-regulation abilities: H_1, Assertions will increase the chances of receiving help when it is needed, while not harming performance; H_2, the Choice group will demonstrate more help avoidance than the Assertion group and worse performance in the posttest due to bad self-regulation choices; and H_3, the Control group will also demonstrate more help avoidance than the Assertion group, and take longer in the training, but have similar performance in the posttest.

2 Deep Thought, our logic tutor

Our propositional logic tutor, Deep Thought, [11] presents proof problems as a set of given logic statements, shown at the top of the workspace and a conclusion to be derived at the bottom of the workspace (see Figure 1). Students solve problems by iteratively deriving new logic statement nodes until they derive and justify the conclusion. To create a

*This material is based upon work supported by the National Science Foundation under Grant No. 582690.

new statement node, students first 'justify' it by selecing 1-2 existing nodes and a rule to apply to them. The tutor is divided into an introduction, pretest, training, and posttest. The introduction includes two worked examples where students click through the derivation and justification of all the nodes, followed by one practice problem to learn the interface. Next, a student takes the pretest problem, which we use to compare the student's incoming proficiency for stratified sampling (see Section 3). Next, the tutor guides students through the **training** section (15 problems) with varying difficulty, where students can request and receive hints. Finally, students take a more difficult non-isomorphic **posttest**, where all students must solve the same set of 4 problems without any tutor assistance. Throughout the tutor, including the pre- and post-test problems, our logic proof tutor provides immediate error feedback for rule application mistakes.

Figure 1: The Deep Thought interface.

2.0.1 Assistance

The tutor uses a data-driven approach based on a modified version of Hint Factory [20, 9] to generate hints from historical student data, resulting in hints on the most *frequent* and *efficient* paths available based on the student's current attempt. Hints provided in the training of the tutor can either be initiated by the student, in which case they are called *on-demand* hints, or they can be initiated by the tutor, in which case they are called *unsolicited* hints. For our hints, we used our recently-designed Assertions interface [9] to place next-step hints in the workspace, which are the next, best statement that can be derived in one rule-application step from the student's current state, as blue nodes marked with a question mark (denoting that they have not been justified) and a 'Goal' label. Although each group received hints through the Assertions interface for a fair comparison, later iterations of the tutor use the Assertion interface only for unsolicited hints. Hints do not tell students which rules or prior nodes can be used to justify the suggested statement and are designed to help students solve problems by suggesting a subgoal statement to help them break down multi-step problems.

3 Methods

The tutor was used as a mandatory, online homework assignment by students in an undergraduate discrete mathematics for computer scientists course (Spring 2019). For this study, we compared 94 students' data from three conditions to investigate the impact of student-choice on performance and behavior. The three conditions were 1) **Control**, 2) **Choice**, and 3) **Assertions**. While all conditions allowed on-demand hints, they differed slightly in unsolicited help. The **Control** group represents the normal conditions in Deep Thought with no unsolicited hints. The Choice group was asked "Would you like a suggestion?" after completing approximately every third step to expose poor self-regulating decisions. We chose this amount to be frequent enough to be comparable to the Assertion group, but not distracting. The Assertions group received periodic unsolicited hints on approximately 40-50% of the steps to produce assistance similar to a partially worked example, or turn-taking tutor where the tutor and the student co-construct a solution to the problem.

We used stratified sampling, splitting students by pretest performance, then randomly assigning them to Assertions ($n = 38$), Choice ($n = 27$), and Control ($n = 29$) to ensure all conditions were balanced in incoming knowledge. The Assertions group was designed to have a slightly larger size to ensure sufficient data collection, and since we felt that this condition would be more beneficial to students than the Choice or Control conditions.

We used each student's pretest **score** to measure incoming knowledge. A student's **score** is a combination of normalized metrics for the pretest *time*, number of *steps*, and *accuracy* on a single problem, which ranks a student based on how fast, efficient, and accurate they are compared to their current peers. To investigate student's performance, we focused on time spent solving a problem, total attempted steps, and accuracy. **Total time** is counted from the moment a problem is on the screen until it is solved by deriving and justifying the conclusion. **Total steps** in a problem include any attempt a student makes at deriving a new step, which includes both correct and incorrect steps (node derivations). **Accuracy** is the total number of correct rule applications divided by all rule application attempts. Note that the tutor is not designed or assumed to promote large improvements in accuracy, since no penalties are assigned for incorrect rule applications, even within the pre- and posttests. We focus on steps and time per problem because it is more difficult for students to learn to determine which steps to derive to achieve shorter, more efficient proofs. Whereas, learning how to apply the rules can be done by memorization and simple practice.

Data were analyzed to compare groups for the pretest, training, and posttest portions of the tutor. ANOVA with Tukey's post hoc tests were used to examine the significance of differences in the means of the populations between pretest groups with Benjamini-Hochberg corrections. For training and posttest metrics, we applied one way ANCOVA using the pretest as a covariate. To check that the data met assumptions, we used the the Shapiro-Wilk's W test, Levene's test, Q-Q plots, and histograms. Data that did not meet the assumptions were transformed using log or square-root transformations, then re-inspected. Data reported in tables are before transformation for clarity. For all tables, at least marginally significant values are bolded ($p <= 0.10$), and significant values are marked with an asterick (* for $p <= 0.05$).

4 Results & Discussion

This section discusses the comparison between the Assertion, Choice, and Control groups, and the differences in performance between students.

4.1 Hint Usage and Help Need

To understand each group's utilization of hints, we examined hint-related metrics. **# Hint Requests** is the total number of hints requested in training. **Hints Received** is the total hints a student received during the tutor, unsolicited and requested. For the Choice group, **# Asked** represents how often they were asked if they would want a hint. Hint **Justification** rate is the percentage of hints received that students connected to their current solution through justification. Table 1 shows the mean, standard deviation and Tukey HSD's results for the hint metrics. ANOVA showed a significant difference in the mean # Hints Received ($F(2,91) = 25.576, p < 0.01$) between the groups. Tukey Contrasts analysis showed significant differences among each comparison (Control-Choice ($p < 0.01$); Choice-Assertions ($p < 0.01$); and Control-Assertions ($p < 0.01$)). We expected these differences because the Assertions group was given frequent, unsolicited hints, the Choice group was asked if they wanted a hint at a slightly lower frequency and was only given a hint if they selected 'Yes', and the Control group received hints only upon request. Since all three groups could request on-demand hints in addition to any the tutor might provide or offer, we compared # Hints Requests, but there were no significant differences between the 3 groups on this metric ($F(2,91) = 0.1816, p = 0.83$)).

Table 1: Mean and Standard Deviation(SD) of the Hint Usage Metrics in the Training.

	Control	**Choice**	**Assertions**
	n = 29	*n = 27*	*n = 38*
Metric	*Mean(SD)*	*Mean(SD)*	*Mean(SD)*
# Asked	34(10)	-	-
# Hints Received	**19(16)***	**35(25)***	**51(12)***
# Hint Requests	21(21)	26(27)	25(23)
Hint Justification Rate	**85%(25)**	**80%(20)***	**84%(6.5)***

The Control group justified 85% of the requested hints, on average, which makes sense as students are more likely to use the hints they request [16]. The mean Hint Justification rates were 84% for Assertions and 80% for the Choice group. ANOVA results revealed a significant difference between groups for the Hint Justification Rate ($F(2,91) = 6.0633, p < 0.01$)). Tukey Contrasts analysis showed significant differences among Control-Choice ($p = 0.03$), and Control-Assertions ($p < 0.01$)), but no significant difference between Choice and Assertions group ($p = 0.79$). This is surprising because we expected the Choice group to have a higher Justification rate than the Assertions group, since, similar to the Control, they chose to get a hint. These results suggest that unsolicited Assertions were just as well received as hints offered as a choice.

Further, we defined measures to address all three hypotheses concerning hint usage: help need, hint abuse, unnecessary hints, and steps in which they received an appropriate level of help (i.e. received a hint when needed and did not receive a hint when not needed). An important goal of this study was to investigate whether periodic unsolicited hints could address help avoidance by increasing the number of times students who needed help received it. Since our hints are partially-worked steps and students could easily ignore them, unsolicited hints should not harm students who do not need them. We determined when a hint was needed vs. not needed via our new Help-Need model described in [8, 10]. The model uses (1) the quality of the current step based on a combined productivity measure of the optimality of their current state (how close it is to the solution based on the Hint Factory [20]), and the time taken to derive it, and (2) a prediction of whether help is needed in the next step (e.g. if the next step is not predicted to be productive, then help is needed). We note that our help-need predictor is not ground truth, but our cited work shows that the Help-Need predictor is correlated with post-test performance. **% Help Needed** is the percentage of total steps our Help-Need model identified as unproductive, where a student could have benefited from a hint, and a hint was not received **% Hint Abuse** is the percent of total steps where our model predicted no Help-Need but a student *requested* a hint, representing a bad help-seeking decision. **% Unnecessary Hint** is the percent of total steps where students *received* a hint on a step where we predicted no Help-Need, including both help abuse requests and the number of times hints were given but not needed. We also included Help Abuse because we wanted to ensure none of the conditions were promoting gaming the system. **% Appropriate Hint** is the percent of steps where Help-Need model aligned with the student need (e.g. a student received a hint when they were predicted to need one or a student did not receive a hint and the model labelled the step as no help-need).

Table 2 shows the differences in these metrics between the groups. With ANCOVA, controlling for the pretest, we found a significant difference between the groups for % Unnecessary Hints ($F(2,91) = 38.35, p < 0.01$) and % Help Needed ($F(2,91) = 10.11, p < 0.01$). For % Unnecessary Hints, Tukey Contrasts analysis revealed significant differences between all 3 groups: Choice-Control ($p = 0.01$), Choice-Assertions ($p < 0.01$), and Control-Assertions ($p < 0.01$). For % Help Needed with the same procedure, we found significant differences between Choice-Assertions ($p = 0.01$) and Control-Assertions ($p < 0.01$); however, there was no significant difference in Control-Choice ($p = 0.45$). There were no significant differences for Hint Abuse ($F(2,91) = 0.04, p < 0.96$) or the Appropriate Hint metrics ($F(2,91) = 0.57, p < 0.56$).

Table 2: Mean and Standard Deviation(SD) of the Help Need Metrics in the Training.

	Control	**Choice**	**Assertions**
	n = 29	*n = 27*	*n = 38*
Metric	*Mean(SD)*	*Mean(SD)*	*Mean(SD)*
% Help needed	20(12)	16(11)	**10(8)***
% Hint Unnecessary	**4(5)***	**7(5)***	**15(4)***
% Help abuse	7(6)	9(9)	7(7)
% Appropriate Hint	72(11)	71(12)	73(7)

The Control group had the lowest percentage of steps with Unnecessary hints, which was expected since they had full autonomy and requested fewer hints than the other groups. The Control group also had the highest percentage of steps where Help-Need was detected, meaning that these students spent more time in steps being unproductive. The Choice group fell in the middle for both % Help Needed and % Unnecessary Hints. H_2, stated that the Choice group would have more help avoidance than the other two groups. The Control group showed similar help avoidance to the Choice group by not requesting hints when needed. However, the Choice group had a significantly higher Help Avoidance than the Assertion group, which provides partial evidence in support of H_2. Additionally, the Control group having a significantly higher % Help Needed partially supports H_3, in which we hypothesized that the Control group would not request hints often enough. The Assertions group decreased steps where students needed help but were not receiving it, confirming H_1. Although more unnecessary hints were provided, our goal was to reduce students being stuck in steps without receiving help, which was achieved even though the frequency of unsolicited hints was not based on an intelligent policy. Incorporating an intelligent policy to determine when to give a hint should result in an even smaller percentage of help need and reduce instances of unnecessary hints. To test whether the larger percentages of Unnecessary Hints would be worse for posttest performance, a simple linear regression was calculated to predict the posttest score based on the % Unnecessary Hints and was not significant ($F(1, 91) = 0.33, p = 0.57$). Therefore, we do not believe these Unnecessary Hints had a significant impact on performance. Another simple linear regression was calculated to predict the posttest score based on the % Help Needed, and a significant regression was found ($F(1, 91) = 8.49, p < 0.01$) providing support that addressing help need is important.

4.2 Evaluating Students' Performance Across the Tutor

To examine the effects on performance each group had, the pretest and posttest performance metrics for the 3 groups were analyzed (see Table 3). ANOVA was performed on pretest metrics to determine if there was a similar distribution of proficiency between the groups. There were no significant differences between the groups on Total Time ($F(2, 91) = 0.28, p = 0.76$) or Total Steps ($F(2, 91) = 1.01, p = 0.37$) in the pretest metrics. There was a marginally significant difference between the groups for accuracy ($F(2, 91) = 2.38, p = 0.09$), but this is not a meaningful difference due to the few number of steps in the pretest and the Choice's group lower average number of steps. Therefore, we concluded that each group had a distribution of students' with similiar incoming proficiency.

For the training and posttest performance metrics, ANCOVA was used controlling for pretest metrics. There were no significant differences between any performance metric in the training portion of the tutor (Total Time ($F(2, 90) = 2.07, p = 0.13$); Total Steps ($F(2, 90) = 1.84, p = 0.16$); Accuracy ($F(2, 90) = 1.34, p = 0.27$)). The posttest metrics show a significant difference in the Total Time ($F(2, 90) = 5.24, p < 0.01$)) between the groups. Tukey Contrast analysis revealed that there was a significant difference between the Assertion and Choice group ($p < 0.01$); however, there was not a significant difference between the Choice and Control ($p = 0.29$) or the Assertion and Control ($p = 0.19$). There was no significant difference between the Total Steps ($F(2, 90) = 2.09, p = 0.13$) or the Accuracy ($F(2, 90) = 0.05, p = 0.95$) between the groups.

These results provide support for H_1 that the students in the Assertions group would perform similarly to the Control group; however, the Control group did not perform worse in the training as expected in H_3. These results along with the results in 2 confirm H_1. Assertions reduced help need without harming performance. These results provide evidence in support of H_2; however, these results do not address why the Choice group performed worse. One theory is that the students could have been making poor self-regulated decisions, supported by Table 2, which may have made them perform worse than the Control even though they both had a choice. The prompts may have lead to the Choice group to make more help-seeking decisions than the Control, where students would have thought about hints less. However, the questions asking whether or not they would like a hint could have also been frustrating or distracting. This distraction could have caused them to lose focus; however, we would have expected the total time in the training to be significantly different in that case.

Table 3: Pretest, Training and Posttest performance metrics for the Assertion, Choice, and Control groups.

		Control $n = 29$	Choice $n = 27$	Assertion $n = 38$
Test	***Metric***	***Mean(SD)***	***Mean(SD)***	***Mean(SD)***
Pretest	*Total Time (min)*	5.8(7)	4.0(2)	6.5(6)
	Total Steps	15(30)	9(7)	11(13)
	Accuracy	40(14)	**35(14)**	**43(17)**
Training	*Total Time (min)*	137(50)	114(49)	122(62)
	Total Steps	374(126)	348(124)	323(118)
	Accuracy	63%(12)	66%(11)	66%(10)
Posttest	*Total Time (min)*	37(29)	**43(34)***	**34(20)***
	Total Steps	104(56)	129(75)	102(47)
	Accuracy	69%(12)	69%(11)	69%(11)

Lastly, one of our concerns was whether students were better at self-regulating than a random proactive policy. The Assertions group was the slowest in the pretest, but they were the fastest in the posttest, shown in Table 3. Their overall hint Justification rate was also high, shown in Table 1. Along with the results confirming H_1 in the Table 2 and Table 3, these results suggest that the Assertions group with unsolicited, tutor-initiated hints did no harm to students in terms of learning outcomes compared to the Control group and produced better learning outcomes than the Choice group. Therefore, these results suggest that proactively adding hints at the very least did no harm.

5 Conclusion

This work contributes an investigation of the effects of three groups with varying levels of autonomy of assistance on learning outcomes and metrics to evaluate hint usage and hint avoidance. The three groups from most autonomous to least: 1) Control, where students could request on-demand hints, 2) Choice, where students were periodically asked if they would like a hint, and 3) Assertions, where hints were periodically added to the student's workspace without any element of student choice. This study sought to determine whether students' autonomy over when and how the interface provides hints affects hint utilization and, in turn, overall success. Our results show that the Assertion and Control group produce similar learning outcomes; however, the Choice group performed worse on the posttest. Overall, our results suggest that unsolicited hints can effectively ensure that more help is delivered when it is needed, reducing autonomy without reducing learning. These results demonstrate that with an effective, machine-learned proactive hint policy, better learning outcomes are possible.

6 References

[1] Aleven, V., McLaren, B., Roll, I., Koedinger, K.: Toward tutoring help seeking. In: International Conference on Intelligent Tutoring Systems, pp. 227–239. Springer (2004)

[2] Aleven, V., Mclaren, B., Roll, I., Koedinger, K.: Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. International Journal of Artificial Intelligence in Education **16**(2), 101–128 (2006)

[3] Azevedo, R., Cromley, J.G.: Does training on self-regulated learning facilitate students' learning with hypermedia? Journal of educational psychology **96**(3), 523 (2004)

[4] Bartholomé, T., Stahl, E., Pieschl, S., Bromme, R.: What matters in help-seeking? a study of help effectiveness and learner-related factors. Computers in Human Behavior **22**(1), 113–129 (2006)

[5] Bunt, A., Conati, C., Muldner, K.: Scaffolding self-explanation to improve learning in exploratory learning environments. In: International Conference on Intelligent Tutoring Systems, pp. 656–667. Springer (2004)

[6] Burton, R.R., Brown, J.S.: An investigation of computer coaching for informal learning activities. International journal of man-machine studies **11**(1), 5–24 (1979)

[7] Katz, I., Assor, A.: When choice motivates and when it does not. Educational Psychology Review **19**(4), 429 (2007)

[8] Maniktala, M., Barnes, T., Chi, M.: Extending the hint factory: Towards modelling productivity for open-ended problem-solving. In: Proceedings of the 13th International Conference on Educational Data Mining (2020)

[9] Maniktala, M., Cody, C., Barnes, T., Chi, M.: Avoiding help avoidance: Using interface design changes to promote unsolicited hint usage in an intelligent tutor. International Journal of Artificial Intelligence in Education (2020 (under review))

[10] Maniktala, M., Cody, C., Isvik, A., Lytle, N., Barnes, T., Chi, M.: Extending the hint factory for the assistance dilemma: A data-driven proactive helpneed model and intervention to improve problem solving. Journal of Educational Data Mining (2020 (under review))

[11] Mostafavi, B., Barnes, T.: Evolution of an intelligent deductive logic tutor using data-driven elements. International Journal of Artificial Intelligence in Education **27**(1), 5–36 (2017)

[12] Peña, A., Kayashima, M., Mizoguchi, R., Dominguez, R.: Improving students' meta-cognitive skills within intelligent educational systems: A review. In: International Conference on Foundations of Augmented Cognition, pp. 442–451. Springer (2011)

[13] Price, T.W., Liu, Z., Catet́e, V., Barnes, T.: Factors influencing students' help-seeking behavior while programming with human and computer tutors. In: Proceedings of the 2017 ACM Conference on International Computing Education Research, pp. 127–135. ACM (2017)

[14] Puustinen, M.: Help-seeking behavior in a problem-solving situation: Development of self-regulation. European Journal of Psychology of education **13**(2), 271 (1998)

[15] RANGANATHAN, R., VANLEHN, K., VAN DE SANDE, B.: What do students do when using a step-based tutoring system? Research & Practice in Technology Enhanced Learning **9**(2) (2014)

[16] Razzaq, L., Heffernan, N.T.: Hints: is it better to give or wait to be asked? In: International Conference on Intelligent Tutoring Systems, pp. 349–358. Springer (2010)

[17] Schank, R.C., Farrell, R.: Creativity in education: A standard for computer-based teaching. Tech. rep., YALE UNIV NEW HAVEN CT DEPT OF COMPUTER SCIENCE (1987)

[18] Schwartz, J.: Intellectual mirrors: A step in the direction of making schools knowledge-making places. Harvard Educational Review **59**(1), 51–62 (1989)

[19] Shute, V., Glaser, R., Raghavan, K.: Inference and discovery in an exploratory laboratory. Tech. rep., PITTSBURGH UNIV PA LEARNING RESEARCH AND DEVELOPMENT CENTER (1988)

[20] Stamper, J., Barnes, T., Lehmann, L., Croy, M.: The hint factory: Automatic generation of contextualized help for existing computer aided instruction. In: Proceedings of the 9th International Conference on Intelligent Tutoring Systems Young Researchers Track, pp. 71–78 (2008)

[21] Vanlehn, K.: The behavior of tutoring systems. International journal of artificial intelligence in education **16**(3), 227–265 (2006)

[22] Zhou, G., Lynch, C., Price, T.W., Barnes, T., Chi, M.: The impact of granularity on the effectiveness of students' pedagogical decisions. In: CogSci (2016)