What Makes Problem-Solving Practice Effective? Comparing Paper and AI Tutoring

Conrad Borchers(✉)[0000-0003-3437-8979], Paulo F. Carvalho[0000-0002-0449-3733], Meng Xia[0000-0002-2676-9032], Pinyang Liu[0000-0002-3842-1017], Kenneth R. Koedinger[0000-0002-5850-4768], and Vincent Aleven[0000-0002-1581-6657]

Carnegie Mellon University, Pittsburgh, PA 15213, USA
{cborcher,pcarvalh,aleven}@cs.cmu.edu,
{pinyangl,mengxia}@andrew.cmu.edu, koedinger@cmu.edu

Abstract. In numerous studies, intelligent tutoring systems (ITSs) have proven effective in helping students learn mathematics. Prior work posits that their effectiveness derives from efficiently providing eventually-correct practice opportunities. Yet, there is little empirical evidence on how learning processes with ITSs compare to other forms of instruction. The current study compares problem-solving with an ITS versus solving the same problems on paper. We analyze the learning process and pre-post gain data from N = 97 middle school students practicing linear graphs in three curricular units. We find that (i) working with the ITS, students had more than twice the number of eventually-correct practice opportunities than when working on paper and (ii) omission errors on paper were associated with lower learning gains. Yet, contrary to our hypothesis, tutor practice did not yield greater learning gains, with tutor and paper comparing differently across curricular units. These findings align with tutoring allowing students to grapple with challenging steps through tutor assistance but not with eventually-correct opportunities driving learning gains. Gaming-the-system, lack of transfer to an unfamiliar test format, potentially ineffective tutor design, and learning affordances of paper can help explain this gap. This study provides first-of-its-kind quantitative evidence that ITSs yield more learning *opportunities* than equivalent paper-and-pencil practice and reveals that the relation between opportunities and learning gains emerges only when the instruction is effective.

Keywords: intelligent tutoring systems, learning curve analysis, mathematics education

1 Introduction

While the COVID-19 pandemic has been a driving force for increased technology use in education, it is an open question to what degree post-pandemic education will be characterized by a "return to normal" or by a reform of past practices with greater technology use [1]. Therefore, it is worthwhile to better understand when technology is most useful, for example, in helping students learn. To address this need, we conducted a classroom study comparing a key TEL system against a representative non-TEL alternative: an intelligent tutoring system (ITS) versus paper.

Intelligent tutoring systems [2], one form of TEL, are increasingly used in educational practice (MATHia, ALEKS, ASSISTments, iReady) [3, 4, 5, 6]. These systems support the deliberate practice of challenging cognitive skills. Many scientific studies confirm the overall effectiveness of ITSs and curricula with ITSs, relative to other TEL and non-TEL environments [7, 8, 9]. Scientific studies have also revealed why ITSs are effective, particularly through (a) step-level feedback and hints within problems that respond to individual learners' strategies and errors [9, 10, 11] and (b) individualized mastery learning based on the assessment of individual knowledge growth [12]. Yet, ITSs have not always been found to be effective [13, 14, 15].

As pen-and-paper practice continues to be prominent in K-12, it is worthwhile to compare tutor practice to paper practice to better understand factors affecting learning in ITSs and math problem-solving practice. Yet, there is a dearth of studies that elucidate practice differences between ITS and paper. Instead, comparisons of ITSs against non-TEL controls have often focused on comparing curricula with varying math content and instructional activities [16] or only compared learning gain or grade outcome differences [17, 18]. What is lacking is a comparison that measures learning process differences next to outcomes. Such a process-level comparison might yield insight into the differential learning affordances of paper and ITS practice. More broadly, it may inform when, how, and for what instructional content the affordances of an ITS, versus those of paper, might be most helpful for learners. Problem-solving on paper represents a lower level of instructional assistance [19] compared to ITSs' assistance (step-level tutoring, individualized mastery learning). The lower level of paper assistance may make it harder for students to complete steps successfully and may slow them down. Further, working on paper, students could skip steps (i.e., make omission errors), reducing the number of successful practice opportunities. Then again, paper practice might have unexpected affordances for learning, for example, freedom in strategy choice. Conversely, high assistance (as in an ITS) may have certain costs. For example, it might open the door to students' gaming-the-system, that is, systematic attempts to get problem steps right by exploiting features in the software that aim to support learning [20].

Process-level data is likely to help understand these differences better. A key theoretical assumption that has emerged from past ITS research is that *eventually-correct steps* are instrumental in student learning from deliberate practice [12, 21, 36]. Following past research, we consider a competency to be learned as comprising multiple Knowledge Components (KCs) [22]. We then consider a "practice opportunity" to be attempts to complete a step in a practice problem that "exercises" one or more KCs, including any feedback or instruction toward completing that step. Such opportunities help students learn and refine the KC(s) needed to perform that step [23]. A practice opportunity is considered successful if the student eventually gets it right, with or without help (e.g., hints or feedback). To the best of our knowledge, no study has rigorously compared eventually-correct steps learning gain differences between ITS and paper practice (or other forms of non-TEL learning) for matched content. We do so using a representative ITS, Mathtutor [24] as a platform. Specifically, the current study tests the following research hypotheses:

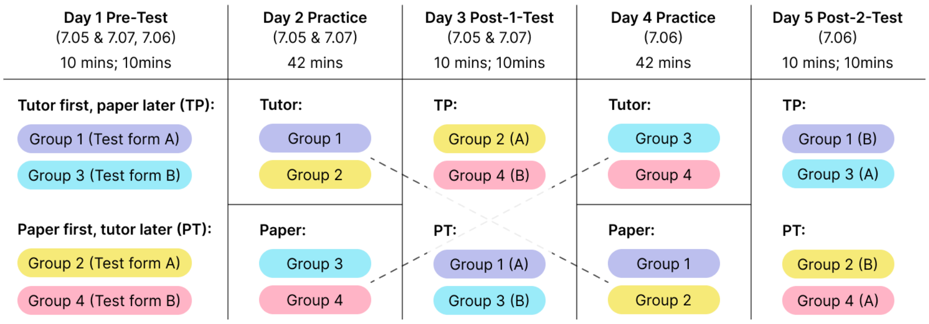
- H1: Problem-solving practice with an ITS yields greater learning gains than doing the same problems on paper due to the higher assistance in an ITS.
- H2: Working on paper, students have fewer eventually-correct KC opportunities, compared to working in an ITS, due to its high assistance.

- H3: The number of eventually-correct steps relates positively to learning gains in both conditions.

2 Method

2.1 Sample and Experimental Design

We conducted a classroom experiment at a suburban school in the United States, with 97 ninth-grade students across seven classes and three teachers participating. The study took place during students' regular mathematics class periods. It employed a within-subject design with tutor and paper practice alternating in a crossover fashion such that each student experienced both conditions [25] (see Figure 1). For each math content unit, classes were assigned to conditions such that approximately half the students solved problems using the tutoring software first, and the other half used paper first. The sequence of problem units was held constant. We opted for this design to minimize the effects of individual differences on the outcome, avoid unfairness from unequal conditions, and increase statistical power.



Day 1 Pre-Test (7.05 & 7.07, 7.06) 10 mins; 10mins	Day 2 Practice (7.05 & 7.07) 42 mins	Day 3 Post-1-Test (7.05 & 7.07) 10 mins; 10mins	Day 4 Practice (7.06) 42 mins	Day 5 Post-2-Test (7.06) 10 mins; 10mins
Tutor first, paper later (TP):	Tutor:	TP:	Tutor:	TP:
Group 1 (Test form A)	Group 1	Group 2 (A)	Group 3	Group 1 (B)
Group 3 (Test form B)	Group 2	Group 4 (B)	Group 4	Group 3 (A)
Paper first, tutor later (PT):	Paper:	PT:	Paper:	PT:
Group 2 (Test form A)	Group 3	Group 1 (A)	Group 1	Group 2 (B)
Group 4 (Test form B)	Group 4	Group 3 (B)	Group 2	Group 4 (A)

Fig. 1. Schematic representation of the experimental setup with three linear graph practice units 7.05, 7.06, and 7.07 across test item counterbalancing conditions A and B.

2.2 Procedure

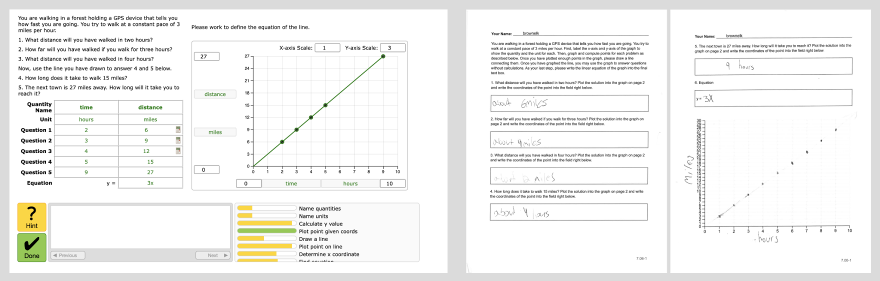
In all, the study lasted five days. On the first day, two research team members gave each class a 10-minute introduction to the tutoring software (described in Section 2.3, see Figure 2). On days one, three, and five, the students took a 20-minute pre-test, post-1-test, and post-2-test (see Figure 1). Every student took each of the three tests in two different formats. Students worked on one test format (tutor or paper, as described below) for 10 minutes, then switched to the other for the remaining 10 minutes. During the practice session on days two and four, students solved math either with the tutor or on paper for the entire 42-minute period. The post-test afterward (i.e., post-1-test on day three and post-2-test on day five) had items for just that content.

We asked the teachers to follow their usual instructional practices as closely as possible during the practice and test sessions. Based on our informal observations, teachers actively moved around the classroom in some classrooms and provided guidance and motivation. In other classrooms, teachers remained seated at their desks, offering guidance whenever students sought help. In the tutor condition, we (informally) noticed that students asked more questions and received more teachers' assistance. This was primarily due to students' unfamiliarity with the tutor or technical issues. Students were allowed to use calculators in all sessions. On paper, students could work in any order or omit any problems (though omitted steps were considered incorrect on the test). In the tutor, the system determined problem order; students needed to complete one problem before advancing to the next. During practice, correctness on all steps was mandatory for proceeding to the next problem.

2.3 Materials: Tests, Tutoring Software, and Paper Practice

Over the two practice days, students practiced problems on linear graphs. From the ITS employed in the study, Mathtutor, we selected three problem sets: 7.05 Graph Interpretation I & 7.07 Graph Interpretation II, and 7.06 Problem Solving, Equations, and Graphs I. The 7.05 problem set focused on quantitative interpretation of graphs by translating given lines to numerical x-y points in a table. Problem set 7.07 consisted of qualitative graph interpretation of relationships between multiple linear graphs. Problem set 7.06 focused on linear graph plotting, deriving the symbolic formula, and using the line to infer an x-value for a given y-value. We devoted one practice period to 7.05 and 7.07 together and one to 7.06, a problem set representing more work. During practice, the tutor (see Figure 2, left) offered step-by-step guidance through hints, correctness feedback, and feedback messages on errors. Furthermore, the tutor displayed a skill bar panel to students representing their mastery of individual KCs. During the computer-based tests, the guidance from the tutoring software was disabled; what was left was the problem-solving interface without hints, feedback, or skill bars. Each test included two problems per practice unit (7.05, 7.06, and 7.07) and was of the same type as practice problems.

We developed an analogous exercise on paper for each tutored problem. We consulted with the participating teachers regarding different exercise designs to simulate the students' default seatwork without feedback (i.e., working independently at desks) as closely as possible while generating equivalent problem-solving steps across tutor and paper for a fair comparison. An example problem from one of the tutor problem sets and its adaptation to paper are shown in Figure 2.



You are walking in a forest holding a GPS device that tells you how fast you are going. You try to walk at a constant pace of 3 miles per hour.

1. What distance will you have walked in two hours?

2. How far will you have walked if you walk for three hours?

3. What distance will you have walked in four hours?

Now, use the line you have drawn to answer 4 and 5 below.

4. How long does it take to walk 15 miles?

5. The next town is 27 miles away. How long will it take you to reach it?

Quantity Name	time	distance
Unit	hours	miles
Question 1	2	6
Question 2	3	9
Question 3	4	12
Question 4	5	15
Question 5	9	27
Equation	y =	3x

Please work to define the equation of the line.

X-axis Scale: 1 Y-axis Scale: 3

Your Name: brownelk

You are walking in a forest holding a GPS device that tells you how fast you are going. You try to walk at a constant pace of 3 miles per hour. First, label the x-axis and y-axis of the graph to show the quantity and the unit for each. Then, graph and compute points for each problem as described below. Once you have plotted enough points in the graph, please draw a line connecting them. Once you have graphed the line, you may use the graph to answer questions without calculations. As your last step, please write the linear equation of the graph into the final text box.

1. What distance will you have walked in two hours? Plot the solution into the graph on page 2 and write the coordinates of the point into the field right below.

about 6miles

2. How far will you have walked if you walk for three hours? Plot the solution into the graph on page 2 and write the coordinates of the point into the field right below.

about 9miles

3. What distance will you have walked in four hours? Plot the solution into the graph on page 2 and write the coordinates of the point into the field right below.

about 12 miles

4. How long does it take to walk 15 miles? Plot the solution into the graph on page 2 and write the coordinates of the point into the field right below.

about 4 hours

7.06-1

Your Name: brownelk

5. The next town is 27 miles away. How long will it take you to reach it? Plot the solution into the graph on page 2 and write the coordinates of the point into the field right below.

9 hours

6. Equation

y = 3X

7.06-1

Fig. 2. Tutor version (left) and paper (right) with example student responses of a tutor problem from problem set 7.06 Problem Solving, Equations, and Graph 1.

We created two test forms, A and B, with isomorphic problems (i.e., same problem type as the practice problems), so students would not see the same test items twice. We counterbalanced the test forms across the pre-test, post-1-test, and post-2-test (i.e., we randomly assigned periods to either the A-B-B or B-A-A sequence of forms on these respective tests; note that the two post-tests together covered the same range of math content as the pre-test). In addition, each test was done partly on paper and partly in the tutor to control for the effects of matching practice and test environment. The test format order was also counterbalanced.

2.4 Data Preprocessing

We analyzed the work on the exercises on paper through a novel approach of entering the paper solutions into the tutoring software; doing so generates log data comparable to that used to analyze the student's work with the tutoring software [26]. Importantly, the tutoring system evaluates the correctness of the student's steps. This way, we can analyze process data in both conditions using log data.

Two research assistants entered the students' paper responses to both test and practice items into the corresponding problem screens in the tutoring software. The order of multiple attempts at a step on paper cannot always be determined from paper responses, so coders were instructed only to enter the attempt in the tutor that they perceived to be the student's final attempt (i.e., only one attempt per step) into the tutor. In this way, to not bias the experiment in favor of the tutor, the student would get credit for self-correcting on paper. The agreement between the two coders was 93.4% based on the coding of the paper test data. Given this high agreement, we present results based on the inputs of one of the two coders in this study.

2.5 Analysis Methods

We analyze tutor log data, data from paper practice, and test data, collected partly in computer-based format and partly on paper, in terms of problem steps, which, following standard practice in the field of Educational Data Mining [23], we define as actions in the tutor interface. These include, for example, entering unit labels in a designated text field, plotting points in a graph, or selecting answers to multiple-choice questions or explanations from a menu. The tutoring system maps steps to KCs as part of its normal operation [27]. Thus, each step is considered an opportunity to apply a particular KC.

To analyze our test data results and H1 (learning gain differences across conditions), we restrict our sample to the students present during both the pre- and post-test of the respective practice section ($N = 81$ for 7.05 & 7.07; $N = 82$ for 7.06). On tests, only the last step attempts were evaluated, and steps without responses were evaluated as errors, including steps of problems that had not been started. This procedure yielded a 2 x 3 x 2 design of Practice Condition (paper vs. tutor practice), Unit (7.05 vs. 7.06 vs. 7.07), and Test Format (interface alignment vs. interface transfer). We conducted an Anova with Condition and Unit as between-subject factors and Test Format as a within-subjects factor. Learning gain functioned as the dependent variable, which we operationalized as the difference in pre-to-post step correctness averages. Two-sided t-tests were employed for post-hoc comparisons, dividing the critical threshold ($\alpha = .05$) by the number of comparisons.

To test H2 (fewer eventually-correct practice opportunities during paper practice), we included all students present on day two ($N = 90$) or day four ($N = 93$).

For each tutor and paper practice unit, we compute, per student, the number of attempted steps, eventually-correct steps based on last step attempts, and the percentage of correct first attempts. To capture hypothesized interface differences between ITS and paper, we aggregated the average step frequency of gaming-the-system behaviors in the tutor (i.e., dividing the number of steps where gaming-the-system was detected by the total number of steps). We employed a detector validated on data from Cognitive Tutor Algebra. The detector takes into account rapid attempts and hint use in the tutor, among others, and is further described in [28]. Furthermore, we aggregated the average number of omission errors on paper per unit. We further analyzed students' knowledge growth on each KC during practice using the AFM model, a logistic regression model commonly used to analyze log data from tutoring systems [29]. This model provides parameter estimates for the initial ease and the learning rate per KC by estimating the log odds of correct steps (i.e., on first attempts) as a function of the opportunity count for the KC being used at the step. Using the tutoring system's KC model, we fit and compare AFM for each of the three problem sets (i.e., 7.05, 7.06, and 7.07). We fit the model separately for paper and tutor logs to estimate learning rates and intercepts averaged by unit and practice condition. Omission errors on paper were treated as learning opportunities.

For H3 (associations between eventually-correct steps and gains), we use the same sample as for H1. We test our hypothesis that eventually-correct steps relate to learning gain via correlation tests. We conduct automated feature selection via stepwise regression to further explore the associations between learning process measures aggregated on the student-level and gain. *AIC*-based backward search was employed to select factors of our experiment (i.e., Condition, Unit, and Test Format), omission errors, and gaming-the-system. The condition factor controls for the absence of gaming-the-system on paper and step omissions in the tutor (coded as 0).

3 Results

3.1 Learning gain differences between tutor and paper (H1)

We start by investigating the test scores across measurement points, conditions, and units (Table 1). The learning gains students exhibited in both the tutor and the paper conditions were statistically significant (see Table 2): tutor, $t(281) = 2.76, p < .001$, paper, $t(287) = 7.94, p < .001$. The gains were also statistically significant for each of the units separately: 7.05, $t(189) = 4.22, p < .001$, 7.06, $t(189) = 4.84, p < .001$, and 7.07, $t(189) = 9.75, p < .001$.

Table 1. Test scores across measurement points and units broken out by condition.

Unit	Paper/Pre	Paper/Post	Tutor/Pre	Tutor/Post
7.05	24.06%	34.52%	25.46%	33.91%
7.06	12.87%	18.9%	11.62%	24.96%
7.07	3.00%	29.53%	3.87%	21.13%

Table 2. Anova table of learning gains across Conditions (paper vs. tutor practice) x Unit (7.05 vs. 7.06 vs. 7.07) x Test Format (interface alignment vs. transfer).

Effect	F	p	η2
Intercept	174.14	<.001	0.18
Condition	0.40	0.526	0.00
Unit	15.61	<.001	0.04
Test Format (Alignment/Transfer)	8.71	0.003	0.02
Condition : Unit	5.28	0.006	0.01
Condition : Test Format	0.61	0.435	0.00
Unit : Test Format	1.65	0.194	0.01
Condition : Unit : Test Format	0.41	0.665	0.00

To test whether problem-solving practice with an ITS yields greater learning gains than problem-solving practice on paper (H1), we looked at the main effect of Condition on learning gain (see Table 2). Contrary to H1, this main effect was not significant, *F*(1, 222) = 0.40, *p* = .526. Overall, learning gains of students practicing with the tutor (*M* = 0.13, *SD* = 0.30) were similar to those of students practicing on paper (*M* = 0.15, *SD* = 0.31). However, there was a significant interaction between Condition and Unit, *F*(2, 222) = 5.28, *p* = .006. As shown in Figure 3, this Condition by Unit interaction reflects gains favoring the tutor in 7.06 but favoring paper in 7.07. Our follow-up *t*-tests of the simple main effects indicate no statistically significant condition differences in learning gains for 7.06 or 7.05 (*p's* > .069) but a significant difference in favor of paper for 7.07 (*p* = .039). Nevertheless, given the reliable interaction, it is worth trying to explain, as we do in the discussion, why the tutor is relatively better for 7.06 while paper is relatively better for 7.07.

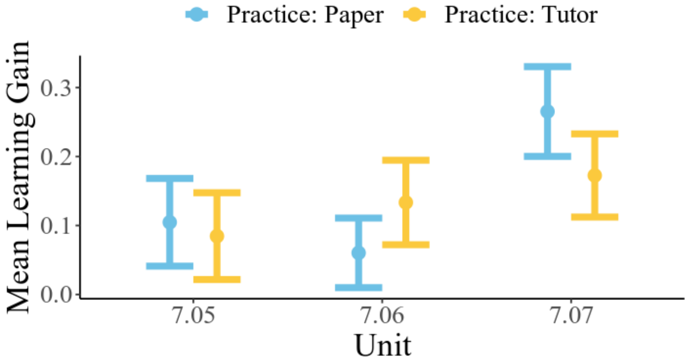


Fig. 3. Relatively greater learning from tutor practice in 7.06 versus from paper practice in 7.07 is illustrated by average learning gains (post minus pre) across Condition and Unit, including 95% confidence intervals.

We note the two other statistically reliable effects in the Anova in Table 2. There was a significant main effect of Test Format, indicating that learning gains were higher when the test format (computer vs. paper) *matched* the practice environment (computer vs. paper) versus when there was no such match and students had to transfer knowledge across formats, $t(567.96) = -3.28$, $p = .001$. Learning gains were nearly twice as high when the test matched the practice format ($M = 0.18$, $SD = 0.30$) than when there was no format match ($M = 0.10$, $SD = 0.30$). All other interactions were not statistically reliable. Thus, there is no evidence, in any Unit, that either tutor or paper produces better transfer to the alternate format than the others.

3.2 Fewer eventually-correct steps on paper (H2)

We now test the hypothesis that students engaging in problem-solving practice with paper have fewer eventually-correct KC opportunities in the same amount of time than students practicing with the tutor (H2). A process-level measure breakdown during practice is in Table 3.

Table 3. Per-condition process measures averaged across students and units.

Condition	N Steps Attempted	% Corrects on First Attempts	Eventually Correct Steps	% Eventually Correct Steps
Paper	91.55	48.86	44.75	48.89
Tutor	98.56	71.20	97.79	99.22

Consistent with H2, students practicing in the tutor experienced 2.19 (97.79/44.75) times more eventually-correct steps than students practicing on paper, $t(119.76) = 7.57$, $p < .001$. This ratio ranged from 2.35 for 7.06 to 1.48 for 7.05. Further, students practicing with the tutor reached a correct answer on almost all steps they attempted (99.2% = 97.79/98.56), whereas students practicing with paper reached a correct answer on less than half of the steps they attempted (48.9% = 44.75/91.55). Steps in the tutor on which students did not reach correctness can be explained by students starting but not finishing their last problem. Notably, students working with the tutoring system had a significantly higher accuracy on first attempts, $t(171.62) = 7.62$, $p < .001$.

To better understand potential interface differences during practice, we also report unit-level differences in omission errors on paper, gaming-the-system in the tutors), and learning rates across practice conditions (see Table 4). We found significantly more frequent tutor gaming-the-system behavior per step in 7.07 than in 7.05, $t(68.55) = 8.22$, $p < .001$, and in 7.06, $t(96.54) = 6.60$, $p < .001$. Furthermore, working on paper, students omitted steps significantly more frequently in 7.06 compared to 7.05, $t(139.57) = 4.67$, $p < .001$, and 7.07, $t(156.92) = 3.76$, $p < .001$. We also found unit-level differences in the analysis of learning rates across practice conditions (see Table 4). In line with the test results, students exhibited the largest advantage from tutoring (compared to paper) on the 7.06 units according to AFM learning rates. Notably, learning rates were low in both conditions in 7.07. Meanwhile, there was comparable learning between the conditions in 7.05.

To better understand where the tutor may be most effective, we focus on unit 7.06, which had the greatest difference in learning gains between tutor and paper (in favor of the tutor). We compared both KC-level learning rates and KC-level pre-post learning gains across conditions (see Table 5). We found, generally, greater KC-level learning rates ($\Delta\beta$) and KC-level test gains (ΔLG) due to tutor practice compared to paper practice. Students showed the largest test improvements due to tutor practice (i.e., ΔLG in Table 5) on the KCs representing point plotting (15% more learning gain), calculating y coordinates given x coordinates (ΔLG = 13%), and line drawing (ΔLG = 12%). For the line-drawing and point-plotting KCs, these test advantages corresponded to within-practice learning rate advantages (denoted $\Delta\beta$) compared to paper ($\Delta\beta = 0.94$ and $\Delta\beta = 0.70$, respectively). The one substantial exception is the KC for calculating y coordinates given x coordinates, where the tutor produced larger learning gains as measured on the tests, but no learning was observed in the learning rate. This KC definition (i.e., the mapping of this KC to problem steps) may be flawed, or the associated items on the tests require other KCs than those on paper.

Table 4. Mean KC-level AFM tutor learning rates and intercepts, tutor gaming-the-system step frequency, and paper step omissions, broken out by units.

Unit	Learning Rate		Learning Intercept		Gaming-the-system	Omissions
	Tutor	Paper	Tutor	Paper	Tutor	Paper
7.05	0.44	0.49	1.78	-0.08	0.00	21.3
7.06	0.19	-0.20	0.51	0.14	0.08	51.1
7.07	0.04	0.06	0.90	-2.16	0.27	26.72

Table 5. KC-level overview of learning slopes and condition-specific learning gains for each condition (including slope differences $\Delta\beta$ and learning gain differences ΔLG) for 7.06.

Knowledge Component (7.06)	Learning Rate			Pre-Post Learning Gain		
	Tutor	Paper	$\Delta\beta$	Tutor	Paper	ΔLG
Plot Points given Coordinates	0.17	-0.53	0.70	16.33%	1.06%	15.26%
Calculate Y given X	-0.07	-0.05	-0.01	20.07%	7.45%	12.62%
Draw a Line	0.55	-0.39	0.94	19.39%	7.45%	11.94%
Name Quantities	0.09	-0.64	0.73	5.26%	1.52%	3.74%
Find Equation	0.51	0.07	0.44	15.31%	12.77%	2.54%
Name Units	0.12	0.20	-0.08	6.68%	5.89%	0.79%
Determine X off Point	0.00	-0.04	0.04	14.29%	15.96%	-1.67%

3.3 Association between eventually-correct steps and and learning gains (H3)

There was no general correlation between the number of eventually-correct steps and learning across the factors of our experiment, $r(568) = 0.00$, $CI = [-0.08, 0.09]$, $p = .902$. Nor was there such a correlation when separating tutor practice ($r(280) = 0.05$, $CI = [-0.06, 0.17]$, $p = .327$) and paper practice ($r(286) = -0.08$, $CI = [-0.19, 0.04]$, $p = .181$). Focusing on the results of the computer test (rather than the combined computer and paper test results), there was a significant positive correlation between the number of eventually-correct steps in practicing with the tutor and gains, *r*(139) = 0.20, *CI* = [0.04, 0.36], *p* = .001. We further break out our correlation analysis by unit (see Table 6) and find that more eventually-correct steps in the tutor were associated with larger gains in 7.05, *r*(90) = 0.19, *CI* = [-0.01, 0.38], *p* = .067 and, when evaluated on the tutor test, also in 7.06, *r*(47) = 0.42, *CI* = [0.16, 0.63], *p* = .003. Conversely, more eventually-correct steps during paper practice related to larger gains in 7.07 when evaluated on paper tests, *r*(47) = 0.29, *CI* = [0.01, 0.53], *p* = .043.

Table 6. Correlation tests between the number of eventually-correct steps and learning gain broken out by unit and transfer condition, including significance levels.

Unit	Tutor Practice			Paper Practice		
	Tutor Test Gain	Paper Test Gain	Overall Gain	Tutor Test Gain	Paper Test Gain	Overall Gain
7.05	0.34*	0.09	0.19 .	0.21	0.14	0.15
7.06	0.42**	-0.23	0.04	-0.28 .	0.02	-0.10
7.07	-0.10	0.20	0.04	-0.18	0.29*	0.05

. *p* < 0.1, * *p* < .05, ** *p* < .01, *** *p* < .001

To better understand what process data features predict pre-to-post learning gains, we performed an automatic feature selection process described in Section 2.5. Of the non-process features (those in Table 2), the Unit and Test Format features were selected as significant control variables. Adjusting for these controls, the selected models indicated that gaming-the-system (in the tutor practice) and omission errors (in the paper practice) were significantly negatively associated with learning gain, β = -0.14, *p* = .006 and β = -0.13, *p* = .012, respectively. Students experienced approximately 0.13 *SD* lower learning gains per additional standard deviation in omission errors or per-step gaming-the-system frequency.

4 Discussion

We present a novel investigation of the relative benefits and limitations of deliberate practice with a computer-based intelligent tutoring system versus on paper. We measured learning gains with pre- and post-tests on both paper and computer. Our experimental method involved identical practice and test problems on paper and tutor with process-level analyses of practice opportunities across interfaces. To our knowledge, this comparison is a first of a kind.

4.1 Why was tutoring not generally better than paper?

The hypothesis that a tutoring system, with its step-level feedback and other guidance, would yield better learning than paper practice (H1) was not confirmed. We saw an increase in performance in all practice conditions, whether done on paper or in the tutor, but not an overall significant effect of condition. We did, interestingly, find a significant interaction between tutor unit and condition such that students learned more from the tutor in the 7.06 unit on graphing linear equations, (significantly) more from paper in the 7.07 unit on qualitative graph interpretation, the same in the 7.05 unit on quantitative graph interpretation. Why did we not find a general tutor advantage? We discuss three possible explanations.

Before doing so, it is worth emphasizing that had our study been done only on unit 7.06, we would see results in both outcome and process analyses consistent with the tutoring benefit posed by H1. In this unit, the tutor condition had marginally significant greater learning gains as well as a higher learning rate and more eventually-correct steps during training. Moreover, students on paper showed the largest omission error frequency on 7.06. In contrast, students working in the tutor appeared to receive enough tutor assistance to grapple with challenging steps, reach correctness on problems, and learn from that assistance.

A first possible explanation for the lack of a general tutor advantage is that the selected tutor units may have design flaws that may hide the tutor benefits of immediate feedback, adaptive hints, and problem selection. These units had not previously benefited from data-driven redesign. It is possible that out-of-the-box tutor design with little data-driven redesign may not give students straightforward practice advantages, despite immediate feedback and assistance [35].

A second possible explanation is that students did not have adequate time to get acquainted with the tutoring software. According to our observation notes, the relatively short time frame for students and teachers to get acquainted with the software in this study (which sparked ample teacher-help behavior during tutor practice) might have diminished learning gain returns on intelligent tutoring. This explanation aligns with prior studies highlighting the importance of integrating a tutor's pedagogy into classroom practices [30].

A third possible explanation for the lack of a general tutor advantage is that paper practice may have distinct benefits for specific mathematical or learning contexts, particularly ones that leverage free-form writing and drawing that paper affords [31]. Or conversely, the tutor may have specific disaffordances. We have not identified why qualitative graph interpretation (unit 7.07) might benefit more from paper affordances than graph production (unit 7.06). Indeed, it would seem that graph production would benefit from free-form entry (e.g., in the plotting of points and drawing lines) more so than graph interpretation, but we found just the opposite -- paper benefited graph interpretation (7.07) more so than graph production (7.06).

Perhaps the converse is true: a tutor may have disaffordances in some contexts. Our process analysis provides some support for this explanation. We found that students exhibited significantly more gaming-the-system behavior during tutor practice in the 7.07 unit than in other units. Such behavior, which is known to be associated with worse learning outcomes [32], might have been invoked by menu-based interface elements in 7.07 not found in the other two units. Such menus have been observed to evoke more gaming-the-system [32] as students are tempted to systematically try multiple options and use the tutor's immediate feedback to get a step correct rather than reasoning towards correctness. Consistent with this explanation, students showed the lowest learning rates in the tutor on 7.07. To be sure, the suggestion here is *not* that menu-based input or multiple-choice questions are bad. Afterall, plenty of tutors with menu-based interactions (e.g., where students explain their steps) have demonstrated enhanced student learning (e.g., [33, 34]). Further, students in the paper condition were also given choices to select from. They could have also taken the easy way out by guessing. The availability of feedback in the tutor is thus a crucial part of this explanation. When immediate feedback allows many fast inputs, students are tempted to turn off their thinking. Yet, as indicated in the studies referenced above, menus with immediate feedback are not generally enough to harm learning. We suspect a third factor is needed: the material is particularly hard. This factor was present in 7.07, where the mean pre-test score was only 3.4% and learning rates were low. In sum, menus *and* immediate feedback *and* insufficient prior preparation may evoke a learning disaffordance from tutoring.

4.2 Does paper practice produce fewer eventually-correct steps?

We found strong evidence for hypothesis H2, in particular, that students are much less likely to produce eventually-correct steps on paper than in the tutor. Indeed, on paper, students achieved less than half the number of eventually-correct steps, in the same amount of time, as students practicing in the tutor. This finding is an important confirmation that feedback and adaptive instruction impact student performance. Tutored students achieved more eventually-correct steps not because they attempted more steps (they did not) but because on the steps they could not solve independently, the tutor aided them through feedback, instructional hints, or a worked example. In contrast, on paper, students had no recourse when they reached steps they could not solve on their own but to omit the step or make a guess. However, students could only partially translate these practice advantages into learning gain advantages. In the next section, we discuss potential reasons why.

4.3 Conditional association between eventually-correct steps and learning

Similar to hypothesis H1, we found that hypothesis H3 was not generally supported but was supported in particular and sensible contexts. Following prior theory [12, 36], H3 predicts that the number of eventually-correct steps a student experiences is associated with their learning gains. We did not find this relationship generally across all tutors, units, and test formats. We did, however, find support for a more specific version of this hypothesis that suggests that in a tutor without a gaming-the-system disaffordance (i.e., not unit 7.07), the number of eventually-correct steps a student experiences is associated with their learning gains in a near-transfer context. This hypothesis is supported by the fact that this correlation was significant in the tutor conditions for units 7.05 and 7.06 when the gains were measured on the matched or near-transfer test (i.e., the computer-based test). Interestingly, the other significant correlation was found for the paper condition in 7.07 on the matched test. In other words, it was found in the Unit where paper affordances for learning were better than the tutor affordances, but again, only on the matched or near-transfer test

This pattern of findings supports the idea that comparing the strength of associations between learning process and outcome data can help predict and explain when an instructional approach, whether paper-based or ITS-based, is working better. In particular, we propose a general conjecture that for a given instructional approach, the number of eventually-correct steps will be predictive of learning gain if and only if that approach is effective in supporting learning.

To support this conjecture, we delineate potential characteristics of suboptimal instruction in our sample. Next to gaming-the-system the system behavior which was previously found to be associated with worse learning outcomes [32], we found omission errors on paper were negatively related to learning gains. What do these errors represent? Perhaps students with higher prior knowledge may be able to generate their own feedback while practicing on paper, but students with lower prior knowledge may not, disengage on paper, and learn less. Furthermore, students can (and do) skip steps on paper, presumably the steps they do not know how to do, which might be the ones they most need practice on. Indeed we found that students with more omission errors on paper learn less from paper. If these errors indicate where students require assistance most, redesigning tutoring systems based on diagnosed needs at paper assessment could guide deliberate practice. For example, skipped steps may guide focused practice practice on difficult steps or motivational support.

Why do we not see correlations on mismatched tests assessing cross-format transfer? Perhaps students who pursue more successful practice (i.e., eventually correct steps) opportunities reach better near-transfer learning because of the benefits of deliberate practice. But, to the extent that they hurry to complete repeated practice, they may not be doing sufficient reflection to produce more general learning needed for farther transfer. Further research is needed to see if such explanations work out.

4.4 Why does this research matter for TEL?

The current study compares two ecological conditions of paper and tutor practice to do research relevant to classroom practice. Although students could reach correctness on more than twice as many steps during tutor practice, our data indicate that the relation between eventually-correct steps and learning gains was conditional on the tutor design. As such, our study gathers novel insights into how and when tutoring systems, a popular form of TEL, might be most effective. These insights emerged through a methodology contribution for comparing process-level learning differences between pen-and-paper practice and tutor practice via generating log data for paper responses on matched content. A lesson learned from this study is that comparing TEL against paper on a process level might be more useful than thought (e.g., comparing how interfaces invoke gaming-the-system), suggesting that data-driven tutor redesign is perhaps more urgent than previously thought.

We see three future directions for refining and applying our contribution of entering paper problem-solving responses into analogous ITS interfaces. First, future work may improve our operationalization of learning opportunities, particularly on paper, via additional data streams (e.g., gaze data, think-aloud data), our counting the number of completed steps achieved through gaming-the-system. Second, future work may reduce the cost of paper entry, for example, through computer vision automation. Two coders worked 6-7 weeks full-time to generate log data for this study. Third, future work may investigate how the affordances of TEL and paper practice may be merged. Free-form input on paper could be leveraged for TEL environments to augment the range of inputs and problem-solving strategies. Conversely, mixed reality may bring personalized instruction and reactive feedback to paper practice.

References

1. Kerres, M., Buchner, J.: Education after the pandemic: What we have (not) learned about learning. Education Sciences. 12(5), 315 (2022)
2. VanLehn, K.: The behavior of tutoring systems. International journal of artificial intelligence in education. 16, 227–265 (2006)
3. Ritter, S., Fancsali, S.: MATHia X: The Next Generation Cognitive Tutor. In: EDM. pp. 624–625. ERIC (2016)
4. Fang, Y., Ren, Z., Hu, X., Graesser, A.C.: A meta-analysis of the effectiveness of ALEKS on learning. Educational Psychology. 39, 1278–1292 (2019)
5. Heffernan, N.T., Heffernan, C.L.: The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. International Journal of Artificial Intelligence in Education. 24, 470–497 (2014)
6. Silva, T.B.: The effects of the i-Ready computer assisted instruction program on the reading and fluency achievement of first graders (2016)
7. du Boulay, B.: Recent meta-reviews and meta–analyses of AIED systems. International Journal of Artificial Intelligence in Education. 26, 536–537 (2016)
8. Kulik, J.A., Fletcher, J.D.: Effectiveness of intelligent tutoring systems: a meta-analytic review. Review of educational research. 86, 42–78 (2016)
9. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational psychologist. 46, 197–221 (2011)
10. Aleven, V., McLaughlin, E.A., Glenn, R.A., Koedinger, K.R.: Instruction based on adaptive learning technologies. Handbook of research on learning and instruction. 2, 522–560 (2016)
11. Corbett, A.T., Anderson, J.R.: Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 245–252 (2001)
12. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: Lessons learned. The journal of the learning sciences. 4, 167–207 (1995)
13. Cabalo, J.V., Ma, B., Jaciw, A.: Comparative Effectiveness of Carnegie Learning's" Cognitive Tutor Bridge to Algebra" Curriculum: A Report of a Randomized Experiment in the Maui School District. Research Report. Empirical Education Inc. (2007)
14. Ritter, S., Kulikowich, J., Lei, P., McGuire, C.L., Morgan, P.: What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. In: Hirashima, T., Hoppe, U., Young, U.U. (eds.) Supporting learning flow through integrative technologies, vol. 162, pp. 13-20. Amsterdam, IOS Press (2007)
15. Steenbergen-Hu, S., Cooper, H.: A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students’ mathematical learning. Journal of educational psychology. 105(4), 970 (2013)
16. Koedinger, K.R., Anderson, J.R., Hadley, W.H., Mark, M.A.: Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education. 8(1), 30-43 (1997)
17. Mendicino, M., Razzaq, L., Heffernan, N.T.: A comparison of traditional homework to computer-supported homework. Journal of Research on Technology in Education. 41(3), 331-359 (2009)
18. Magalhães, P., Ferreira, D., Cunha, J., Rosário, P.: Online vs traditional homework: A systematic review on the benefits to students' performance. Computers & Education. 152, 103869 (2020)
19. Koedinger, K.R., Aleven, V.: Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review. 19, 239–264 (2007)

20. Baker, R.S., Corbett, A.T., Koedinger, K.R.: Detecting student misuse of intelligent tutoring systems. In: Intelligent Tutoring Systems: 7th International Conference, ITS 2004, Maceió, Alagoas, Brazil, August 30-September 3, 2004. Proceedings 7. pp. 531–540. Springer (2004)
21. Anderson, J.R., Conrad, F.G., Corbett, A.T.: Skill acquisition and the LISP tutor. Cognitive Science. 13, 467–505 (1989)
22. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive science. 36, 757–798 (2012)
23. Koedinger, K.R., Baker, R.Sj., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the EDM community: The PSLC DataShop. Handbook of educational data mining. 43, 43–56 (2010)
24. Aleven, V., McLaren, B.M., Sewall, J.: Scaling up programming by demonstration for intelligent tutoring systems development: An open-access web site for middle school mathematics learning. IEEE transactions on learning technologies. 2, 64–78 (2009)
25. Sibbald, B., Roberts, C.: Understanding controlled trials crossover trials. Bmj. 316, 1719–1720 (1998)
26. Koedinger, K.R., Stamper, J.C., Leber, B., Skogsholm, A.: LearnLab's DataShop: A Data Repository and Analytics Tool Set for Cognitive Science. Top. Cogn. Sci. 5, 668–669 (2013)
27. Koedinger, K.R., Anderson, J.R.: Illustrating principled design: The early evolution of a cognitive tutor for algebra symbolization. Interactive Learning Environments. 5, 161–179 (1998)
28. Paquette, L., de Carvahlo, A., Baker, R., Ocumpaugh, J.: Reengineering the feature distillation process: A case study in detection of gaming the system. In: Educational data mining 2014 (2014)
29. Cen, H., Koedinger, K., Junker, B.: Learning factors analysis–a general method for cognitive model evaluation and improvement. In: Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006. Proceedings 8. pp. 164–175. Springer (2006)
30. Pane, J.F., McCaffrey, D.F., Slaughter, M.E., Steele, J. L., Ikemoto, G.S.: An experiment to evaluate the efficacy of cognitive tutor geometry. Journal of Research on Educational Effectiveness. 3(3), 254-281 (2010)
31. Anthony, L., Yang, J., Koedinger, K.R.: Towards the application of a handwriting interface for mathematics learning. In: 2006 IEEE International Conference on Multimedia and Expo. pp. 2077-2080 (2006)
32. Huang, Y., Dang, S., Richey, J.E., Asher, M., Lobczowski, N.G., Chine, D., McLaughlin, E.A., Harackiewicz, J.M., Aleven, V., Koedinger, K.: Item Response Theory-Based Gaming Detection. International Educational Data Mining Society (2022)
33. Aleven, V.A., Koedinger, K.R.: An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. Cognitive science. 26(2), 147-179 (2002)
34. Atkinson, R.K., Renkl, A., Merrill, M.M.: Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. Journal of educational psychology. 95(4), 774 (2003)
35. Huang Y, Lobczowski NG, Richey JE, McLaughlin EA, Asher MW, Harackiewicz JM, Aleven V, Koedinger KR (2021) A general multi-method approach to data-driven redesign of tutoring systems. In: LAK21: 11th International Learning Analytics and Knowledge Conference. pp 161–172
36. Koedinger KR, McLaughlin EA, Jia JZ, Bier NL (2016) Is the doer effect a causal relationship? How can we tell and why it's important. In: Proceedings of the sixth international conference on learning analytics & knowledge. pp 388–397