How and When: The Impact of Metacognitive Knowledge Instruction and Motivation on Transfer Across Intelligent Tutoring Systems

Article *in* International Journal of Artificial Intelligence in Education · September 2023

DOI: 10.1007/s40593-023-00371-0

CITATIONS
12

READS
558

3 authors:

Mark Abdelshiheed
University of Colorado Boulder
34 PUBLICATIONS **225** CITATIONS
SEE PROFILE

Tiffany Michelle Barnes
North Carolina State University
399 PUBLICATIONS **5,670** CITATIONS
SEE PROFILE

Min Chi
North Carolina State University
131 PUBLICATIONS **1,933** CITATIONS
SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

How and When: The Impact of Metacognitive Knowledge Instruction and Motivation on Transfer across Intelligent Tutoring Systems

Mark Abdelshiheed, Tiffany Barnes and Min Chi

Received: date / Accepted: date

Abstract Two metacognitive knowledge types in deductive domains are procedural and conditional. This work presents a preliminary study on the impact of metacognitive knowledge and motivation on transfer across two Intelligent Tutoring Systems (ITSs), then two experiments on metacognitive knowledge instruction. Throughout this work, we trained students on a logic ITS that supports a *default* forward-chaining and an alternative backward-chaining (BC) strategy, then a probability ITS that only supports BC. Students were grouped into those with conditional knowledge who know *how* and *when* to use each strategy ($StrBoth$), those with procedural knowledge who know *how only* ($StrHow$), and the rote students who persist in the default strategy and know neither how nor when ($Rote$). The online traces' initial accuracy was used to further split students into high- and low-motivation groups. The preliminary study showed that only high-motivation $StrBoth$ students transferred their metacognitive knowledge across the two ITSs. The two experiments provided metacognitive knowledge instruction for $StrHow$ and $Rote$ students to

M. Abdelshiheed (Corresponding) · T. Barnes · M. Chi
Computer Science, North Carolina State University, Raleigh, 27695, NC, USA
E-mail: {mnabdels, tmbarnes, mchi}@ncsu.edu

catch up with their *StrBoth* peers. In *Exp.* 1, we utilized prompted nudges to teach when to use BC, and in *Exp.* 2, we combined nudges with worked examples to teach how and when to use BC. Based on our findings, we propose a ***M**etacognitive knowledge, initial **M**otivation, and instructional **I**nterventions (MMI)* framework for transfer across ITSs. The framework suggests that the key factors for facilitating transfer are the motivation for *StrBoth* students, nudges for their *StrHow* peers, and the combination of worked examples and nudges for *Rote* students.

Keywords Metacognitive Knowledge · Motivation · Metacognitive Interventions · Transfer · Intelligent Tutoring Systems

Introduction

One essential priority of education is preparing students for future learning in that they would transfer acquired skills and problem-solving strategies across different domains (Bransford & Schwartz, 1999). While achieving transfer can be challenging (Detterman & Sternberg, 1993), substantial research has shown that it can be facilitated by possessing metacognitive knowledge (Zepeda et al., 2015; Chi & VanLehn, 2010) or motivation (Belenky & Nokes-Malach, 2013; Nokes-Malach & Belenky, 2011). Metacognitive knowledge refers to what individuals know about themselves as cognitive processors and about different approaches used for problem-solving and learning a particular task (Schraw & Dennison, 1994). Our perspective of approaching motivation is defining it as the desire to learn without concern for ulterior motives.

Considerable work has highlighted the impact of the interaction of metacognitive knowledge and motivation on self-regulated learning (Azevedo et al., 2017; Zimmerman, 2011). In this work, we focus on two metacognitive knowledge types related to problem-solving strategies: procedural and conditional

(Krathwohl, 2002). Procedural knowledge is needed to understand how to use each strategy, while conditional knowledge relates to how and when to use each strategy. To infer the students' motivation, we consider their desire to learn by leveraging the initial accuracy of their trace logs without concern for their ulterior motives. This work operationalizes metacognitive knowledge and motivation, which can be considered a step toward objective methods for concepts that rely on theories and subjective self-report measures in literature.

Substantial work has demonstrated the significance of teaching how and when to use each strategy on subject-matter knowledge transfer (de Boer et al., 2018; Schraw & Gutierrez, 2015). We focus on two interventions for teaching how and when to use a strategy: teaching by example (Likourezos & Kalyuga, 2017; Glogger-Frey et al., 2015) and prompted nudges (Richey et al., 2015; Belenky & Nokes-Malach, 2009). While prior work has used interventions that included hints, feedback, nudges, and worked examples, such interventions were provided in classroom settings by human experts and self-guided packets, or their effectiveness on transfer was not evaluated. To the best of our knowledge, prior work has yet to investigate the impact of instructional interventions to teach metacognitive knowledge on transfer across intelligent tutoring systems.

Intelligent Tutoring Systems (ITSs) are interactive e-learning environments that provide instruction, scaffolded practice, and immediate help and feedback to students without requiring intervention from a human teacher (Vanlehn, 2006). Throughout this work, we leveraged two ITSs: logic and probability. We trained students first on a logic ITS that supports a default forward-chaining and an alternative backward-chaining (BC) strategy, then on a probability ITS six weeks later that only supports BC. Students were grouped into those with conditional knowledge who know *how* and *when* to use each strategy ($StrBoth$), those with procedural knowledge who know *how only* ($StrHow$),

and the rote students who stick to the default strategy and know neither how nor when (*Rote*). Students were further split into high- and low-motivation groups based on the initial accuracy of their trace logs.

A preliminary study on 495 undergraduates across three semesters showed that only high-motivation *StrBoth* students transferred their metacognitive knowledge across the two ITSs. We conducted two consecutive experiments, each in a semester, to provide metacognitive knowledge instruction for *StrHow* and *Rote* students to catch up with their *StrBoth* peers. In *Exp.* 1, we leveraged prompted nudges to teach when to use BC, and in *Exp.* 2, we combined nudges with worked examples to teach how and when to use BC. Based on our findings, we propose a ***M**etacognitive knowledge, initial **M**otivation, and instructional **I**nterventions (MMI)* framework for transfer across ITSs. The framework suggests that the key factors for facilitating transfer are the motivation for *StrBoth* students, nudges for their *StrHow* peers, and the combination of worked examples and nudges for *Rote* students.

The main **contributions** of this work are:

1. Operationalizing metacognitive knowledge and motivation to promote objective rather than subjective self-report measures in educational domains.
2. Demonstrating the impact of combining metacognitive knowledge and motivation on quantifying, predicting, and capturing transfer across ITSs.
3. Showing the significance of providing metacognitive instructional interventions for students with low metacognitive knowledge on transfer across ITSs.
4. Proposing a transfer framework across ITSs based on metacognitive knowledge, initial motivation, and instructional interventions.

The remaining sections are divided into the background and related work, our research questions, the methods, the preliminary study, two sections for

the two experiments, a post-hoc analysis by combining the two experiments' results, the general discussion, the proposed ***MMI*** framework, and the limitations and broader impacts of this work.

Background & Related Work

The term "knowledge transfer" has been considerably used in literature to refer to broad concepts that can be summarized into two views: an algorithmic model-based view referred to as *Transfer Learning* (Weiss et al., 2016) and a human-based view known as *Preparation for Future Learning* (Bransford & Schwartz, 1999). Transfer learning is a machine learning technique that adapts a model trained on one task to perform well on a related task, which is **not** in the scope of our work. We adopt the **second** view of knowledge transfer as it addresses preparing individuals for future learning.

Bransford and Schwartz (1999) proposed a view of transfer as preparation for future learning, which assumes that students continue to learn by transferring acquired skills and strategies across different domains. Much research has shown that transfer can be accelerated by obtaining metacognitive knowledge (Zepeda et al., 2015; Chi & VanLehn, 2010) or influenced by the students' motivation (Belenky & Nokes-Malach, 2013; Nokes-Malach & Belenky, 2011).

Metacognitive Knowledge and Strategy Instruction

Metacognition indicates cognition about cognition and the ability to conceive, monitor and regulate knowledge (Livingston, 2003; Roberts & Erdos, 1993). Two types of metacognitive knowledge are procedural and conditional (Krathwohl, 2002). Procedural knowledge is related to the understanding of *how* to use different problem-solving strategies and learning approaches without conscious attention or reasoning about their rationale (Willingham et al.,

1989; Georgeff & Lansky, 1986). Conditional knowledge is a higher form of knowledge, as it requires understanding *how* and *when* to use each strategy (Schraw, 1998; Schraw & Moshman, 1995).

Many studies have shown that mastering how and when to use each strategy yields subject-matter knowledge transfer (Zepeda et al., 2015; Chi & VanLehn, 2010; Wagster et al., 2007). Chi and VanLehn (2010) found that students who mastered the principle-emphasis instruction on how and when to apply each principle transferred their problem-solving strategy from a probability to a physics ITS. Wagster et al. (2007) showed that students who knew how to apply different strategies to construct concept maps on a biology ITS outperformed their peers on a transfer task. Zepeda et al. (2015) demonstrated that students who knew how to plan outperformed their peers on a novel self-guided control of variables learning task.

Prior research has shown the significance of metacognitive strategy instruction in regulating strategy use (de Boer et al., 2018; Schraw & Gutierrez, 2015). Schraw and Gutierrez (2015) argued that metacognitive strategy instruction should provide knowledge about *how*, *when*, and *why* to use a given strategy. They suggested that such instruction should distinguish the merit of each strategy and compare strategies according to their feasibility and familiarity from the learner's perspective. de Boer et al. (2018) investigated the long-term effects of metacognitive strategy instruction on academic performance. They found that students who were given interventions that included when, why, how and which strategy to use outperformed their peers on a post-test task and a far follow-up test. de Boer et al. (2018) argued that only learning *how* to use each strategy in multi-strategy domains is insufficient. Rather, it is equally important to learn *when* to use each.

Considerable work has explored many metacognitive interventions for strategy instruction and highlighted their tradeoffs. We focus on two interventions:

teaching a strategy by example (Likourezos & Kalyuga, 2017; Glogger-Frey et al., 2015) and prompted nudges (Richey et al., 2015; Belenky & Nokes-Malach, 2009) Glogger-Frey et al. (2015) found that students receiving worked examples of journal extracts reviews outperformed their peers, who had to come up with the reviews, on post-test performance. However, Likourezos and Kalyuga (2017) reported no detectable difference between students who received fully-guided worked examples, partially-guided ones and unguided assistance on post-test geometry tasks. Belenky and Nokes-Malach (2009) showed that students who were prompted with metacognitive nudges outperformed their peers on a permutation transfer task. Conversely, Richey et al. (2015) found no detectable difference between students who were instructed to study the worked examples and their peers, who received the same examples with tutoring nudges, on near, intermediate and far transfer electric circuit tasks.

In brief, prior work has shown that knowing how and when to use each strategy facilitates metacognitive knowledge transfer. Hence, many interventions have been investigated for strategy instructions, such as worked examples and prompted nudges. However, as far as we know, there is no agreement on the most effective combination of the two interventions, and no work has directly used these interventions to teach metacognitive knowledge across ITSs.

Motivation as Desire to Learn

Eccles (1983) defined motivation as a process that combines the individual's perception of three factors: expectations for success, subjective task value, and intrinsic interest. Touré-Tillery and Fishbach (2014) stated that motivation is the psychological force that enables action. The multiple definitions and perspectives to approach motivation prompted various motivation theories,

such as achievement goal theory (Elliot, 2005; Dweck, 1986) and expected value theory (Eccles & Wigfield, 2020; Eccles, 1983).

The motivation theories relied on self-report measures, such as surveys and questionnaires, to determine the ulterior motives of learners, such as interest, mastery, or performance. For example, the achievement goal theory proposes a 2 X 2 framework to reflect the learner's goal orientation: {mastery, performance} X {approach, avoidance} (Elliot, 2005; Dweck, 1986). The mastery aims to understand and master the task, while the performance reflects the desire to outperform others. The approach and avoidance capture approaching success and avoiding failure, respectively. Hence, mastery-approach students are those who stated in the questionnaire that they want to master the task to achieve success. However, one issue in self-report measures is that they are subjective and could be inaccurate (Fulmer & Frijters, 2009); for example, two students reporting that they want to learn as much as possible could have different intentions or interpretations.

In recent years, digital technologies such as ITSs made it possible to measure motivation objectively using students' online trace logs and found contrasting results (Fancsali et al., 2014; Otieno et al., 2013; M. Zhou & Winne, 2012). M. Zhou and Winne (2012) compared a self-report survey and online traces in examining achievement goals while studying a multimedia-formatted article. Traces were collected when students applied tags to text selections or clicked hyperlinks in the article. The tag labels were similar to the self-report items in the survey, except that they were associated with particular text or links. The results showed that the traces were stronger predictors of achievement than self-reports. Otieno et al. (2013) investigated whether the use of hints and glossaries in a geometry ITS can be used as an online measure for goal orientation. They found that the online traces differed from self-report data, as the former was more predictive of post-test scores than the latter.

Conversely, Fancsali et al. (2014) re-examined the use of hints and glossaries as online measures of goal orientation. They argued against Otieno et al. (2013) by finding that the online traces were weakly associated with self-efficacy judgments measured via embedded questionnaires.

Despite the conservative opinions on the use of trace logs in learning analytics (see Winne (2020) for trace logs reliability), we believe it is worth taking an extra step toward operationalizing motivation via objective measures; hence, we use the online trace logs for inferring motivation across ITSs. Specifically, we define motivation as simply the desire to learn without concern for ulterior motives, such as interest, mastery, performance, or expected value.

Research Questions

This work addresses three research questions:

- **(RQ1)** How would combining metacognitive knowledge and motivation impact transfer across ITSs?
- **(RQ2)** Would providing instructional interventions for low metacognitive knowledge students facilitate their transfer across ITSs?
- **(RQ3)** Which factors impact transfer for metacognitive knowledge groups?

Methods

We describe the two tutors in this section. We note that two problems are isomorphic if their solutions require the same set of rules or principles.

As the results are reported in multiple sections, we state common notes in this paragraph. The term **"learning performance"** encapsulates the students' scores, which are measured using pre- and post-test, isomorphic scores, and the normalized learning gain (NLG) defined as $NLG = \frac{Post-Pre}{\sqrt{100-Pre}}$, where 100 is the maximum post-test score. NLG measures the improvement from pre-

to post-test, the higher the better (Abdelshiheed, Maniktala, et al., 2022; Abdelshiheed et al., 2021, 2020; Hake, 1998). For reporting results conveniently, we refer to pre-test, post-test and NLG scores as Pre, $Post$ and NLG, respectively. Results with $p < \alpha$ are referred to as **"detectable,"** where $\alpha = .05$ unless Bonferroni correction is used. Finally, when reporting ANCOVA and ANOVA results, all statistical assumptions, such as normality and homoscedasticity, are satisfied but unreported to avoid redundancy and congesting the article with more numeric results.

Logic Tutor

The logic tutor (Barnes et al., 2008) teaches students propositional logic proofs through a standard sequence of pre-test, training and post-test. The three phases share the same interface, but training is the *only* one where students can seek and get help. The pre-test has two problems, while the post-test is harder and has six problems; the first two are isomorphic to the pre-test problems. Training consists of five ordered levels with an *incremental degree of difficulty*, and each level consists of four problems. A problem consists of given nodes at the top and a target node at the bottom, and one needs to derive intermediate nodes by applying valid logic rules such as Modus Ponens and Addition. Each level teaches a new rule for students to apply. Every problem has a score based on students' time, accuracy and solution length. The *pre-* and *post-test* scores are calculated by averaging their pre- and post-test problem scores. The problem score formula is shown in the *Supplementary Materials*.

Throughout the tutor, a student can solve any problem by either a *forward-chaining (FC)* or a *backward-chaining (BC)* strategy. Students know about the two strategies from the Discrete Mathematics lectures they take before the tutor assignment. Figure 1a shows that for *FC*, one must derive the conclusion

at the bottom from givens at the top, while Figure 1b shows that for *BC*, students need to derive a contradiction from givens and the *negation* of the conclusion. Problems are presented by *default* in FC, but students can switch to BC by clicking a button in the tutor interface. The intelligent features of the logic tutor are described in the *Supplementary Materials*.

Probability Tutor

The probability tutor (Chi & VanLehn, 2010) is a web- and text-based tutor that teaches how to solve probability problems using 10 major principles, such as the Complement Theorem and Bayes' Rule. It consists of four sections: textbook, pre-test, training on ITS, and post-test. Similar to the logic tutor, training is the only section for students to receive and ask for hints, and the post-test is harder than the pre-test.

In the textbook, students study the domain principles; In pre- and post-test, students solve 14 and 20 open-ended problems, respectively, that require them to derive an answer by writing and solving one or more equations. Each pre-test problem has a corresponding isomorphic post-test problem. Students' answers are graded in a double-blind manner by experienced graders using a partial-credit rubric, where grades are based *only* on accuracy. The *pre-* and *post-test* scores are the average grades in their respective sections. The details of grading and isomorphic problems on the probability tutor are provided in the *Supplementary Materials*.

The training section interface is seen in Figure 2. It consists of 12 problems, each of which can *only* be solved by BC in that it requires students to derive an answer by writing and solving equations until the target is ultimately reduced to the givens. For each training problem, the tutor records intelligent features for state representation for each student, as described by G. Zhou et al. (2022).

Preliminary Study

We collected data from an undergraduate Discrete Mathematics Computer Science course at North Carolina State University across three semesters. A total of 495 students finished both tutors: $N = 151$ for Fall 2017, $N = 128$ for Spring 2018, and $N = 216$ for Fall 2018. The students' demographics were as follows: age (26.4±4.6, min: 21, max: 60), gender (82% Male, 18% Female), and race (54% White, 15% Asian, 5% Hispanic, 4% Black or African American, 22% Other/Multi/Unknown). We found no detectable difference in the distribution of demographic attributes within and across groups. All students went over the logic tutor described in the **Logic Tutor** section. Six weeks later, students were trained on the probability tutor following the procedure described in the **Probability Tutor** section.

Inferring Metacognitive Knowledge

As discussed in the **Methods** section, students can choose to switch problem-solving strategies *only* on the logic tutor. Thus, we inferred students' metacognitive knowledge based on their interactions with the logic tutor alone. Each problem can be solved by either following the default FC or switching to BC. However, most problems, especially the higher-level ones, can be solved much more efficiently by BC (Abdelshiheed, 2023; Abdelshiheed, Hostetter, Yang, et al., 2022; Abdelshiheed, Hostetter, Shabrina, et al., 2022) and we expect that effective problem solvers should switch their strategy on these problems, and more importantly, they should switch it early when solving them. Thus, our metacognitive knowledge measurement is a combination of ***how*** to use each strategy (Zepeda et al., 2015; Wagster et al., 2007), and ***when*** to use each (de Boer et al., 2018; Winne & Azevedo, 2014). After analyzing log data on the logic tutor (shown in the *Supplementary Materials*), we considered two factors

in learning when to use a strategy: one is that a student should switch in later levels (harder training problems) where the savings will be significant, and the other is that students should switch early (when convenient) while solving a problem. On average, students take 210 actions to solve a problem, and the median number of actions that a student takes before switching is 30.

As stated in the **Logic Tutor** section, training has an incremental degree of difficulty, as each level introduces a new logic rule. Since the rate of change of rules is constant, the tutor difficulty is assumed to be linear in terms of the levels, and therefore, we weighted the **how** and **when** components of learning each strategy by the corresponding level number. Therefore, the metacognitive score (MetaScore) for a student i was calculated[1] as:

$$MetaScore_i = \sum_{L=1}^{5}[\sum_{p=1}^{4}[L * How_{ip} * When_{ip}]] \quad (1)$$

where $How_{ip} = 1$ indicates that student i sustained a switch to use BC when solving problem p at level L, while 0 means unsustained or no switch; $When_{ip} = 1$ if the student i switched early on problem p ($\leq$ median [30 actions]) and $When_{ip} = -1$ for late switch ($>$ median [30 actions]). Based on this formula, $MetaScore_i > 0$ indicates that student i knows how and when to use each strategy; if $MetaScore_i < 0$, it indicates that student i knows how but not good at knowing when; finally, if $MetaScore_i = 0$, this suggests that student i knows neither how nor when by persisting in the default FC settings.

Categorizing Metacognitive Knowledge in the Preliminary Study: Based on $MetaScore$s, students are divided into three groups: those who possess conditional knowledge by knowing how and when ($MetaScore > 0$) are referred to as the ***StrBoth*** group ($N = 145$); those who possess procedural

[1] See the *Supplementary Material* for the rationale of defining the *MetaScore* using this approach.

knowledge by knowing how only ($MetaScore < 0$) as ***StrHow*** ($N = 166$); and the rote students ($MetaScore = 0$) as ***Rote*** ($N = 184$).

Inferring Motivation

Inspired by prior research on behavioral measures of motivation (Touré-Tillery & Fishbach, 2014), we inferred students' motivation by tracking the accuracy of applying principles in their online trace logs. By doing so, we factor in the fact that students often have various ulterior motives. Similar to prior work (Vollmeyer & Rheinberg, 2006; Rheinberg et al., 2000), the motivation in this work is defined based on the initial interactions in the early stages of *each* tutor. In other words, our measured students' *initial* motivation does not consider the fact that students' motivation may change over time. We found that the percentage of correct rule applications in the first two problem-solving questions resulted in a bimodal distribution of students on each tutor —described in the *Supplementary Materials*— and hence was used as the behavioral measure of inferring motivation on each tutor. Due to the bimodal distribution, students were divided into high- and low-motivation groups through a median split; for logic: HM_{Logic} ($N = 248$) and LM_{Logic} ($N = 247$) and for probability: HM_{Prob} ($N = 249$) and LM_{Prob} ($N = 246$). A chi-square test showed no detectable evidence on students staying at the same motivation level across the two tutors: $\chi^2(1,\ N = 495) = 1.26,\ p = 0.26$. In other words, students' motivation levels may change over a semester or change based on subjective domains. Additionally, our motivation definition differs from students' incoming competence in that one-way ANOVA showed no detectable difference on Pre between high- and low-motivation students: $F(1, 493) = 0.7,\ p = .17$ for logic and $F(1, 493) = 0.001,\ p = .98$ for probability.

Results (Prelim. Study)

We examine the impact of 1) metacognitive knowledge alone, 2) motivation alone, and 3) the interactions of the two on students' learning on both tutors.

Metacognitive Knowledge Results (Prelim. Study)

Table 1 illustrates the metacognitive groups' learning performance on the logic and probability tutors. It shows the mean and standard deviation of Pre, $Post$, NLG and isomorphic scores ($Iso.\,Post$ and $Iso.\,NLG$). For the logic tutor, while no detectable difference was found among the three groups on Pre, a one-way ANCOVA analysis using Pre as covariate and metacognitive group as factor showed a detectable difference on $Post$: $F(2, 491) = 17.3$, $p < .001$, $\eta^2 = .3$. Subsequent contrast analyses with Bonferroni[2] adjustment ($\alpha = .05/3 = .016$) showed that $StrBoth$ scored detectably higher than $Rote$: $t(327) = 3.8$, $p < .001$, $d = 2.9$ and $StrHow$: $t(309) = 5.8$, $p < .0001$, $d = 4.5$. Additionally, $Rote$ scored higher than $StrHow$: $t(348) = 2.2$, $p = .03$, $d = 1.6$. While a one-way ANOVA showed no detectable difference among the three groups on NLG, subsequent contrast analyses showed that $StrBoth$ scored higher than $StrHow$: $t(309) = 2.4$, $p = .02$, $d = 3.6$. For the probability tutor, however, no detectable difference was found between any pair of the three groups on any scores.

To summarize, our results suggest that knowing *how* to use each strategy alone can not lead students to learn better on logic; students need to know *when* to use each strategy as well. Additionally, while $StrBoth$ learns detectably better than the other two groups on logic, they did not outperform other groups on probability.

[2] Bonferroni Correction was used for conservative results.

Motivation Level Results (Prelim. Study)

Table 2 compares the high- and low-motivation groups' learning performance across the two tutors. As stated in the **Inferring Motivation** section, no detectable difference was found between high- and low-motivation groups on Pre on each tutor. As expected, a one-way ANCOVA with Pre as covariate and motivation as factor showed that on both tutors, high-motivation students detectably outperformed their low peers on $Post$: $F(1, 492) = 15.8$, $p < .001$, $\eta^2 = .17$ for logic and $F(1, 492) = 24.5$, $p < .001$, $\eta^2 = .17$ for probability. While we found no detectable difference between their NLG on the logic tutor, a one-way ANOVA showed that highly motivated students had detectably higher NLG than their low peers on the probability tutor: $F(1, 493) = 7.6$, $p < .01$, $\eta^2 = .12$.

In short, these results suggest that our motivation measure is reasonable in that the highly motivated students indeed outperformed their low peers on the post-test of each tutor. The former also had detectably higher NLG than the latter on the probability tutor.

Results of Interaction Between Metacognition and Motivation (Prelim. Study)

Logic Tutor:

Combining the metacognitive groups $\{Rote, StrHow, StrBoth\}$ with the logic motivation $\{HM_{Logic}, LM_{Logic}\}$ resulted in six groups. A chi-square test showed that students' motivation did not differ detectably across the three metacognitive groups: $\chi^2(2, N = 495) = 2.87, p = 0.24$. Additionally, no detectable difference was found among the six groups on logic Pre: $F(2, 489) = 0.69$, $p = .49$.

Figure 3a shows the groups' performance on logic *Post*. A two-way ANCOVA using *Pre* as covariate, and metacognitive group and motivation as factors, showed no detectable interaction effect. However, there was a main effect of metacognitive group: $F(2, 488) = 16.6$, $p < .0001$ and motivation: $F(1, 488) = 16.7$, $p < .0001$. Particularly, in each metacognitive group, the HM_{Logic} group outperformed (Bonferroni-corrected ($\alpha = .05/15 = .003$)) the corresponding LM_{Logic} group: $t(182) = 2.1$, $p = .03$, $d = 1.4$ for *Rote*, $t(164) = 3.1$, $p = .002$, $d = 2.4$ for $StrHow$ and $t(143) = 2$, $p = .04$, $d = 1.4$ for $StrBoth$. Among the three HM_{Logic} groups, high-motivation $StrBoth$ outperformed (Bonferroni-corrected ($\alpha = .05/15 = .003$)) their peers: $t(165) = 2.8$, $p = .006$, $d = 2.1$ against high-motivation *Rote* and $t(160) = 3.8$, $p < .001$, $d = 3$ against high-motivation $StrHow$. Among the three LM_{Logic} groups, the same pattern persisted, as the low-motivation $StrBoth$ surpassed their two low-motivation peer groups.

Similarly, for logic NLG (Figure 3b), a two-way ANOVA using the same two factors found no detectable interaction or main effect. However, among the HM_{Logic} groups, high-motivation $StrBoth$ outperformed (Bonferroni-corrected ($\alpha = .05/15 = .003$)) high-motivation *Rote*: $t(165) = 2.2$, $p = .03$, $d = 3.5$ and high-motivation $StrHow$: $t(160) = 2.3$, $p = .03$, $d = 5.4$. We found no detectable difference among the LM_{Logic} groups. In short, our results suggest that the high-motivation $StrBoth$ group performs the best among the six groups in terms of *Post* and NLG on the logic tutor.

Probability Tutor:

Similarly, the metacognitive groups {*Rote*, $StrHow$, $StrBoth$} were combined with the probability motivation {HM_{Prob}, LM_{Prob}} resulting in six groups. A chi-square test showed that students' motivation on probability did not detectably differ across the three metacognitive groups: $\chi^2(2, N = 495) =$

0.53, $p = 0.76$. Moreover, no detectable difference was found among the six groups on probability Pre: $F(2, 489) = 0.5$, $p = .63$. Figures 4a and 4b illustrate the groups' probability $Post$ and NLG, respectively.

Starting by Figure 4a, a two-way ANCOVA with metacognitive knowledge and motivation as factors, and Pre as covariate, showed a detectable interaction effect on $Post$: $F(2, 488) = 3.8$, $p = .02$, $\eta^2 = .09$. Additionally, there was a main effect of motivation in that high-motivation students detectably outperformed their low-motivation peers: $F(1, 488) = 24.4$, $p < .0001$. Among the HM_{Prob} groups, $StrHow$ and $StrBoth$ outperformed (Bonferroni-corrected ($\alpha = .05/15 = .003$)) $Rote$: $t(171) = 2.4$, $p = .02$, $d = 1.7$ and $t(163) = 2.4$, $p = .02$, $d = 1.9$, respectively. However, no detectable difference was found among the LM_{Prob} groups.

Regarding NLG, seen in Figure 4b, a two-way ANOVA using the same two factors showed a detectable interaction effect: $F(2, 489) = 6.4$, $p < .01$, $\eta^2 = .16$. Subsequent contrast analyses with Bonferroni adjustment ($\alpha = .05/15 = .003$) showed that, within $StrHow$ and $StrBoth$, high-motivation students surpassed their low peers: $t(164) = 2.2$, $p = .03$, $d = 2.9$ and $t(143) = 3.8$, $p < .001$, $d = 4.4$, respectively. Across the HM_{Prob} groups, both $StrHow$ and $StrBoth$ had higher NLG than their $Rote$ peers: $t(171) = 2$, $p = .04$, $d = 2.5$ and $t(163) = 3$, $p = .003$, $d = 4.2$, respectively. In brief, on the probability tutor, the high-motivation $StrHow$ and $StrBoth$ groups outperform their peers on $Post$ and NLG.

From a Preliminary Study to Two Experiments

Based on the preliminary study results, we aimed to boost the performance of $Rote$ and $StrHow$ students so they can catch up with their $StrBoth$ peers. Hence, we conducted two experiments to investigate the impact of providing

interventions on the logic tutor to *Rote* and *StrHow* students. In Experiment 1, we provided prompted nudges to recommend switching to BC when proper, and in Experiment 2, we combined nudges with worked examples to teach how and when to use BC. *StrBoth* students received no interventions throughout the experiments, as we feared an expertise reversal effect (Kalyuga, 2009). This effect is based on the finding that instructional guidance may have negative consequences on experienced learners.

We aimed to balance the metacognitive groups $\{Rote, StrHow, StrBoth\}$ across the conditions $\{Experimental, Control\}$. While assigning students to the two conditions at the beginning of the logic tutor is desired, a student's metacognitive group can be determined **only** at the end of the logic training (see Eqn. 1) when it becomes too late to intervene. Hence, we had to find a method to early predict the metacognitive group before the condition assignment.

Early Prediction of Metacognitive Group

We performed a $75-25$ train-test split on students from the preliminary study and trained a random forest classifier (RFC), which takes the mean feature vector collected during the logic pre-test for a student[3] and returns a predicted metacognitive group. In order to avoid overfitting, pruning was performed to limit the excessive use of features within any branch. Moreover, semester-based cross-validation was performed, in which we trained the classifier on two semesters and validated its performance on the third semester. Overall, the RFC achieved a 96.7% accuracy on the testing dataset, as shown in the confusion matrix in Table 3. The RFC was used for early prediction of the metacognitive group in Experiments 1 and 2.

[3] For each problem, we record 152 features per student, as described in the *Supplementary Materials*.

Experiment 1: Prompted Nudges (Exp.1: Nudge)

Consider the number of times you flag an email as spam after seeing the prompt "Report as Spam" or when you remember to buy a grocery item after your spouse's reminder. There are countless similar situations where a nudge influences your decision-making, even when individuals are unaware of such influence, such as the 2016 US elections.

The nudge theory defines a nudge as any factor that alters behavior in a predictable manner without excluding alternatives (Thaler et al., 2013; Thaler & Sunstein, 2008). The theory suggests that nudges have an essential role in behavioral economics (Simon & Tagliabue, 2018) and influence individuals' social and cognitive behaviors (Smith et al., 2013). Thaler wrote in an article for the New York Times that nudging should be guided by three principles: transparency, ease of opting out, and improving the welfare of the individuals being nudged (Thaler, 2015). Considerable research has accommodated the nudge theory in educational research to promote concepts, recommendations, and strategies (Zepeda et al., 2015; Belenky & Nokes-Malach, 2009).

The preliminary study showed that the early switch from forward- to Backward-Chaining (BC) is a desired behavior on the logic tutor. We aimed to boost the performance of *Rote* and *StrHow* students by leveraging the nudge theory for its non-confrontational indirect suggestions. Hence, we conducted an experiment to investigate the impact of showing *Rote* and *StrHow* students **prompted nudges** that recommend switching to BC when proper to do so. We first describe the participants, followed by instructional intervention on the logic tutor, procedure and results.

Participants (Exp.1: Nudge)

Similar to the preliminary study, our participants were students from the same undergraduate course at the same university in Spring 2020. The students' demographics were as follows: age (24 ± 4.1, min: 20, max: 58), gender (80% Male, 20% Female), and race (52% White, 16% Asian, 6% Hispanic, 3% Black or African American, 23% Other/Multi/Unknown). We found no detectable difference in the distribution of demographic attributes within and across conditions and groups.

A total of 64 students completed both tutors and was divided by the RFC into 27 *Rote*, 29 $StrHow$ and 8 $StrBoth$ students who were excluded due to their small sample size. For *Rote* and $StrHow$ students, they were **randomly** assigned to two conditions: $N = 28$ for $Experimental$ —Nudge— (13 $Rote_{Nud}$ + 15 $StrHow_{Nud}$) and $N = 28$ for $Control$ (14 $Rote_{Ctrl}$ + 14 $StrHow_{Ctrl}$). A chi-square test showed no detectable difference in the distribution of *Rote* and $StrHow$ students across the conditions: $\chi^2(1, N = 56) = 0.07$, $p = .79$. To measure the quality of the early metacognitive group predictions, the RFC performance was evaluated on $Control$ students, who received no intervention, yielding a 96.4% accuracy.

Instructional Intervention and Procedure (Exp.1: Nudge)

For this experiment, we modified our logic tutor by offering *prompted nudges* (the text in the black box in Figure 5) that recommend switching to the BC strategy when it is proper to do so. We adhered to the three principles suggested by Thaler (2015) in designing the nudges: transparency, ease of opting out, and improving the welfare of the individuals being nudged. To satisfy transparency, we ensured the text within the nudge was straightforward and explicit, such as "It will save you time if you switch to BC" and "The

tutor thinks switching to BC is an easier option." The nudge box was small in size and placed on the bottom of the interface with a "Close(X)" button to ensure the ease of opting out. We aimed to make the nudges improve the welfare of students being nudged by adopting a data-driven approach to decide **when** to provide a nudge.

Figure 6 shows in green the six problems in which the nudges could be displayed. These problems were determined to be "proper" to be solved by BC; we followed a data-driven approach to select these problems rather than using expert rules or providing nudges based on the student's demands. We analyzed the strategy switch behavior from our preliminary study to guide us *which "proper"* problems to display the nudges in by picking the most frequently switched problems, and *when* to display them, by learning a probability distribution of the duration lengths that students take before switching. More specifically, 55% of the time, the tutor would wait for 1.5 minutes before displaying the nudge, 35% for 3 minutes, and only 10% for 6 minutes. For the remaining problems (colored in white in Figure 6), the tutor behaves the same as the original tutor.

Our goal is to investigate whether recommending students to switch to BC would boost the performance of *Rote* and *StrHow* students, allowing them to catch up with the *StrBoth* group. Therefore, *only* the *Experimental* condition ($Rote_{Nud}$ and $StrHow_{Nud}$) were trained on the modified logic tutor (shown in Fig. 6), while the *Control* ($Rote_{Ctrl}$ and $StrHow_{Ctrl}$) students received no such intervention. Table 4 summarizes our procedure, which is similar to the preliminary study, except for adding the gray section to distinguish the two conditions.

Results (Exp.1: Nudge)

The results are discussed from two perspectives: the **learning performance**, where scores on each tutor are reported for different conditions, metacognitive and motivation groups, and the **strategy switch behavior**, where the students' choices of switching strategy on the logic tutor are analyzed.

Learning Performance (Exp.1: Nudge):

Table 5 compares the learning performance of the four groups on each tutor. A two-way ANOVA using condition $\{Experimental, Control\}$ and metacognitive group $\{Rote, StrHow\}$ as factors showed no detectable difference on Pre on each tutor: $F(1, 52) = 0.1$, $p = .71$ for logic and $F(1, 52) = 0.7$, $p = .41$ for probability. Next, the scores are analyzed on each tutor separately.

Starting with logic, a two-way ANCOVA using Pre as covariate, and condition as well as metacognitive group as factors, found a detectable interaction effect on $Post$: $F(1, 51) = 6.4$, $p = .01$, $\eta^2 = .14$. Subsequent contrast analyses with Bonferroni adjustment ($\alpha = .05/6 = .008$) showed that $StrHow_{Nud}$ had higher $Post$ than their $Control$ peers $StrHow_{Ctrl}$: $t(27) = 2.4$, $p = .02$, $d = 2.5$ and $Experimental$ peers $Rote_{Nud}$: $t(26) = 2.3$, $p = .03$, $d = 1.9$. Analyzing logic NLG, a two-way ANOVA using the same two factors showed a detectable interaction effect: $F(1, 52) = 9.4$, $p < .01$, $\eta^2 = .23$. Follow-up Bonferroni-corrected analyses ($\alpha = .05/6 = .008$) revealed that $StrHow_{Nud}$ outperformed $StrHow_{Ctrl}$ and $Rote_{Nud}$: $t(27) = 2.7$, $p = .01$, $d = 1.4$ and $t(26) = 2.4$, $p = .02$, $d = 1.8$, respectively.

Similar patterns were found on probability but with more detectable results. A two-way ANCOVA using the same two factors with Pre as covariate found a detectable interaction effect on $Post$: $F(1, 51) = 10.3$, $p < .01$, $\eta^2 = .37$, in that $StrHow_{Nud}$ had detectably higher (Bonferroni-corrected ($\alpha =$

.05/6 = .008)) *Post* than $StrHow_{Ctrl}$: $t(27) = 3.2$, $p = .003$, $d = 2.2$ and $Rote_{Nud}$: $t(26) = 3.1$, $p = .004$, $d = 1.9$. For NLG, we observed a detectable interaction effect after carrying out a two-way ANOVA using the same factors: $F(1, 52) = 11.2$, $p = .001$, $\eta^2 = .4$, as $StrHow_{Nud}$ detectably surpassed (Bonferroni-corrected ($\alpha = .05/6 = .008$)) $StrHow_{Ctrl}$: $t(27) = 4.6$, $p < .0001$, $d = 2$ and $Rote_{Nud}$: $t(26) = 3$, $p = .006$, $d = 1.7$. On both tutors, no detectable difference was found between $Rote_{Nud}$ and $Rote_{Ctrl}$ in any of the measures shown in Table 5.

To sum up, $StrHow_{Nud}$ students detectably outperformed their peers on $Post$ and NLG scores on both tutors, while $Rote_{Nud}$ did not benefit from our intervention, as they did not show any detectable advantage over their *Control* peers.

Strategy Switch Behavior (Exp.1: Nudge)

To investigate whether the instructional intervention impacted the students' strategic behaviors, we analyzed such behaviors on the logic tutor. Figure 7 shows the strategy switch behaviors of the four groups (from FC to BC); we compared their decisions during the logic tutor $Training$, where only $Rote_{Nud}$ and $StrHow_{Nud}$ were offered nudges, and $Post$ (post-test), where no student got nudges. By following the definitions described after Eqn. 1, we display the percentage of *No Switches* (sticking to the default strategy), *Early Switches* (switching within the first 30 actions), and *Late Switches* (switching after the first 30 actions). In Figure 7, the four groups are ordered by the percentage of *Early Switches*, the most desired behavior.

A two-way ANOVA using condition and metacognitive group as factors showed a detectable interaction effect on *Early Switches*: $F(1, 52) = 14.1$, $p < .001$, $\eta^2 = .26$ for $Training$ and $F(1, 52) = 12.6$, $p < .001$, $\eta^2 = .19$ for $Post$. Subsequent Bonferroni-corrected analyses ($\alpha = .05/6 = .008$) showed

that $StrHow_{Nud}$ made early strategy switches detectably more than the other three groups; for instance, they detectably surpassed $Rote_{Nud}$ on $Training$ and $Post$: $t(26) = 5.4$, $p < .0001$, $d = 1.4$ and $t(26) = 4.9$, $p < .0001$, $d = 1.3$, respectively. On the other hand, no detectable differences were found between the two *Control* groups, or between the two *Rote* groups.

In other words, $StrHow_{Nud}$ students showed substantial compliance with the nudges and were able to switch strategy early throughout the tutor, even on post-test questions, where no nudges were given. However, their peers failed to switch strategy due to the lack of knowledge about BC ($Rote_{Nud}$), lack of nudges ($StrHow_{Ctrl}$), or lack of both ($Rote_{Ctrl}$).

Experiment 2: Worked Examples & Nudges (Exp.2: Instruction)

Our findings from Experiment 1 suggest that recommending students to switch to BC might not help all students, as *Rote* students lack knowledge about the BC strategy. Therefore, we reinforced our intervention by adding worked examples (WE) to teach students how to solve problems using BC. In cognitive load theory, the worked-example effect refers to the observed learning outcome from teaching with worked examples compared to other approaches, such as problem-solving (Renkl, 2005). A Worked Example (WE) is a step-by-step solution that solves a problem or completes a task. Renkl (2005) stated that WEs are designed to support metacognitive skills acquisition by introducing a problem, its solution steps, and the final solution. Substantial work has leveraged WEs to enhance students' problem-solving skills and prepare them for explicit instruction (Likourezos & Kalyuga, 2017; Glogger-Frey et al., 2015).

In Experiment 2, we investigated how explicit instruction on *how* (WE) and *when* (nudges) to use BC would impact students' learning across the two tutors

and, more importantly, whether such instruction would further eliminate the gap among different learners.

Participants (Exp.2: Instruction)

The participants were from the same undergraduate course at the same university in Fall 2020. The demographics were as follows: age (23.9 ± 4.7, min: 20, max: 56), gender (82% Male, 18% Female), and race (54% White, 22% Asian, 5% Hispanic, 3% Black or African American, 16% Other/Multi/Unknown). No detectable difference was found in the distribution of demographic attributes within and across conditions and groups.

A total of 128 students completed both tutors, and our RFC divided them into 60 *Rote*, 42 *StrHow* and 26 *StrBoth* students. Like Experiment 1, *Rote* and *StrHow* students were **randomly** assigned to two conditions[4]: $N = 61$ for *Experimental* —Instruction— (35 $Rote_{Ins}$ + 26 $StrHow_{Ins}$) and $N = 41$ for *Control* (25 $Rote_{Ctrl}$ + 16 $StrHow_{Ctrl}$). No detectable difference was found in the distribution of *Rote* and *StrHow* students across the two conditions: $\chi^2(1, N = 102) = 0.13$, $p = .72$. Regarding *StrBoth* students, they used the original logic tutor without any intervention. The accuracy of the RFC is further evaluated on the *Control* and *StrBoth* students since they received no intervention. Our results showed that the RFC achieved 95.5% accuracy, similar to its performance in the preliminary study and Experiment 1.

Instructional Intervention and Procedure (Exp.2: Instruction)

Compared to Experiment 1, we modified our logic tutor by adding two WEs on explicit BC strategy instruction as the first problem in the first two levels,

[4] The difference in size is due to the fact that we prioritized having a sufficient number of *Experimental* students to perform a meaningful analysis of our intervention.

as shown in Figure 8. Our goal is to investigate whether explicit BC strategy instruction using WEs combined with nudges would make *Rote* and *StrHow* catch up with *StrBoth*. We expect that the former two groups would benefit from our intervention designed to scaffold the metacognitive knowledge that they lack. On the other hand, for *StrBoth* students, we expect that providing them with additional scaffolding could interfere with their existing metacognitive knowledge. Therefore, *only* the *Experimental Rote* and *StrHow* groups will get the treatment shown in Figure 8, while the *Control Rote* and *StrHow* groups and the *StrBoth* group will get no treatment. Experiment 2 procedure is similar to Experiment 1 (Table 4), except that for Experiment 2, the interventions are shown in Figure 8 and *StrBoth* students are included and receive the original tutor, like their *Control* peers.

Results (Exp.2: Instruction)

The results are organized into two sections similar to Experiment 1: **learning performance** and **strategy switch behavior**. The first section is divided into several parts; first, we compare the *Experimental* and *Control* conditions. Then we break down the conditions into metacognitive groups and compare them with each other and the *StrBoth* group. Next, the motivation distribution is shown for all groups, and finally, the impact of motivation on the groups' performance is discussed.

Learning Performance (Exp.2: Instruction):

Experimental vs Control

Table 6 compares the two conditions across the tutors showing various metrics' mean and standard deviation. The last column shows the one-way ANOVA

comparisons between the two conditions, including the effect size η^2. As shown in the table, while no detectable difference was found between the two conditions on Pre: $F(1, 100) = 0.8$, $p = .38$ for logic and $F(1, 100) = 2.7$, $p = .11$ for probability, $Experimental$ detectably outperformed $Control$ in all other aspects.

Comparing Metacognitive Groups within Conditions

To investigate whether $Rote$ and $StrHow$ students benefited from our intervention, we compared their performance across the two conditions, as shown in the first five columns in Table 7.

Regarding the logic tutor performance, A two-way ANOVA using condition $\{Experimental, Control\}$ and metacognitive group $\{Rote, StrHow\}$ as factors showed no detectable difference on Pre: $F(1, 98) = 0.28$, $p = .6$. A two-way ANCOVA using the same factors with Pre as covariate found a detectable interaction effect on $Post$: $F(1, 97) = 17.3$, $p < .0001$, $\eta^2 = .06$. Follow-up contrast analyses with Bonferroni adjustment ($\alpha = .05/6 = .008$) revealed that while no detectable difference was found between the $Control$ groups, a detectable difference was found between the $Experimental$ groups: $Rote_{Ins} > StrHow_{Ins}$ ($t(59) = 2.9$, $p = .005$, $d = 4.3$). Additionally, each $Experimental$ group detectably surpassed its respective $Control$. These findings show that $Rote_{Ins} > StrHow_{Ins} > Rote_{Ctrl}, StrHow_{Ctrl}$. A two-way ANOVA using the same factors showed similar patterns on NLG, except that $StrHow_{Ins}$ learned no different from the $Control$ groups —$Rote_{Ctrl}$ and $StrHow_{Ctrl}$.

On the probability tutor, a two-way ANOVA using condition and metacognitive group as factors showed no detectable difference on Pre: $F(1, 98) = 0.05$, $p = .82$. Additionally, a two-way ANCOVA using the same factors with Pre as covariate showed no detectable interaction effect on $Post$. Subsequent contrast analyses showed that no detectable difference was found between

the *Experimental* groups, or between the *Control* groups. However, each *Experimental* group detectably outperformed its respective *Control*, suggesting that $Rote_{Ins}$, $StrHow_{Ins} > Rote_{Ctrl}$, $StrHow_{Ctrl}$. Similar findings were observed on NLG.

Comparing with StrBoth Group

The last column in Table 7 shows the performance of $StrBoth$ across all measures. We further explored the effectiveness of our intervention from two aspects: 1) whether it would make *Rote* and $StrHow$ students in the *Experimental* condition catch up with $StrBoth$, and 2) whether, without such intervention, the students in the *Control* condition would perform worse than $StrBoth$.

As for the first aspect, our results show that the *Experimental* condition performed as well as or better than $StrBoth$ in that no detectable difference was found between the two on all measures on logic and probability. Next, we individually compared the two *Experimental* groups, $Rote_{Ins}$ and $StrHow_{Ins}$, against $StrBoth$. We found no detectable difference between the three groups on logic and probability Pre, as shown in Table 7. While no detectable difference was found between $StrBoth$ and $StrHow_{Ins}$ on logic, $Rote_{Ins}$ outperformed (Bonferroni-corrected ($\alpha = .05/3 = .016$)) $StrBoth$ on logic $Post$: $t(59) = 3.2$, $p = .002$, $d = 2$. For probability, no detectable difference was found among $Rote_{Ins}$, $StrHow_{Ins}$, and $StrBoth$ across all measures.

As for the second aspect, as expected, *StrBoth* outperformed *Control* on *Post* on both tutors: $t(65) = 2.4$, $p = .02$ for logic and $t(65) = 3.8$, $p < .001$ on probability. After individually comparing the two *Control* groups against $StrBoth$, we found no detectable difference between the three groups on logic and probability Pre, as shown in Table 7. However, $StrBoth$ detectably outperformed (Bonferroni-corrected ($\alpha = .05/3 = .016$)) the two *Control* groups on logic $Post$: $t(49) = 3$, $p = .004$, $d = 1.7$ for $Rote_{Ctrl}$

and $t(40) = 3.9$, $p < .001$, $d = 1.1$ for $StrHow_{Ctrl}$ and probability $Post$: $t(49) = 4.7$, $p < .0001$, $d = 3.1$ for $Rote_{Ctrl}$ and $t(40) = 3.5$, $p < .001$, $d = 2.2$ for $StrHow_{Ctrl}$.

To summarize, these findings show that with our instructional intervention, $Experimental$ indeed caught up with $StrBoth$ as the former performed at least as well as $StrBoth$ on both tutors. On the other hand, without the intervention, $StrBoth$ outperformed $Control$ on the two tutors. Specifically, our results showed that the intervention was most beneficial to $Rote$ students as $Rote_{Ins}$ surpassed all other groups, including $StrBoth$, on logic and continued to perform well on probability.

Impact of Motivation on Performance

Figure 9 shows the Pre and $Post$ scores for the two conditions (on left) and the $StrBoth$ group (on right). For the logic tutor (Fig. 9a), we analyzed the left subfigure by performing a two-way repeated measures ANOVA on the scores using condition $\{Ins, Ctrl\}$ and logic motivation $\{HM_{Logic}, LM_{Logic}\}$ as factors. While we found no detectable interaction effect, there was a main effect of condition: $F(1, 98) = 8.08$, $p < .01$, in that the two $Experimental$ groups detectably outperformed their $Control$ peers on logic $Post$. Regarding $StrBoth$, shown on right, we carried out a one-way repeated measures ANOVA on the scores using logic motivation as factor. The results showed that the high-motivation $StrBoth$ detectably outperformed their low-motivation peers on logic $Post$: $F(1, 24) = 6$, $p = .02$.

Analyzing the probability scores (Fig. 9b), we found similar patterns to those observed on logic (Fig. 9a). For the left subfigure, a two-way repeated measures ANOVA on the scores, using condition $\{Ins, Ctrl\}$ and probability motivation $\{HM_{Prob}, LM_{Prob}\}$ as factors, showed no detectable interaction effect. However, similar to Figure 9a, there was a main effect of condition in

favor of the two *Experimental* groups in their *Post*: $F(1, 98) = 5.6$, $p = .02$. Regarding the right subfigure, a one-way repeated measures ANOVA on the scores, using probability motivation as factor, showed that the high-motivation *StrBoth* group had a detectably higher probability *Post* than their low-motivation peers: $F(1, 24) = 19.3$, $p < .01$.

To sum up, it seems that motivation played an important role in distinguishing high and low learners in the *StrBoth* group. On the other hand, for *Experimental* and *Control* conditions, no detectable difference was observed in the learning performance between high- and low-motivation students.

Strategy Switch Behavior (Exp.2: Instruction)

Figure 10 shows the strategy switch behavior of the five groups. The groups are ordered by the percentage of *Early Switches*, and *StrBoth* is highlighted in bold as the gold standard. A one-way ANOVA found that the switch behaviors differed detectably among the five groups: $F(4, 123) = 71.2$, $p < .0001$, $\eta^2 = .7$ for $Training$ and $F(4, 123) = 62.6$, $p < .0001$, $\eta^2 = .67$ for $Post$. More importantly, the behaviors of each group were very similar between $Training$ and $Post$. Subsequent contrast analyses showed that while no detectable difference was observed between $Rote_{Ins}$ and $StrBoth$ on their switch behaviors, both groups switched early detectably more than the other three groups: $StrHow_{Ins}$, $StrHow_{Ctrl}$ and $Rote_{Ctrl}$. For example, $StrBoth$ switched early detectably (Bonferroni-corrected ($\alpha = .05/10 = .005$)) more than $StrHow_{Ins}$: $t(50) = 5.4$, $p < .0001$, $d = 1.4$ for $Training$ and $t(50) = 4.9$, $p < .0001$, $d = 1.3$ for $Post$.

In short, analyzing strategy switch behaviors confirms that $Rote_{Ins}$ indeed caught up with $StrBoth$, as the former showed very similar behaviors to the latter during the training when the intervention was available and, more importantly, during the post-test when such intervention was not present. On the

other hand, much to our surprise, the strategy switch behaviors of $StrHow_{Ins}$ stayed similar to their *Control* peers, $StrHow_{Ctrl}$.

Post-hoc Analysis

We present a post-hoc analysis by combining the two experiments' results. Table 8 summarizes the logic and probability scores for seven groups: two from Experiment 1 ($Rote_{Nud}$, $StrHow_{Nud}$), two from Experiment 2 ($Rote_{Ins}$, $StrHow_{Ins}$), two for *Control* ($Rote_{Ctrl}$, $StrHow_{Ctrl}$) and the *StrBoth* group. For the *Control* groups, we combined the data from the two experiments, and for *StrBoth*, we added the eight students excluded from Experiment 1 due to their small sample size.

A one-way ANOVA using group as factor found no detectable difference between all groups on *Pre*: $F(6, 185) = 1.1$, $p = .36$ for logic and $F(6, 185) = 0.8$, $p = .57$ for probability. Next, we compare the conditions then we compare the metacognitive groups. We will refer to experimental students in Experiments 1 and 2 as Nud and Ins, respectively.

Comparing Conditions: Nud vs Ins vs Control

We performed a series of pairwise contrast analyses (Bonferroni-corrected ($\alpha = .05/3 = .016$)) to compare *Post* and NLG for the three conditions $\{Nud, Ins, Control\}$ on the two tutors. Starting with the logic tutor, we found that $\boldsymbol{Ins > Nud > Control}$ on *Post* and NLG; specifically, $Ins > Nud$ ($t(87) = 2.4$, $p = .02$ for *Post* and $t(87) = 2.5$, $p = .01$ for NLG) and $Nud > Control$ ($t(95) = 2.1$, $p = .03$ for *Post* and $t(95) = 2.2$, $p = .02$ for NLG).

For the probability tutor, we found that $\boldsymbol{Ins, Nud > Control}$, as no detectable difference was found between Nud and Ins on *Post* or NLG, while

both conditions outperformed $Control$: $Ins > Control$ ($t(128) = 3.3$, $p < .01$ for $Post$ and $t(128) = 3.1$, $p < .01$ for NLG) and $Nud > Control$ ($t(95) = 2.2$, $p = .03$ for $Post$ and $t(95) = 2.3$, $p = .02$ for NLG).

Comparing Metacognitive Groups

We conducted a series of pairwise contrast analyses to compare $Post$ and NLG for the three $Rote$ groups alone $\{Rote_{Nud}, Rote_{Ins}, Rote_{Ctrl}\}$, the three $StrHow$ groups alone $\{StrHow_{Nud}, StrHow_{Ins}, StrHow_{Ctrl}\}$, and then compare all groups with $StrBoth$. Similar patterns between $Post$ and NLG were found, which will be summarized next.

Regarding the $Rote$ groups, the comparisons revealed that $\boldsymbol{Rote_{Ins} > Rote_{Nud}, Rote_{Ctrl}}$ on logic and $\boldsymbol{Rote_{Ins} > Rote_{Nud} > Rote_{Ctrl}}$ on probability. In other words, $Rote_{Ins}$ detectably outperformed $Rote_{Nud}$ on both tutors. For the $StrHow$ groups, the analyses showed that $\boldsymbol{StrHow_{Nud}}$, $\boldsymbol{StrHow_{Ins} > StrHow_{Ctrl}}$ on both tutors, as no detectable difference was found between $StrHow_{Nud}$ and $StrHow_{Ins}$.

Comparing with $StrBoth$, we found $\boldsymbol{Rote_{Ins} > StrBoth}$, $\boldsymbol{StrHow_{Nud}}$, $\boldsymbol{StrHow_{Ins} > Rote_{Nud}, Rote_{Ctrl}, StrHow_{Ctrl}}$ on logic, while $\boldsymbol{StrBoth}$, $\boldsymbol{Rote_{Nud}, StrHow_{Nud}, Rote_{Ins}, StrHow_{Ins} > Rote_{Ctrl}, StrHow_{Ctrl}}$ on probability. In brief, $StrBoth$ detectably outperformed the $Control$ groups on both tutors as expected. While the experimental $Rote$ and $StrHow$ groups caught up with $StrBoth$ on probability, $Rote_{Ins}$ surprisingly outperformed $StrBoth$ on logic.

General Discussion, Our MMI Framework, and Broader Impacts

We summarize our findings by addressing the research questions of this work.

RQ1 (Combining Metacognitive Knowledge and Motivation):

The preliminary study confirms the importance of motivation in that on both tutors, the impact of metacognitive knowledge on student learning appeared only among the highly motivated students. In contrast, for low-motivation students, no detectable difference was found within the three metacognitive groups. These findings confirm that our choice of using the accuracy of online traces on the first two questions is a reasonable way to measure students' initial motivation levels.

Our results demonstrate the distinction between knowing how and when to use each strategy. Students who knew both ($StrBoth$) transferred their conditional knowledge across the two tutors, which confirms our $MetaScore$ definition of measuring metacognitive knowledge (see Equation 1). Finally, we emphasize the impact of combining metacognitive knowledge and motivation on facilitating transfer. We found that high-motivation $StrBoth$ students consistently performed best on both tutors despite receiving no interventions.

RQ2 (Providing Instructional Interventions for Students with Low Metacognitive Knowledge):

The two experiments aimed to provide instructional interventions for $StrHow$ and *Rote* students based on the metacognitive knowledge they lack. $StrHow$ students leveraged the nudges on both experiments, which taught them *when* to use backward chaining. Meanwhile, *Rote* students benefited most from Experiment 1 due to the combination of worked examples and nudges, which taught them *how* and *when* to use backward chaining, respectively. However, *Rote* students did not perform well in Experiment 2, as the nudges alone would not teach both the how and when.

RQ3 (Factors Impacting Transfer for Metacognitive Groups):

We discuss our findings for each metacognitive knowledge group, then present a framework that summarizes these results.

StrBoth: Although $StrBoth$ and $Control$ students received no interventions on both experiments, the former detectably outperformed the latter on both tutors. This finding supports the belief that some students learn regardless of the environment (Kanfer & Ackerman, 1989), and we argue that metacognitive knowledge plays a vital role in this outcome. Additionally, motivation was a decisive factor among $StrBoth$ students, as highly motivated $StrBoth$ students consistently outperformed their low-motivation peers in the preliminary study and Experiment 2 on both tutors.

Rote: Their best performance occurred in Experiment 2 due to receiving prompted nudges and worked examples; $Rote_{Ins}$ detectably outperformed all groups on logic, including $StrBoth$ —the gold standard— and caught up with $StrBoth$ on probability. Therefore, it is evident that $Rote$ students benefited from the ***combination*** of worked examples and prompted nudges that taught them how and when to use the backward-chaining strategy and, as a result, encouraged them to try a strategy other than the default forward chaining. However, when the worked examples disappeared —as in Experiment 1— the students' performance deteriorated, as the nudges only recommended the strategy switch, which is irrelevant without knowing such a strategy.

StrHow: Unlike their $Rote$ peers, $StrHow$ students benefited equally from the two experiments; $StrHow_{Nud}$ and $StrHow_{Ins}$ detectably surpassed their respective control peers and caught up with $StrBoth$ on both tutors, while there was no detectable difference between $StrHow_{Nud}$ and $StrHow_{Ins}$ on either tutor. Hence, $StrHow$ students leveraged the ***prompted nudges*** in both experiments that taught them *when* to switch to BC, which is the conditional metacognitive knowledge that they lacked before.

Based on these findings, we propose a ***M**etacognitive knowledge, initial **M**otivation, and instructional **I**nterventions (MMI)* framework for transfer across ITSs. Table 9 summarizes our framework by highlighting which factors in each metacognitive knowledge group will likely facilitate transfer. The factors include the students' motivation level and the provided interventions in the two experiments. Note that the logic and probability motivation were combined into one row, as this framework makes the same claim about them. The green and red table entries claim whether students within a metacognitive group will likely transfer their knowledge by possessing high motivation or receiving intervention(s). The gray entries are empty to reflect that no claim can be made due to the absence of experimentation. In other words, no interventions were provided to $StrBoth$ students, as we prioritized treating them as the gold standard to compare them to our experimental students. Our framework can be summarized as follows:

1. Possessing high motivation is a detectable measure of transfer only among $StrBoth$. However, this is not the case for their *Rote* and $StrHow$ peers regardless of receiving interventions (in the two experiments) or not (in the preliminary study).
2. $StrHow$ students show signs of transfer when receiving prompted nudges about *when* to switch strategy, while *Rote* students require the combination of nudges (when) and worked examples (how) to achieve transfer. These outcomes are observed regardless of the motivation level on both tutors.

Limitations and Broader Impacts

Despite these findings, we emphasize that our work had at least four caveats. First, our measurement of the students' motivation used the first two problems on each tutor and did not consider that students' motivation may vary

during the training. Second, splitting students into experimental and control conditions resulted in small sample sizes in Experiment 1. Third, the logic tutor offered a default strategy, and the probability tutor supported only one strategy. A more convincing testbed would be having the tutors support both strategies, where students are asked to choose the default strategy. Finally, our framework is based on our definitions of metacognitive knowledge and motivation, the provided interventions, and the two ITSs.

The future work involves utilizing adaptive methodologies (Abdelshiheed et al., 2023a, 2023b; Hostetter et al., 2023) to determine when to prompt a nudge. We believe our work has **broader impacts** that can be summarized as follows:

1. Our operationalization of motivation was based on the initial accuracy of trace logs, which contributed to predicting and quantifying transfer across ITSs. On the other hand, motivation theories, such as Achievement Goal Theory (AGT), rely on self-report measures that are subjective and hard to generalize. We speculate that initial accuracy could, with further research, be used to measure achievement goals and hence substitute self-report measures. We believe that AGT is the most explicit theory in relating motivation to ***performance***, whether performing the task extremely well (known as mastery in AGT) or performing the task better than other individuals (known as performance in AGT). We believe AGT and initial accuracy would likely align, especially since students with high mastery or performance approach would likely be cautious (accurate) in applying each step during the initial phases of learning. The remaining motivation theories are less explicit about performance; instead, they measure other dimensions of motivation, like the expected value, fulfillment, or satisfaction (Eccles & Wigfield, 2020; Eccles, 1983).

2. Our work showed a significant distinction between knowing how and when to use each strategy, which highlights the need for careful consideration in designing interventions and ITSs. Considering the nudge intervention as an example, it is cognitive if it tells a student to apply or undo a rule, but what makes it **metacognitive** in our work is that it tells students to reconsider their thinking of the strategy from the root, and it fulfills the realization of **when** to tell them so. Despite being non-confrontational, a nudge will likely prompt a set of ***meta-questions*** that are not limited to the current strategy but whether picking that strategy from the beginning was the right choice. Examples of such questions include "Why did the tutor ask me to reconsider my strategy instead of helping with my current solution? Does this mean the default strategy is the wrong one from the beginning?", "Why did this nudge appear exactly now? Why not ten seconds ago or one minute later? Was there a certain behavior I did wrong with the current strategy?", and "My current progress in the default (FC) strategy looks like this. If I switch to the alternative (BC) strategy, I think it will look like that. What makes the new strategy easier?" Several interventions and ITSs are designed to teach how to apply or memorize a procedure without making students question the rationale and timing of using each procedure and subroutine.

Supplementary Materials. Our *Supplementary Materials* can be found here: **10.13140/RG.2.2.35859.45601/1**.

Acknowledgments. This research was supported by the NSF Grants: 1651909, 1660878, 1726550, and 2013502.

Statements and Declarations

– **Conflict of Interests**: The authors have no conflict of interests.

– **Authors' Contribution**: All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Mark Abdelshiheed. The first draft of the manuscript was written by Mark Abdelshiheed, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

Abdelshiheed, M. (2023). *Combining reinforcement learning and three learning theories to achieve transfer and bridge metacognitive knowledge gap.* North Carolina State University.

Abdelshiheed, M., Hostetter, J. W., Barnes, T., & Chi, M. (2023a). Bridging declarative, procedural, and conditional metacognitive knowledge gap using deep reinforcement learning. In *Proceedings of the 45th annual conference of the cognitive science society.*

Abdelshiheed, M., Hostetter, J. W., Barnes, T., & Chi, M. (2023b). Leveraging deep reinforcement learning for metacognitive interventions across intelligent tutoring systems. In *Proceedings of the 24th international conference on artificial intelligence in education.*

Abdelshiheed, M., Hostetter, J. W., Shabrina, P., Barnes, T., & Chi, M. (2022). The power of nudging: Exploring three interventions for metacognitive skills instruction across intelligent tutoring systems. In *Proceedings of the 44th annual conference of the cognitive science society* (pp. 541–548).

Abdelshiheed, M., Hostetter, J. W., Yang, X., Barnes, T., & Chi, M. (2022). Mixing backward- with forward-chaining for metacognitive skill acquisition and transfer. In *Proceedings of the 23rd international conference on artificial intelligence in education* (pp. 546–552).

Abdelshiheed, M., Maniktala, M., Barnes, T., & Chi, M. (2022). Assessing competency using metacognition and motivation: The role of time-awareness in preparation for future learning. In *Design recommendations for intelligent tutoring systems* (Vol. 9, pp. 121–131). US Army Combat Capabilities Development Command–Soldier Center.

Abdelshiheed, M., Maniktala, M., Ju, S., Jain, A., Barnes, T., & Chi, M. (2021). Preparing unprepared students for future learning. In *Proceedings of the 43rd annual conference of the cognitive science society* (pp. 2547–2553).

Abdelshiheed, M., Zhou, G., Maniktala, M., Barnes, T., & Chi, M. (2020). Metacognition and motivation: The role of time-awareness in preparation for future learning. In *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 945–951).

Azevedo, R., Taub, M., & Mudrick, N. V. (2017). Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies. In *Handbook of self-regulation of learning and performance* (pp. 254–270). Routledge.

Barnes, T., Stamper, J. C., Lehmann, L., & Croy, M. J. (2008). A pilot study on logic proof tutoring using hints generated from historical student data. In *Edm* (pp. 197–201).

Belenky, D., & Nokes-Malach, T. (2009). Examining the role of manipulatives and metacognition on engagement, learning, and transfer. *The Journal of Problem Solving*, *2*(2), 6. doi: 10.7771/1932-6246.1061

Belenky, D., & Nokes-Malach, T. (2013). Mastery-approach goals and knowledge transfer: An investigation into the effects of task structure and framing instructions. *Learning and individual differences*, *25*, 21–34.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of research in education*, *24*(1),

61–100. doi: 10.3102/0091732X024001061

Chi, M., & VanLehn, K. (2010). Meta-cognitive strategy instruction in intelligent tutoring systems: How, when, and why. *Educational Technology & Society*, *13*(1), 25–39.

de Boer, H., Donker, A. S., Kostons, D. D., & van der Werf, G. P. (2018). Long-term effects of metacognitive strategy instruction on student academic performance: A meta-analysis. *Educational Research Review*, *24*, 98–115.

Detterman, D. K., & Sternberg, R. J. (1993). *Transfer on trial: Intelligence, cognition, and instruction.* Ablex Publishing.

Dweck, C. S. (1986). Motivational processes affecting learning. *American psychologist*, *41*(10), 1040.

Eccles, J. (1983). Expectancies, values and academic behaviors. *Achievement and achievement motives*.

Eccles, J., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary educational psychology*, *61*, 101859.

Elliot, A. J. (2005). A conceptual history of the achievement goal construct.

Fancsali, S., Bernacki, M., Nokes-Malach, T., Yudelson, M., & Ritter, S. (2014). Goal orientation, self-efficacy, and "online measures" in intelligent tutoring systems. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Fulmer, S. M., & Frijters, J. C. (2009). A review of self-report and alternative approaches in the measurement of student motivation. *Educational Psychology Review*, *21*(3), 219–246. doi: 10.1007/s10648-009-9107-x

Georgeff, M. P., & Lansky, A. L. (1986). Procedural knowledge. *Proceedings of the IEEE*, *74*(10), 1383–1398.

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., & Renkl, A. (2015).

Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learning and Instruction*, *39*, 72–87.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, *66*(1), 64–74.

Hostetter, J. W., Abdelshiheed, M., Barnes, T., & Chi, M. (2023). A self-organizing neuro-fuzzy q-network: Systematic design with offline hybrid learning. In *Proceedings of the 22nd international conference on autonomous agents and multiagent systems (aamas)* (pp. 1248–1257).

Kalyuga, S. (2009). The expertise reversal effect. In *Managing cognitive load in adaptive multimedia learning* (pp. 58–80). IGI Global.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of applied psychology*, *74*(4), 657. doi: 10.1037/0021-9010.74.4.657

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory into practice*, *41*(4), 212–218.

Likourezos, V., & Kalyuga, S. (2017). Instruction-first and problem-solving-first approaches: alternative pathways to learning complex tasks. *Instr. Sci.*, *45*, 195–219. doi: 10.1007/s11251-016-9399-4

Livingston, J. A. (2003). *Metacognition: An overview.* ERIC.

Nokes-Malach, T., & Belenky, D. (2011). Incorporating motivation into a theoretical framework for knowledge transfer. *Cognition in Education*, 109. doi: 10.1016/B978-0-12-387691-1.00004-1

Otieno, C., Schwonke, R., Salden, R., & Renkl, A. (2013). Can help seeking behavior in intelligent tutoring systems be used as online measure for goal orientation? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Renkl, A. (2005). The worked-out-example principle in multimedia learning.

The Cambridge handbook of multimedia learning, 229–245.

Rheinberg, F., Vollmeyer, R., & Rollett, W. (2000). Motivation and action in self-regulated learning. In *Handbook of self-regulation* (pp. 503–529). Elsevier. doi: 10.1016/B978-012109890-2/50044-5

Richey, J. E., Zepeda, C. D., & Nokes-Malach, T. (2015). Transfer effects of prompted and self-reported analogical comparison and self-explanation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 37).

Roberts, M. J., & Erdos, G. (1993). Strategy selection and metacognition. *Educational Psychology*, *13*, 259–266. doi: 10.1080/0144341930130304

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional science*, *26*(1-2), 113–125.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary educational psychology*, *19*(4), 460–475.

Schraw, G., & Gutierrez, A. P. (2015). Metacognitive strategy instruction that highlights the role of monitoring and control processes. In *Metacognition: Fundaments, applications, and trends* (pp. 3–16). Springer.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational psychology review*, *7*, 351–371.

Simon, C., & Tagliabue, M. (2018). Feeding the behavioral revolution: Contributions of behavior analysis to nudging and vice versa. *Journal of Behavioral Economics for Policy*, *2*(1), 91–97.

Smith, N. C., et al. (2013). Choice without awareness: Ethical and policy implications of defaults. *Journal of Public Policy & Marketing*, *32*(2), 159–172.

Thaler, R. (2015). The power of nudges, for good and bad. *The New York Times*. (Available at: `https://www.nytimes.com/2015/11/01/upshot/the-power-of-nudges-for-good-and-bad.html`)

Thaler, R., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health,*

wealth, and happiness. HeinOnline.

Thaler, R., Sunstein, C. R., & Balz, J. P. (2013). Choice architecture. *The behavioral foundations of public policy*, *25*, 428–439.

Touré-Tillery, M., & Fishbach, A. (2014). How to measure motivation: A guide for the experimental social psychologist. *Soc. Personal. Psychol.*, *8*, 328–341. doi: 10.1111/spc3.12110

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

Vollmeyer, R., & Rheinberg, F. (2006). Motivational effects on self-regulated learning with different tasks. *Educational Psychology Review*, *18*(3), 239–253. doi: 10.1007/s10648-006-9017-0

Wagster, J., Tan, J., Wu, Y., Biwas, G., & Schwartz, D. (2007). Do learning by teaching environments with metacognitive support help students develop better learning behaviors? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 29).

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 1–40.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of experimental psychology: learning, memory, and cognition*, *15*(6), 1047.

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior*, *112*, 106457.

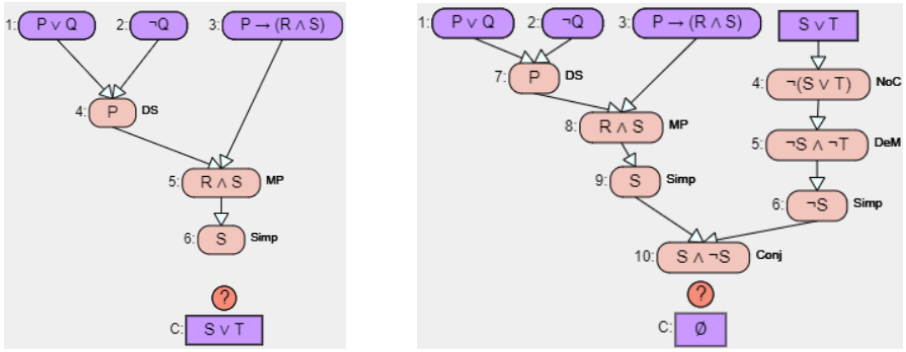
Winne, P. H., & Azevedo, R. (2014). Metacognition. In *The cambridge handbook of the learning sciences* (pp. 63–87).

Zepeda, C. D., Richey, J. E., Ronevich, P., & Nokes-Malach, T. (2015). Direct instruction of metacognition benefits adolescent science learning, transfer, and motivation: An in vivo study. *Journal of Educational Psychology*, *107*(4), 954. doi: 10.1037/edu0000022

Zhou, G., Azizsoltani, H., Ausin, M. S., Barnes, T., & Chi, M. (2022). Leveraging granularity: Hierarchical reinforcement learning for pedagogical policy induction. *International journal of artificial intelligence in education*, *32*(2), 454–500.

Zhou, M., & Winne, P. H. (2012). Modeling academic achievement by self-reported versus traced goal orientation. *Learning and Instruction*, *22*(6), 413–419. doi: 10.1016/j.learninstruc.2012.03.004

Zimmerman, B. J. (2011). Motivational sources and outcomes of self-regulated learning and performance. *Handbook of Self-Regulation of Learning and Performance*, 49. doi: 10.4324/9780203839010



(a) Forward Chaining (b) Backward Chaining

Fig. 1: Logic Tutor Problem-Solving Strategies

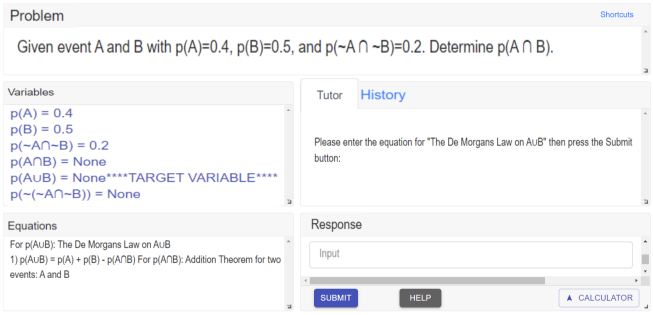


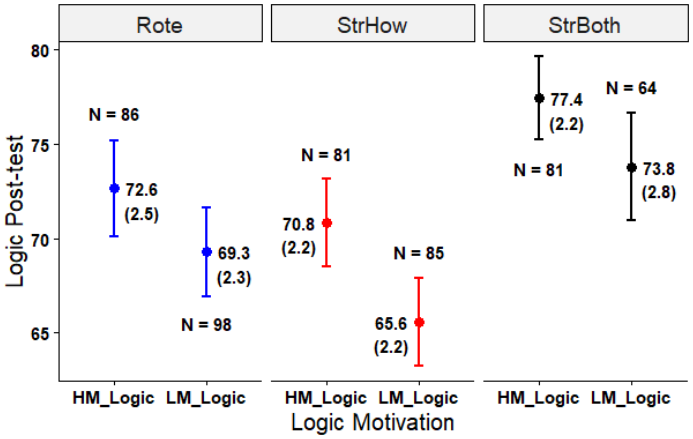
Fig. 2: Probability Tutor Training Interface

Table 1: Comparing the Metacognitive Groups (Prelim. Study)

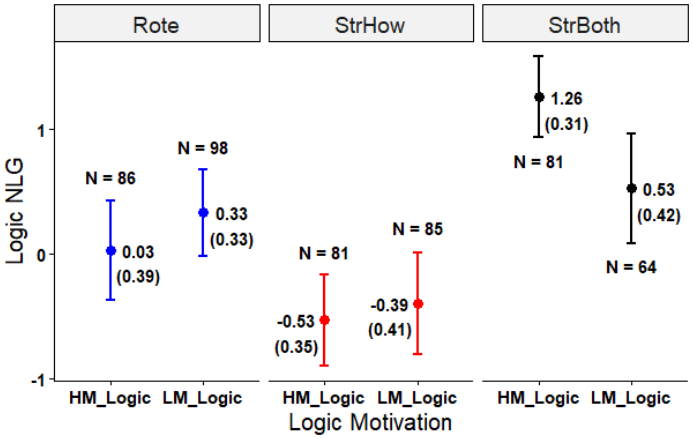
Group	*Pre*	*Iso. Post*	*Iso. NLG*	*Post*	*NLG*
	Logic Tutor				
Rote ($N = 184$)	75.5 (2.8)	69.8 (1.91)	0.14 (.31)	70.9 (1.68)	0.19 (.393)
StrHow ($N = 166$)	74.9 (3)	67.7 (1.96)	-0.49 (.42)	68.2 (1.67)	-0.46 (.39)
StrBoth ($N = 145$)	78.4 (3.2)	76.2 (1.9)	0.96 (.43)	75.8 (1.7)	0.94 (.395)
	Probability Tutor				
Rote	71.8 (2.6)	73.7 (2.8)	0.002 (.06)	73.4 (2.6)	-0.007 (.05)
StrHow	72.1 (2.5)	73.9 (3.2)	0.008 (.07)	74 (2.8)	0.01 (.05)
StrBoth	72.3 (2.8)	75.1 (3.4)	0.01 (.07)	75.5 (3)	0.02 (.06)

Table 2: Comparing the Motivation Level (Prelim. Study)

Group	*Pre*	*Iso. Post*	*Iso. NLG*	*Post*	*NLG*
	Logic Tutor				
HM_{Logic} ($N = 248$)	78.9 (5.3)	73.8 (1.6)	0.27 (.09)	73.6 (1.4)	0.25 (.06)
LM_{Logic} ($N = 247$)	73.4 (5.5)	70.2 (1.8)	0.17 (.1)	69.2 (1.4)	0.14 (.07)
	Probability Tutor				
HM_{Prob} ($N = 249$)	81.7 (4.2)	79.6 (2.3)	0.08 (.05)	79 (1.8)	0.05 (.04)
LM_{Prob} ($N = 246$)	77 (4.4)	68.7 (2.9)	-0.05 (.04)	69 (2.5)	-0.03 (.04)



(a) *Post*

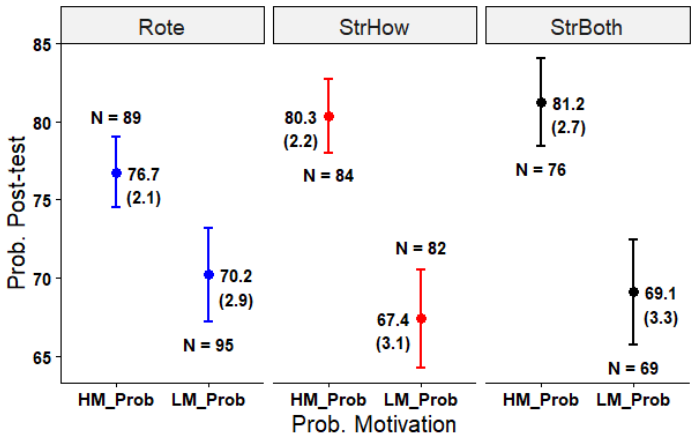


(b) *NLG*

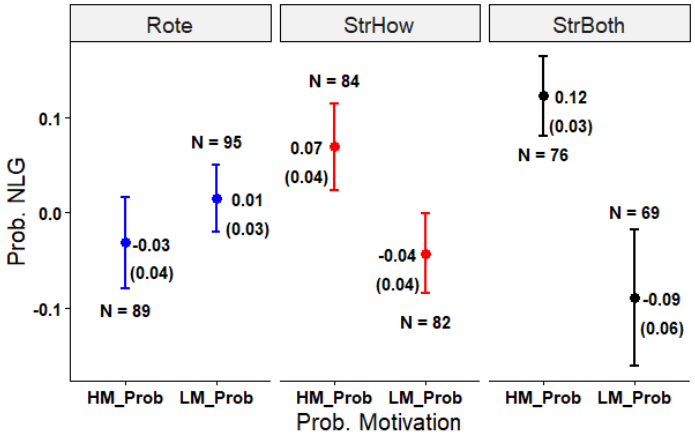
Fig. 3: Metacognitive Knowledge and Motivation on Logic (Prelim. Study)

Table 3: Confusion Matrix of Testing Dataset

Prediction / Truth	*Rote*	*StrHow*	*StrBoth*
Rote	45	1	0
StrHow	2	40	0
StrBoth	0	1	35



(a) *Post*



(b) *NLG*

Fig. 4: Metacognitive Knowledge and Motivation on Prob. (Prelim. Study)

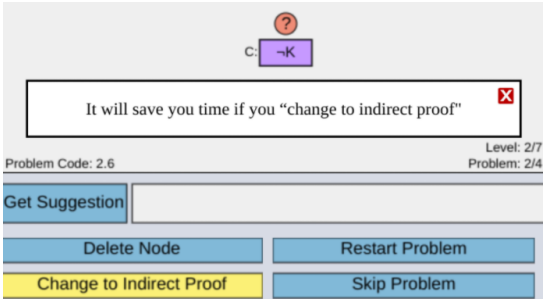


Fig. 5: Prompted Nudge

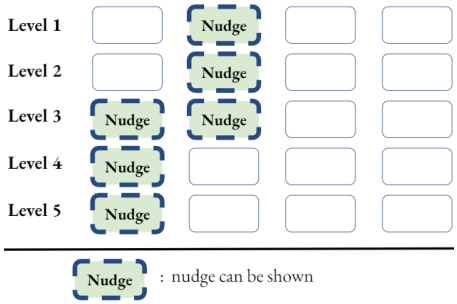


Fig. 6: Training on the Modified Logic Tutor (Exp.1: Nudge)

Table 4: Overview of Procedure (Exp.1: Nudge)

Logic	Pre-test (2 problems)
	Training (20 problems):
	Nudge (*Experimental*) $\Longrightarrow$ Intervention (Fig. 6)
	Control $\Longrightarrow$ Original
	Post-test (6 problems, including 2 isomorphic)
Six weeks later	
Prob.	Textbook
	Pre-test (14 problems)
	Training (12 problems)
	Post-test (20 problems, including 14 isomorphic)

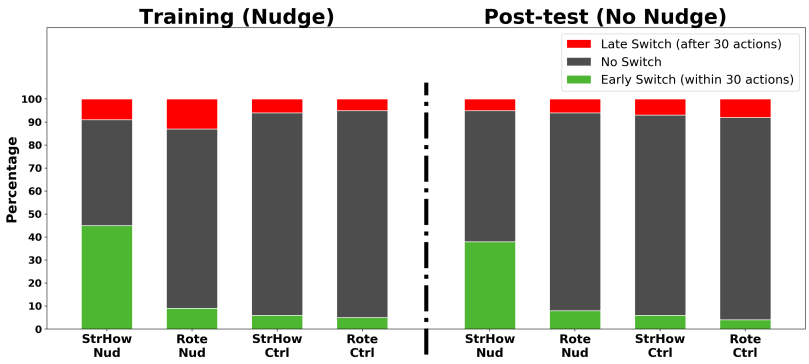


Fig. 7: Strategy Switch Behavior on Logic (Exp.1: Nudge)

Table 5: Comparing The Groups' Scores (Exp.1: Nudge)

	Nudge (*Experimental*)		*Control*	
	$Rote_{Nud}$ ($N = 13$)	$StrHow_{Nud}$ ($N = 15$)	$Rote_{Ctrl}$ ($N = 14$)	$StrHow_{Ctrl}$ ($N = 14$)
Logic Tutor				
Pre	66.5 (17)	65.6 (19)	64.1 (21)	63.9 (20)
Iso. Post	66.9 (5.6)	74.7 (6.1)	65.6 (5.2)	61.8 (4.7)
Iso. NLG	0.02 (.1)	0.12 (.07)	0.05 (.08)	-0.04 (.1)
Post	67.3 (5.1)	76.8 (4.9)	64.5 (4.3)	65.4 (4.2)
NLG	0.03 (.08)	0.18 (.09)	-0.01 (.16)	0.02 (.13)
Probability Tutor				
Pre	75.3 (13)	74.9 (15)	75.3 (17)	76.1 (14)
Iso. Post	81.1 (7.1)	92.9 (5.7)	79.1 (6.5)	80.3 (6.8)
Iso. NLG	0.1 (.18)	0.33 (.18)	0.07 (.17)	0.09 (.2)
Post	79.7 (5.9)	90.5 (5.4)	76.2 (6.1)	78.5 (5.6)
NLG	0.07 (.16)	0.29 (.08)	0.04 (.13)	0.05 (.15)

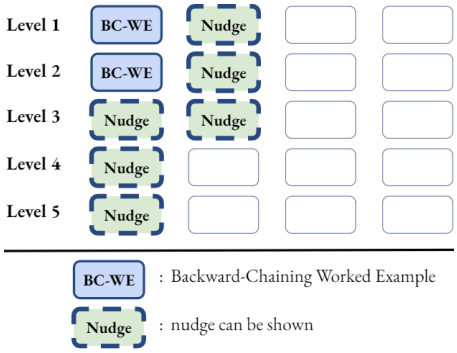


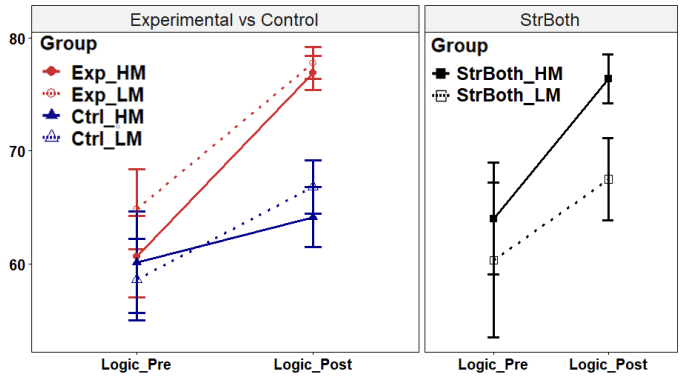
Fig. 8: Training on the Modified Logic Tutor (Exp.2: Instruction)

Table 6: Comparing the Conditions across Tutors (Exp.2: Instruction)

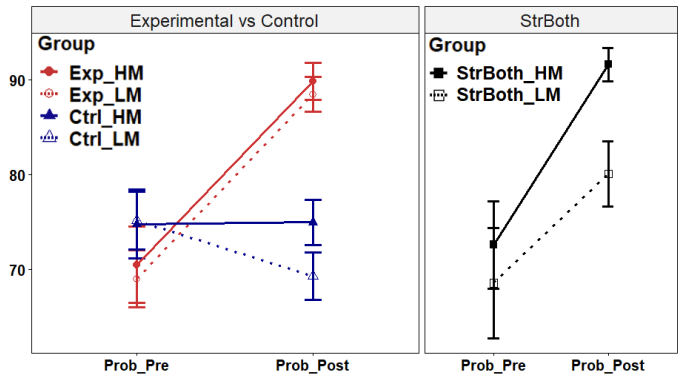
	Experimental ($N = 61$)	*Control* ($N = 41$)	1-way ANOVA
		Logic Tutor	
Pre	62.8 (19.6)	59.4 (18.4)	$p = .38$
Iso. Post	75.9 (2.5)	62.4 (4)	$F(1, 100) = 15.3$, $p <$ **.001**, $\eta^2 = .13$
Iso. NLG	0.19 (.03)	0.03 (.06)	$F(1, 100) = 8.2$, $p =$ **.005**, $\eta^2 = .08$
Post	77.4 (3.6)	65.4 (5.2)	$F(1, 100) = 38.9$, $p <$ **.001**, $\eta^2 = .28$
NLG	0.2 (.05)	0.05 (.06)	$F(1, 100) = 10.6$, $p =$ **.002**, $\eta^2 = .1$
		Probability Tutor	
Pre	69.6 (18.6)	74.9 (15.1)	$p = .13$
Iso. Post	92.9 (2.8)	84.4 (4.1)	$F(1, 100) = 18.5$, $p <$ **.001**, $\eta^2 = .16$
Iso. NLG	0.4 (.05)	0.11 (.08)	$F(1, 100) = 24.3$, $p <$ **.001**, $\eta^2 = .2$
Post	88.9 (5)	72.2 (6.1)	$F(1, 100) = 58.2$, $p <$ **.001**, $\eta^2 = .37$
NLG	0.33 (.06)	-0.18 (.21)	$F(1, 100) = 51.4$, $p <$ **.001**, $\eta^2 = .34$

Table 7: Comparing The Groups' Scores (Exp.2: Instruction)

	Instruction (*Experimental*)		*Control*		
	$Rote_{Ins}$ ($N = 35$)	$StrHow_{Ins}$ ($N = 26$)	$Rote_{Ctrl}$ ($N = 25$)	$StrHow_{Ctrl}$ ($N = 16$)	$StrBoth$ ($N = 26$)
			Logic Tutor		
Pre	61.8 (23)	64.2 (14)	60.1 (20)	58.3 (16)	62.3 (21)
Iso. Post	78.9 (1.9)	71.9 (1.6)	64.2 (3.8)	59.4 (4.2)	73.1 (5.3)
Iso. NLG	0.25 (.04)	0.1 (.04)	0.05 (.05)	-0.02 (.06)	0.08 (.05)
Post	80.3 (1.7)	73.4 (1.5)	64.3 (3.5)	67.2 (2.9)	72.3 (5.5)
NLG	0.25 (.03)	0.13 (.03)	0.02 (.04)	0.09 (.07)	0.11 (.06)
			Probability Tutor		
Pre	67 (20)	73.1 (16)	73.2 (15)	77.7 (15)	70.6 (19)
Iso. Post	92.5 (3.4)	93.5 (3.3)	82.5 (3.9)	87.4 (5.8)	91.7 (6.2)
Iso. NLG	0.43 (.06)	0.37 (.12)	0.09 (.21)	0.14 (.23)	0.37 (.16)
Post	88 (3.1)	90.2 (3.1)	71.3 (3.5)	73.5 (5.5)	85.8 (5.7)
NLG	0.35 (.05)	0.3 (.08)	-0.16 (.23)	-0.21 (.21)	0.24 (.15)



(a) On Logic



(b) On Probability

Fig. 9: The performance of Motivation Groups (Exp.2: Instruction)

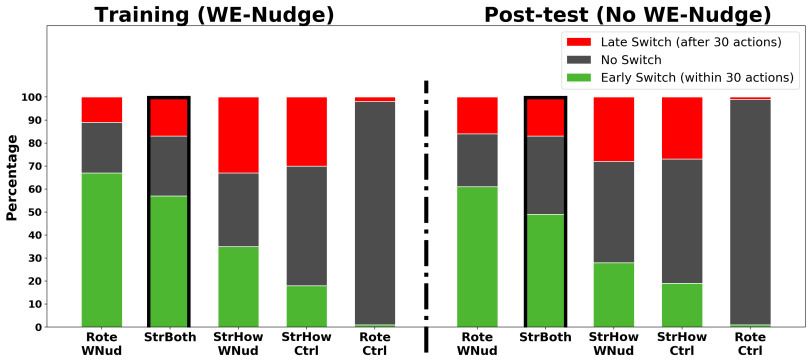


Fig. 10: Strategy Switch Behavior on Logic (Exp.2: Instruction)

Table 8: Comparing The Groups' Scores (Post-hoc Analysis)

	Group	*Pre*	*Iso. Post*	*Iso. NLG*	*Post*	*NLG*
		Logic Tutor				
Exp.1: Nud.	$Rote_{Nud}$ ($N = 13$)	66.5 (17)	66.9 (5.6)	0.02 (.1)	67.3 (5.1)	0.03 (.08)
	$StrHow_{Nud}$ ($N = 15$)	65.6 (19)	74.7 (6.1)	0.12 (.07)	76.8 (4.9)	0.18 (.09)
Exp.2: Ins.	$Rote_{Ins}$ ($N = 35$)	61.8 (23)	78.9 (1.9)	0.25 (.04)	80.3 (1.7)	0.25 (.03)
	$StrHow_{Ins}$ ($N = 26$)	64.2 (14)	71.9 (1.6)	0.1 (.04)	73.4 (1.5)	0.13 (.03)
Control	$Rote_{Ctrl}$ ($N = 39$)	61.5 (20)	64.7 (4.2)	0.05 (.06)	64.4 (3.7)	0.01 (.07)
	$StrHow_{Ctrl}$ ($N = 30$)	60.9 (17)	60.5 (4.4)	-0.03 (.07)	66.4 (3.7)	0.06 (.1)
	StrBoth ($N = 34$)	62.7 (20)	73.6 (5.1)	0.1 (.05)	72.9 (5.2)	0.14 (.05)
		Probability Tutor				
Exp.1: Nud.	$Rote_{Nud}$	75.3 (13)	81.1 (7.1)	0.1 (.18)	79.7 (5.9)	0.07 (.16)
	$StrHow_{Nud}$	74.9 (15)	92.9 (5.7)	0.33 (.18)	90.5 (5.4)	0.29 (.08)
Exp.2: Ins.	$Rote_{Ins}$	67 (20)	92.5 (3.4)	0.43 (.06)	88 (3.1)	0.35 (.05)
	$StrHow_{Ins}$	73.1 (16)	93.5 (3.3)	0.37 (.12)	90.2 (3.1)	0.3 (.08)
Control	$Rote_{Ctrl}$	74 (15)	81.3 (4.8)	0.08 (.2)	73.1 (4.9)	-0.09 (.19)
	$StrHow_{Ctrl}$	77 (15)	84.1 (6.2)	0.12 (.22)	75.8 (5.5)	-0.09 (.18)
	StrBoth	70.4 (20)	92.1 (5.9)	0.38 (.14)	86.4 (5.7)	0.27 (.15)

Table 9: ***MMI*** Framework for Transfer

	Metacognitive Knowledge	Rote (Neither)	Procedural (How)	Conditional (How & When)
Motivation		No	No	Yes
Intervention	Nudges (When)	No	Yes	-
	WE & Nudges (How & When)	Yes	Yes	-

Red/Green cells answer this: "Would possessing high motivation or receiving interventions facilitate transfer for a metacognitive knowledge group?"