1-13-2020

The curious case of loops

Briana B. Morrison

Lauren E. Margulieux

Adrienne Decker

The curious case of loops

Briana B. Morrisonhttps://orcid.org/0000-0003-4260-4278, Computer Science Department, University of Nebraska Omaha, Omaha, NE, USA

Lauren E. Margulieuxhttps://orcid.org/0000-0002-8800-2398, Department of Learning Sciences, Georgia State University, Atlanta, GA, USA

and

Adrienne Decker, Department of Engineering Education, University at Buffalo, Buffalo, NY, USA

ABSTRACT

BACKGROUND AND CONTEXT

Subgoal labeled worked examples have been extensively researched, but the research has been reported piecemeal. This paper aggregates data from three studies, including data previously unreported, to holistically examine the effect of subgoal labeled worked examples across three student populations and across different instructional designs.

OBJECTIVE

By aggregating the data, we provide more statistical power for somewhat surprising yet replicable results. We discuss which results generalize across populations, focusing on a stable effect size for subgoal labels in programming instruction.

METHOD

We use descriptive and inferential statistics to examine the data collected from different student populations and different classroom instructional designs. We concentrate on the effect size across samples of the intervention for generalization.

FINDINGS

Students using two variations of subgoal labeled instructional materials perform better than the others: the group that was given the subgoal labels with farther transfer between worked examples and practice problems and the group that constructed their own subgoal labels with nearer transfer between worked examples and practice problems.

ARTICLE HISTORY

Received 8 May 2019
Accepted 18 December 2019

KEYWORDS

Worked example, subgoal label, experiment, CS1

Introduction

Subgoal-labeled worked examples have been effective for teaching computing concepts, but the research to date has been reported in a piecemeal fashion. Pieces of three experiments using subgoal labeled worked examples for learning loop constructs have been reported in various conference proceedings (Morrision, Marguliuex, & Guzdial, 2015; Margulieux, Morrison, Catrambone, & Guzdial, 2016; Morrison, Decker & Marguliuex, 2016). The current paper aggregates these pieces and reports on new data from the experiments to examine more holistically the effect of subgoal labeled worked examples across three student populations and across different instructional designs. The different instructional designs include the first instance of testing student-generated subgoal labels and the first instance of testing differing amounts of transfer between worked examples and practice problems, in any discipline. By aggregating data from all three studies, including data that has not been reported before, we provide more statistical and explanatory power for somewhat surprising yet replicable results. We discuss which results generalize across populations, focusing on a stable effect size to be expected when using subgoal labels in programming instruction.

Literature review/background

This section reviews the current literature for subgoal learning along with some background in cognitive load theory to allow for framing the studies. We first present a common instructional design tool, *worked examples*, before presenting *cognitive load theory*, as the examples given to illustrate cognitive load involve worked examples. We then focus on subgoal label research (in worked examples) conducted within the computing discipline.

Worked examples

Worked examples are a type of instructional material used to teach procedural problem-solving processes. Worked examples give learners concrete examples of the procedure being used to solve a problem, showing the explicit steps in the problem-solving process. Eiriksdottir and Catrambone (2011) argue that learning primarily from worked examples may result in better initial performance as the worked examples are more easily mapped to the problems to be solved. They further posit, however, that learning from worked examples is less likely to result in retention and transfer of knowledge than learning from more abstract instructions. When studying worked examples, learners tend to focus on incidental features rather than the fundamental features of the problem. This occurs because the incidental features are easier to grasp for novices as they do not yet have the necessary domain knowledge to recognize the fundamental features of the worked examples (Chi, Bassok, Lewis, Reimann, & Glaser, 1989). For example, when studying physics worked examples, learners are more likely to recognize that the

example has a ramp than that the example uses Newton's second law (Chi et al., 1989). Therefore, while worked examples can improve initial performance, when learners focus on incidental features, they ineffectively organize and store information, leading to ineffective recall and transfer (Bransford, 2000).

Cognitive load

Cognitive load can be defined as the load imposed on an individual's working memory by a particular learning task (van Gog & Paas, 2012). The cognitive load imposed on the learner can directly affect knowledge retention and performance scores. Cognitive Load Theory (CLT) is grounded in the human architecture of the brain, which has a limited capacity for working memory. All the information is processed in working memory before being stored in long-term memory. If the total amount of processing required to learn exceeds the limited capacity of working memory, then learning is impaired (Plass, Moreno, & Brünken, 2010). Current thinking defines two different types of cognitive load on a student's working memory: intrinsic load and extraneous load (Kalyuga, 2011; Sweller, 2010; Sweller, van Merriënboer, & Paas, 1998; van Merriënboer & Sweller, 2005).

Intrinsic load is a combination of the innate difficulty of the material being learned combined with the learner's existing knowledge. For example, a conceptual understanding of a loop and the individual programming constructs to write a loop are intrinsic load for a problem that uses a loop. Extraneous cognitive load occurs when the learner is presented with information that does not directly contribute toward learning and is thus, extraneous. For example, while studying a worked example of a loop for calculating the average of a group of scores, the details of how a specific score is calculated are necessary for processing the worked example but not intrinsic to understanding how to solve a problem using a loop. Thus, the incidental details of worked examples are often extraneous. Working memory resources that are devoted to information that is relevant or germane to learning are referred to as *germane resources* (Sweller, Ayres, & Kalyuga, 2011).

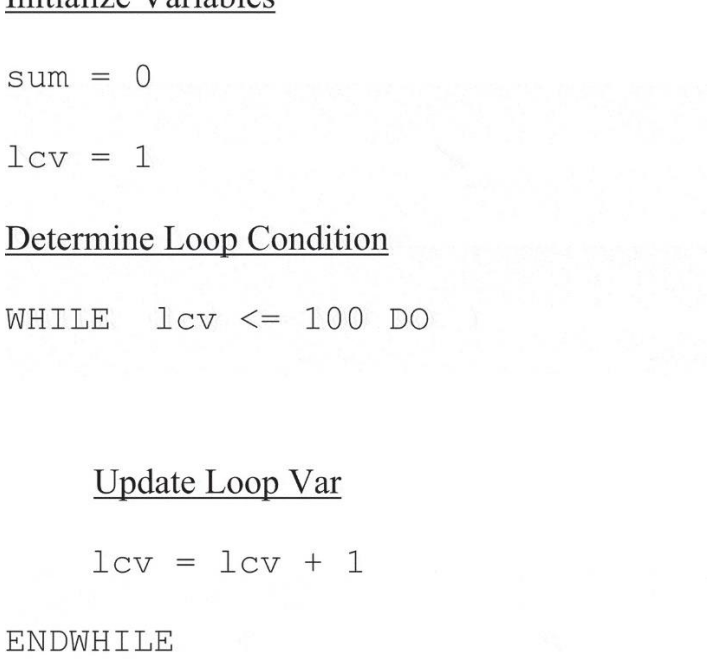
The intrinsic and extraneous loads may be moderated through the careful design of the instructional materials. The intrinsic load should be managed so that learners are not given too much new information to process at once. While some extraneous load is inevitable, instructional materials should attempt to eliminate unnecessary extraneous load. Worked examples, when carefully designed, can accomplish both of these goals (Sweller et al., 1998).

Subgoal labels

To guide learners' attention away from incidental details and promote deeper processing of worked examples for improved recall and transfer, the subgoal learning framework can be used to design worked examples that emphasize problem-solving structure. The subgoal learning framework is a strategy used predominantly in STEM fields to help students deconstruct problem-solving procedures into subgoals, or the functional parts of the problem-solving procedure, to better recognize the fundamental components of the problem-solving process (Atkinson, Catrambone, & Merrill, 2003). Subgoals can be thought of as the building blocks of procedural problem solving and they exist for all problem-solving procedures except the simplest ones.

Subgoal labeling is a specific technique used to promote subgoal learning. It has been used to help learners recognize the fundamental structure of the problem-solving procedure being illustrated in a worked example (Catrambone, 1994, 1996, 1998). Subgoal labels are function-based instructional phrases that explain to the learner the purpose of that step, or subgoal, in the problem-solving process. In Figure 1, the first two lines of code have the subgoal label "Initialize Variables." This label provides information about the purpose of that subgoal and the function behind the steps within it. Studies (Atkinson, 2002; Atkinson & Derry, 2000; Catrambone, 1994, 1996, 1998; Margulieux & Catrambone, 2014; Margulieux, Guzdial, & Catrambone, 2012) have consistently found that subgoal-oriented instructions improved problem-solving performance across a variety of STEM domains, such as programming (e.g. (Margulieux et al., 2012)) and statistics (e.g. (Catrambone, 1998)).

Figure 1. Partial worked example illustrating subgoal labels. Subgoal labels are underlined.



Giving subgoal labels in worked examples improves learner performance while solving novel problems without increasing the amount of time learners spend studying instructions or working on problems (Margulieux et al., 2012). From a cognitive perspective, it is thought that subgoal labels are effective because they visually group the problem-solving steps within the worked examples into subgoals and give meaningful labels to the groups (Atkinson et al., 2003). This subgoal-labeled format highlights the structure of the examples, helping students to focus on the structural features of the problem and allows the learner to more effectively organize the information (Atkinson, Derry, Renkl, & Wortham, 2000). Because learners are more focused on the structural features of the worked example allowing more effective organization of the information, subgoal labels may reduce the extraneous cognitive load that can hinder learning but is inherent in worked examples (Renkl & Atkinson, 2002).

Subgoal labels that are context-independent are the most effective type of subgoal labels (Catrambone, 1995, 1998). Catrambone found that learners who were given abstract labels (e.g. Ω) and had sufficient prior knowledge performed better than those who were given context-specific labels (e.g. initialize accumulation loop variables) on problem-solving tasks done after a week-long delay or in problems that required using the problem-solving procedure differently than demonstrated in the examples (Catrambone, 1998). Catrambone explained this finding by arguing that learners with sufficient prior knowledge could correctly explain to themselves the purpose of the abstract subgoal and that they presumably had to self-explain due to the abstract nature of the label. He argued that the self-explanation was more effective than providing context-specific labels.

Self-explanation

A common and effective type of constructive learning that might help learners understand subgoals is self-explanation. Self-explanation is a learning strategy in which students use prior knowledge and logical reasoning to make sense of new information and gain knowledge. A review of self-explanation studies found it is effective across a range of domains if the domain has logical rules with few exceptions (Wylie & Chi, 2014).

Self-explanation of a worked example's solution identifies structural features and reasons for the function of the problem-solving steps (Bielaczyc, Pirolli, & Brown, 1995). The purpose of self-explanation is similar to that of subgoal learning. By self-explaining worked examples, learners are more likely to recognize structural versus superficial features. However, learners do not often engage in self-explanation without explicit

prompting. Many studies (e.g. Chi et al., 1989) found that 10% or less of learners self-explained examples without external prompting. Most of the time learners can self-explain if they devote additional resources to the task (Wylie & Chi, 2014) and if they are reminded and guided to do so. Research has found little difference in the learning outcomes of students who self-explain on their own or are prompted to self-explain (e.g. Bielaczyc et al., 1995). This suggests that self-explanation itself is the cause of learning benefits.

Parsons problems

Before describing how subgoals have been used in computer science education, we should explain a type of assessment used in this research, Parsons problems. When learning programming, students must learn a new language – the programming language used to communicate instructions to the computing agent – with its own unique syntax. This level of intrinsic cognitive load can overwhelm the learner, so researchers have sought ways to eliminate or reduce the learning of programming language syntax (Resnick et al., 2009). For text-based programming languages, one way to assess student knowledge without requiring syntax knowledge is to use Parsons problems (Parsons & Haden, 2006). In Parsons problems, the correct code is broken into code fragments that students then put into the correct order with the correct indentation. Parsons problems require a lower cognitive load on the learner because the search space is limited to only the code fragments in the problem and there is no possibility of syntax errors. Using Parsons problems for assessment of student knowledge allows students without syntax knowledge of the programming language to demonstrate procedural problem-solving knowledge.

Subgoals in computer science

Subgoal learning was first applied to programming education in the context of an experimental laboratory with psychology undergrads as participants. Due to this context, the programming procedure being taught had to be accessible to absolute novices. Thus, participants were taught to create apps in Android App Inventor. In this highly controlled environment, subgoal-labeled worked examples were found to improve problem-solving performance by 8% (Margulieux et al., 2012). From that experiment, research has focused on testing subgoal labeled worked examples in more authentic programming education environments, including online learning with K-12 teachers (Margulieux, Catrambone, & Guzdial, 2016), a game-based K-3 setting (Joentausta & Hellas, 2018), and in open educational resources that crowdsource subgoal labels (Kim, Miller, & Gajos, 2013). Our research applies subgoal learning to an introductory

programming course, specifically to students who were learning to solve problems using while loops.

Our first study (*Study 1*) (Morrison et al., 2015) tested hypotheses related to whether using subgoal labels to teach while loops would produce results similar to those achieved in other disciplines. Learning to use while loops is cognitively demanding, and the study proposed that using subgoal labels to help students learn would reduce the cognitive load imposed during learning. Because students were several weeks into an introductory programming course, we also recognized that they would have some prior knowledge that was relevant to solving the loop problems. For this reason, we hypothesized that students might better learn the subgoals of the procedure if they were prompted to self-explain the subgoals, rather than being given subgoal labels that were already defined. Self-explaining the subgoals, if students were able to do it, would encourage active learning of the subgoals and lead to deeper learning than viewing existing subgoal labels, which would lead to passive learning.

To test this hypothesis, the study divided the participants into three treatment groups, each with its own instructional materials: learning with no subgoal labels (*No Subgoal*), learning with given pre-defined subgoal labels (*Given*), and asking participants to generate their own subgoal labels after some initial training (*Generate*). Each treatment group was then subdivided into two sections: isomorphic (near) or contextual (far) transfer between worked examples and practice problems (see Method-Design for more information on transfer). Like self-explaining subgoals, contextual (far) transfer between worked examples and practice problems was expected to promote deeper learning and improve later problem-solving performance, if students could successfully engage in it. The contextual transfer was also expected to be highly cognitively demanding and perhaps unachievable for many students.

This first study found that students who learned with subgoal labels (either given or generated) performed better on the code-writing assessments than those who learned without subgoal labels. Within the given and generated groups, the best performing group depended on the type of transfer between worked example and practice problems that they received.

The unexpected results occurred with the given subgoal label group. Cognitive Load Theory predicts that learning with given subgoal labels and no contextual transfer should impose lower cognitive processing than learning with given subgoal labels and contextual transfer and thus result in better learning. The contextual transfer would require additional working memory to process, reducing learning. However, the results from the first study directly contradict this prediction. *Study 1* found, unlike the other two treatment groups, that participants who learned with given subgoal labels and contextual transfer significantly outperformed the given subgoal labels with isomorphic

problems, completely opposite from what Cognitive Load Theory predicts. We examined whether this main finding was an anomaly or if it could be replicated.

In a follow-up paper (*Study 1 follow up)* (Morrison, Margulieux, et al., 2016), we examined the performance of students on a Parsons problem assessment, after having learned loop problem-solving in one of the treatment groups (with no subgoal labels, with given subgoal labels, or generating their own subgoal labels). We found that students who were given subgoals performed statistically significantly better than those who had no subgoals or who generated their own subgoals, regardless of transfer condition. Participants that were given subgoal labels performed overall better than those that did not have subgoal labels and those that generated their own subgoal labels. Though participants in the generate labels and no labels conditions performed equally, participants who generated their own labels completed the task faster than those who did not receive labels.

In *Study 2* (Margulieux, Morrison, Catrambone, & Guzdial, 2016), the examination of the quality of the learner-generated labels from a new population of students and how this affected problem-solving performance was reported. *Study 2* found that twice as many participants generated specific labels than general labels, but a larger percentage of participants who received contextual transfer generated general labels than those who had an isomorphic transfer. Participants who learned with isomorphic transfer and generated their own labels performed relatively well, regardless of the specificity of their labels. For those that learned with contextual transfer, their performance depended on whether they created specific or general labels. Those who created specific labels performed as poorly as the worst-performing group, those who received no subgoal labels with the contextual transfer. On the other hand, participants who created general labels with contextual transfer performed better than any other group.

Study 3 (Morrison et al., 2016) paper replicated *Study 1* (Morrison et al., 2015) with a third population of students. The results supported the findings from the previous studies: participants who learn by generating subgoal labels (using isomorphic worked example – practice problem pairs) performed the best, and statistically better than if they had been worked example – practice problem pairs with the contextual transfer. Despite the previous publications that report results of each of the three experiments individually, we have yet to report all of the data from these three experiments or examine them holistically to determine the cross-population effects of the subgoal labeled worked examples. This paper addresses this gap.

Present study

In this paper, we examine the effect of learning subgoals through different instructional methods (i.e. given labels versus generated labels compared to unlabeled) and transfer

distance between worked examples and practice problems (i.e. isomorphic or contextual transfer) across three separate, but comparable, populations. This new analysis of the data allows us to report findings that were excluded from previous conference proceedings and explore the average effect of the interventions to determine a stable effect size across populations. We have the following research questions:

RQ1*: How do different instructional methods of learning with subgoals (either given or learner generated) affect problem-solving performance?*

RQ2*: How does transfer distance (i.e. isomorphic transfer (changing the values in a problem with the same context) or contextual transfer (changing the context, or cover story)) from worked examples to paired practice problems affect problem-solving performance?*

To measure performance, we used three different assessments: (1) four novel coding writing problems, (2) one Parsons problem, and (3) a post-test of five multiple-choice questions, none of which contained the subgoal labels. The three assessments were chosen to represent three levels of difficulty and application of knowledge. Code writing was intended to be the most difficult and required students to recall the problem-solving process from memory. The Parsons problem was intended to assess knowledge of the problem-solving process while allowing students to recognize, rather than recall, the procedure. Furthermore, students do not have to determine how to apply a conceptual understanding to a new context in Parsons problems because the lines of code are provided for them. Therefore, increasing the transfer distance between worked examples and practice problems might not necessarily improve Parsons problem performance, though it was expected to improve code-writing performance. The multiple-choice questions required students to trace the code and determine which answers containing possible outputs were correct. These questions were intended to be the easiest questions and a learning check to identify participants who were not engaging in the instruction. Additionally, we measured cognitive load related to the instructional materials using the (Morrison, Dorn, & Guzdial, 2014) instrument and time on task for both the learning period and each assessment.

Method

Design

The experiment had two manipulations: the format of worked examples and the transfer distance between worked examples and practice problems. The worked example either

had no subgoal labels (i.e. *No Subgoal*), had subgoal labels created by experts (i.e. *Given*), or included a placeholder for the participant to fill in their own subgoal label (i.e. *Generated*). In the *No Subgoal* condition (Control Group A in the Supplementary), the worked example is presented in a step by step solution of how to develop the code solution for the problem, including code comments. In the *Given* condition (Subgoal Given, group B in the Supplementary), the worked example is the same but broken into groups and labeled by the subgoal associated with the task. One subgoal may include more than one step. Code comments were identical to the control condition. For the *Generated* condition (Subgoal Generate group C in the Supplementary), the worked example was broken into groups, as in the Given condition, but instead of including the expert-created subgoal label, a blank space was included to allow the participant to type in their own subgoal explaining what the pieces of code accomplished.

The second manipulation involved the differences between the worked example and practice problem given to the students. As can be seen in the Supplementary (worked examples compared to practice problems), for the isomorphic (near) transfer problems the context of the problem for the worked example and the practice problem is identical, and only the values being manipulated change. For the contextual (far) transfer problems, the context of the worked example and the practice problem are different; however, the solution has an identical format. The experiment measured performance with pre- and post-tests, problem-solving tasks (both writing code and completing a 13-step Parsons problem), self-reported cognitive load on the (Morrison et al., 2014) instrument, and time on task.

Instructional materials

In this study, we developed instructional materials to teach introductory programming students to solve programming problems using while loops. We selected the topic of writing indefinite loops for several reasons: (1) based on experience we know that students can struggle with the introduction of repetition statements, (2) while loops are the most general form of a repetition control structure allowing any type of loop to be written, and (3) teaching of this topic occurs in the early part of the term allowing us to reach the maximum number of students – typically before the withdrawal date for the term passed.

The materials used pseudocode so that students from multiple universities and courses that used different programming languages could participate. Pseudocode is easy for students to understand regardless of the programming language that they are learning (Tew & Guzdial, 2011). The first two experiments started before students had learned to use while loops in their courses, and the third experiment was conducted after students had been introduced to while loops. The procedure took about two hours

to complete. In most cases, the experimenters conducted the experiment in a regularly scheduled lab for the programming courses from which they recruited participants. The labs were held in closed classrooms with at least one computer per student. Some participants completed the procedure as an at-home assignment.

The instructional materials were three separate worked examples interleaved with a practice problem after each worked example. The format of the worked examples can be seen in the Supplementary (Worked Examples). Each worked example appeared on one screen, followed by the practice problem on the next screen. Students could go back and forth between the worked example and the practice problem during the instructional period. Once the student reached the assessment portion of the study, they could not go back to the instructional materials.

At the beginning of the session, the experimenters introduced the study explaining that they would learn to solve problems using while loops and that the materials they received would help them to achieve this. The experimenters then gave students a link to a SurveyMonkey survey where all of the materials and assessments were hosted. Participants worked independently and could ask for help from the experimenter on administrative tasks (e.g. “What is my participant number?”) but not for help on the programming tasks (e.g. “How do I increase the loop control variable?”). Because students worked independently, some completed the tasks faster than others, and SurveyMonkey recorded how quickly each student progressed through the various stages of the experiment.

Assessments

After completing the instructional period with worked examples and practice problems, participants were asked to solve four novel problems using while loops. All the assessment problems required a contextual transfer from the worked examples and practice problems that participants used to learn the procedure. No subgoal labels appeared in any of the assessment problems. We scored participants’ problem-solving solutions to create a problem-solving score. We evaluated the solutions line-by-line rather than as a whole to provide more sensitivity in the score. Each correct line of code earned one point for a maximum score of 44 points across four questions. Lines of code were considered correct if they were conceptually correct, regardless of typos or syntax errors. Logic errors (e.g. having < rather an ≤) made the line incorrect. We decided to score for conceptual and logical accuracy rather than absolute accuracy because the participants were inexperienced programmers.

We also measured participants’ problem-solving procedural knowledge with a Parsons problem. We scored participants’ Parsons problem answers for correct order. Because the Parsons problem had 13 pieces of code to rank order, the maximum score

was 13. Participants earned one point for each piece of code that was in the correct order relative to the piece before it. For example, if a participant's solution ranked the 6th, 7th, and 8th pieces of code in the 7th, 8th, and 9th positions, they would lose only the first point because it did not follow the 5th piece of code. The 7th and 8th pieces would still be in order, relative to other pieces of code, and counted as correct. This scoring scheme was considered better than scoring for exact order because it does not penalize later pieces of code for earlier mistakes.

Procedure

Most participants completed the experiment during one of their lab sessions in a computer laboratory. Students had an option to complete an alternative assignment, but none selected that option. Participants worked independently, and each session included between 15 and 30 people. The sessions typically lasted between 1 and 1.5 h, depending on the rate at which participants completed the tasks. For students in the lab setting, a few stragglers were asked to leave at the end of 2 h due to the next class arriving.

First, participants completed a demographic questionnaire and the pre-test. Next, they began the instructional period. The instructional period began with training. Participants who were going to generate their own subgoal labels received training to create subgoal labels (see the Supplementary – How to Make Subgoal Labels). The training included instructions about creating subgoal labels, examples of a subgoal labeled worked example, and activities to practice creating subgoal labels on simple algebra problems designed to be easy for any college student so that they could focus on creating labels. Participants who did not generate their own subgoal labels received training to complete verbal analogies (available in the Supplementary – Verbal Analogies). Verbal analogies (e.g. water: thirst: food: hunger) were considered a comparable task to subgoal label training because they both require analyzing text to determine an underlying structure. Like the subgoal label training, the analogy training included instructions, worked examples, and activities to practice.

Following the training, the instructional period provided worked examples and practice problems to help participants learn to use while loops to solve problems. Once participants completed the instructional period, they started the assessment period. Throughout the procedure, the time taken to complete each task was recorded. A diagram of the entire study procedure can be seen in Figure 2.

Figure 2. Complete study procedure. Items with * are provided in the Supplementary.

	No Subgoal Labels		Subgoal Labels Given		Subgoal Labels Generated	
	Consent					
	Demographics*					
	Pre test*					
	Training Problem – summing					
	Analogy Training & Activity*				Subgoal Training & Activity*	
Groups	**No Subgoal Isomorphic Transfer**	**No Subgoal Contextual Transfer**	**Given Isomorphic Transfer**	**Given Contextual Transfer**	**Generate Isomorphic Transfer**	**Generate Contextual Transfer**
Worked Example 1 - Calculate Average Tip	(no subgoal labels)*		(subgoal labels given)*		(space to generate subgoal labels)*	
Problem 1 - Calculate Average	P1 - Tip* (no subgoals)	P1A - Rainfall* (no subgoals)	P1 - Tip* (given)	P1A - Rainfall* (given)	P1 - Tip* (generate)	P1A - Rainfall* (generate)
Worked Example 2 - Count Matching Values (dice - 7)	(no subgoal labels)*		(subgoal labels given)		(space to generate subgoal labels)	
Problem 2 - Count Matching Values	P2 - Dice-2* (no subgoals)	P2A - < 3 (no subgoals)*	P2 - Dice-2 (given)	P2A - < 3 (given)	P2 - Dice - 2 (generate)	P2A - < 3 (generate)
Worked Example 3 - Count Prime Numbers (1-100)	(no subgoal labels)*		(subgoal labels given)		(space to generate subgoal labels)	
Problem 3 - Count	P3 - Primes (100-200) (no subgoals)	P3A - Unique Values (no subgoals)	P3 - Primes (100-200) (given)	P3A - Unique Values (given)	P3 - Primes (100-200) (generate)	P3A - Unique Values (generate)
	Cognitive Load Measurement					
	Problem Solving Assessment (4 problems; 2 near transfer, 2 far transfer)*					
	Parsons Problem Assessment *					
	Post Test *					

Participants

Participants across the three experiments were 220 students recruited through programming courses and offered course credit for completing a lab activity as compensation. To account for possible effects of prior experience, participants reported whether they had experience with programming and/or using loops during high school (AP courses or otherwise) and college. Other learner characteristics that participants provided were gender, age, academic major, high school grade point average (GPA), college GPA, whether English was their primary language, number of years in college, self-reported comfort with computers, expected difficulty of completing the programming task, and prior courses in programming. Participants were randomly assigned to intervention conditions to avoid possible confounds caused by learner characteristics. To ensure that there were no confounds, learner characteristics and problem-solving performance were correlated using Pearson's r for continuous learner variables and Spearman's ρ for dichotomous learner variables. The results of these analyses are reported in Table 1, 2, and 3.

Table 1. Learner characteristics in experiment 1 and their relationship to performance. (Table view)

Learner characteristic	Mean/proportion	Std. deviation	Correlation with problem-solving performance
Gender	84% male	-	$\rho = -.02$, $p = .90$
Age	21	4	$r = -.06.$, $p = .60$

Learner characteristic	Mean/proportion	Std. deviation	Correlation with problem-solving performance
Major	50% CS major	-	$\rho = -.03$, $p = .78$
High School GPA	3.40	0.58	$r = -.06$, $p = .59$
College GPA	3.08	0.55	$r = .18$, $p = .13$
English is primary language	91% yes	-	$\rho = .06$, $p = .61$
Years in college	2.4	1.4	$r = -.03$, $p = .81$
Comfort with computers*	4.2	1.5	$r = .46$*, $p < .001$
Expected difficulty of task*	4.0	1.4	$r = .29$*, $p = .007$
Prior course in programming	42% yes	-	$\rho = .37$*, $p < .001$

*The question about comfort with computers asked student to rate how comfortable they were using a computer on a 7-point scale that ranged from "1 – not comfortable at all" to "7 – very comfortable." The question about expected difficulty of task used a 7-point scale that ranged from "1 – very difficult" to "7 – very easy".

Table 2. Learner characteristics in experiment 2 and their relationship to performance. (Table view)

Learner characteristics	Mean/proportion	Std. deviation	Correlation with problem-solving performance
Gender	40% male	-	$\rho = .05$, $p = .79$
Age	22	6.9	$r = .05.$, $p = .79$
Major	23% CS major	-	$\rho = .01$, $p = .94$
High School GPA	3.81	0.37	$r = .43$*, $p = .02$
College GPA	3.33	0.53	$r = .15$, $p = .42$
English is primary language	91% yes	-	$\rho = .32$, $p = .06$
Years in college	3.3	1.8	$r = -.14$, $p = .42$
Comfort with computers*	3.5	1.2	$r = .07$, $p = .68$
Expected difficulty of task*	3.2	1.3	$r = .24$, $p = .17$
Prior course in programming	29% yes	-	$\rho = .11$, $p = .52$

Table 3. Learner characteristics in experiment 3 and their relationship to performance. (Table view)

Learner characteristics	Mean/proportion	Std. deviation	Correlation with problem-solving performance
Gender	71% male	-	$\rho = -.02$, $p = .90$
Age	19	3	$r = .08.$, $p = .46$
Major	33% New Media 63% Game Design	-	$\rho = .11$, $p = .30$

Learner characteristics	Mean/proportion	Std. deviation	Correlation with problem-solving performance
High School GPA	3.61	0.32	$r = .05$, $p = .70$
College GPA	3.47	0.62	$r = -.06$, $p = .64$
English is primary language	96% yes	-	$\rho = -.15$, $p = .15$
Years in college	1.9	0.8	$r = .06$, $p = .57$
Comfort with computers*	5.3	1.4	$r = .52$*, $p < .001$
Expected difficulty of task*	4.5	1.4	$r = .31$*, $p = .002$
Prior course in programming	94% yes	-	$\rho = .30$*, $p = .003$

In addition to asking students about their prior experiences with programming and using loops, participants completed a pre-test to measure their prior knowledge of solving problems using while loops. The pre-test included five multiple-choice questions from AP CS A exams. Participants who answered more than two questions on the pre-test correctly were excluded from the analysis to reduce potential error because the instructional materials were intended for novices. Participants who did not complete all components of the experiment were also excluded from the analysis. The numbers of students excluded were relatively low and detailed in the following sections.

Experiment 1 participants

Participants were 66 students from 1 of the 4 introductory programming courses at a technical university in the southeast United States. The experiment occurred before students learned about loops in their courses. Students performed poorly on the pre-test, $M = 1.2$ out of 5 points, and 32% of participants earned no points. Six students (out of 72, 8%) were excluded from analysis for high pre-test scores. No statistically significant relationships between all assessments and learner characteristics were found for most variables. Comfort with computers, expected difficulty of task, and taking a prior course, however, correlated with problem-solving performance. To ensure that no conditions had an advantage over the others based on these learner variables, we inspected the means for each of these learner variables within each condition. We found no meaningful differences (i.e. more than a few decimal points) among conditions.

Experiment 2 participants

Participants were 54 students from introductory programming courses at a different technical university in the southeast United States. Unlike in Experiment 1, only 23% of participants were computer science majors. The majority of students were taking a Computational Media course. Many of them were likely taking the course because the

university requires that all students take a programming course, and this course is designed specifically for students not majoring in computing. This sample characteristic explains the relatively high average age and the number of years in college for participants.

The average score on the pre-test was low, *M* = 1.6 out of 5, and 23% of students earned no points. Five students (out of 59, 8%) were excluded from analysis for high pre-test scores. The only learner variable that correlated with assessment scores was a high school GPA. The mean high school GPA for each experimental group was inspected to ensure that no groups had an advantage over the others. Each mean was within a few decimal points of the others.

Experiment 3 participants

The last site used to collect data was a technical university in the northeast United States. The final experiment had a larger number of participants than the first two, 100 students. The final experiment also included students from first-semester introductory programming courses, like the first two experiments, and students in a second-year course. Collecting data from both the first-semester and second-year course in the computing curriculum allowed us to explore how prior knowledge impacted the results because the students in the second-year course would have already learned, practiced, and been tested on solving problems with while loops in a previous course (*Study 3*) (Morrison et al., 2016).

In this experiment, both first-semester and second-semester students had already learned to use while loops. To account for prior knowledge, participants completed the same pre-test as Experiments 1 and 2. The average score was *M* = 2.3 out of 5. Participants were not excluded from analyses based on their pre-test scores, unlike in the previous two experiments. At the University for this study, students are not given credit for AP CS courses. This led to a large number of students in the first-semester course having prior programming knowledge. If we had excluded students based on their pre-test scores, there would not have been enough statistical power in the analyses. Additionally, this manuscript aggregates the effect of subgoal labels across different populations; having more knowledgeable students represents a unique population compared to the first two studies. As in Experiment 1, comfort with computers, expected difficulty of task, and taking a prior course correlated with problem-solving performance. We again inspected the means for each of these learner variables within each condition to ensure that no condition had an inherent advantage over the others. No meaningful differences were found among conditions.

Results

The data used for this paper have been partially reported in previous papers as independent experiments. The problem-solving, post-test, and time on task data for Experiment 1 were published in (*Study 1*) (Morrison et al., 2015). The Parsons problem data for Experiment 1 were published in (*Study 1 follow-up*) (Morrison, Margulieux, et al., 2016). For both Experiments 1 and 2, the problem-solving, Parsons problem, quality of generated labels, and time on task were published in (*Study 2*) (Margulieux et al., 2016). For Experiment 3, the problem-solving and time on task data were published in (*Study 3*) (Morrison et al., 2016). For some of the analyses reported in these papers, the differences among groups had meaningful effect sizes but were not statistically significant. By analyzing the data together, the sample size, and thus statistical power, will be large enough to produce reliable effect sizes and, if the differences are large enough, statistical significance.

In addition to adding statistical power to our analyses, this paper will include data that has not been reported before due to space constraints. The new data included in this analysis are cognitive load data for all three experiments, Parsons problem data from Experiment 3, and post-test data for Experiments 2 and 3.

For all dependent variables (i.e. problem-solving performance, post-test, Parsons problem, cognitive load, and time on task), we analyzed the distribution of scores for skewness and kurtosis to ensure normal distribution and, therefore, that parametric statistical tests, such as ANOVA, were appropriate. In addition, we visually inspected the histograms of scores for each measurement. In all cases, the skewness and kurtosis were within normal bounds (i.e. between −2 and 2 (Gravetter & Wallnau, 2016)) and histograms followed a normal distribution. Therefore, no outliers were excluded from analyses, and parametric tests are appropriate for analyses of the measurements.

Performance data

For our inferential statistics, we report two types of effect sizes. The first, est. ω^2, is for only omnibus analyses (i.e. ANOVAs) and describes how much of the variation in scores can be attributed to the manipulation (i.e. proportion of variance accounted for, PVAT). For example, for the problem-solving tasks, an est. ω^2 of .06 means that 6% of the variation in performance can be attributed to the instructional manipulations. In the social sciences, an est. ω^2 of .06 is considered a medium-sized effect (Cohen, 1969). The second effect size, *f* or *d*, was used for only our *post hoc* analyses to describe the difference between groups using the standard deviation as the unit of measurement. For example, for the problem-solving tasks, a *d* of .5 would mean that the difference between the means of two groups is half of the standard deviation for those groups. The statistic *d* is used for t-tests, and the statistic *f* is used for ANOVAs and is equal to 2*d* (Cohen, 1988). For example, an *f* of .25 is equal to a *d* of .5, and both indicate that

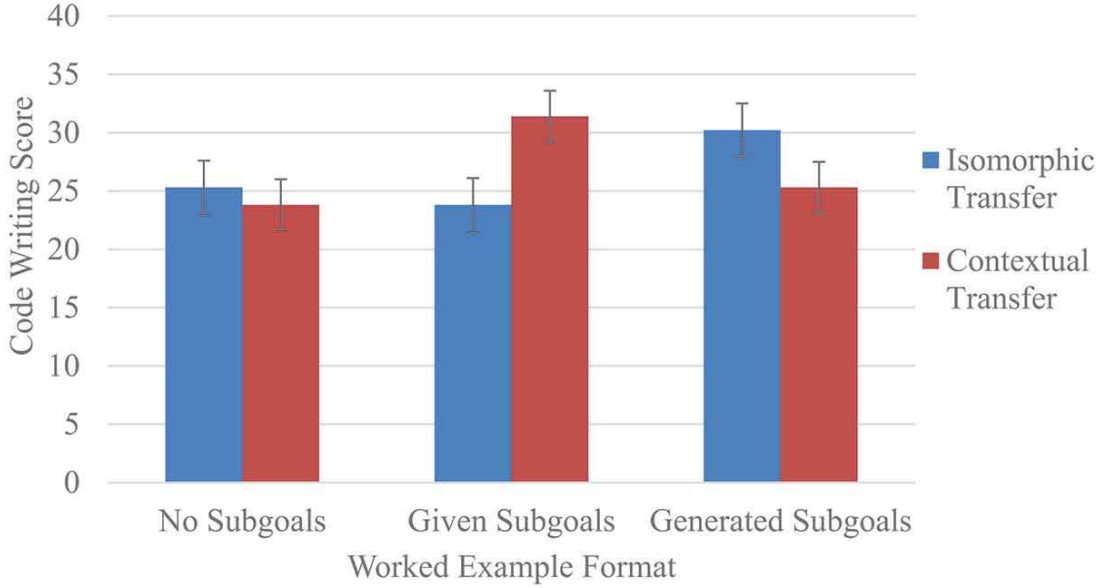
the difference between means is half of a standard deviation, which is considered a medium-sized effect (Cohen, 1969).

Problem-solving score

The main dependent variable, score on problem-solving tasks, had a maximum score of 44. The overall mean score was 26.58, and the standard deviation was 14.05. For the omnibus ANOVA analyses of these data, worked example format and transfer distance were treated as randomly assigned variables. In addition, university, which was different for each experiment, was treated at a quasi-experimental variable. This nested design allows us to combine the data from the three experiments while still accounting for possible differences among universities.

Problem-solving score depended on the interaction of the worked example format and transfer distance, F (2, 188) = 5.23, p = .028, est. ω^2 = .08 (see Figure 3), matching previous results from independent experiments (*Study 1, Study 2, Study 3*) (Margulieux et al., 2016; Morrison et al., 2016, 2015). Due to the interaction, the main effects of worked example format and transfer distance will not be reported to avoid confusion in interpretation (Maxwell & Delaney, 2004). Instead, pairwise comparisons will be used as post hoc tests to explore the pattern of results. Exploring the effect of the university, there was no interaction of university and worked example format, p = .37, university and transfer distance, p = .65, nor university, worked example format, and transfer distance, p = .20. In addition, there was no main effect of university, p = .12; therefore, the combined data from all three universities were used for the *post hoc* tests.

Figure 3. Performance across six groups on problem-solving (code writing) tasks.



For *post hoc* analysis, we used simple main effects. Simple main effects analyze the effect of one independent variable for each level of the other independent variables. For example, simple main effects analysis will explore the effect of worked example format twice, once within the isomorphic transfer and once within the contextual transfer. Because the worked example format had three levels, the effect is analyzed with pairwise comparisons among each of the levels. The full results can be found in Table 4. The only two comparisons that were statistically significant were those within isomorphic transfer between given labels and generated labels and within contextual transfer between no labels and given labels. These results suggest that there are two levels of performance, low and high. The two lowest-scoring groups performed statistically worse than the two highest-scoring groups (see Figure 3). The two groups in the middle did not perform statistically different than the others, but they are numerically close to the lowest-scoring groups and had higher mean differences and effect sizes from the highest-scoring groups. Thus, we consider the two middle groups as low-performance groups.

Table 4. Pairwise comparisons evaluating simple main effect of worked example format. (Table view)

Transfer distance	Worked example format comparison	Std. error	Mean difference	Significance	Effect size (*d*)
Isomorphic	*No Subgoal* to *Given*	3.34	1.55	.64	.10
	No Subgoal to *Generate*	3.34	−4.80	.15	.37
	Given to *Generate*	3.52	−6.36	.02	.44
Context	*No Subgoal* to *Given*	3.01	−7.62	.01	.59
	No Subgoal to *Generate*	3.15	−1.49	.64	.11
	Given to *Generate*	3.24	6.14	.06	.47

To further explore performance, we split the problem-solving tasks into nearer (i.e. switched context) and farther (i.e. deviate from exact procedural steps) transfer from the instructional tasks. Switched context meant that we used the same type of contextual transfer as we used between the worked example and practice problem pairs. In this case, it describes transfer between the instructional tasks (i.e. worked example and practice problems) and the problem-solving tasks in this assessment. Procedure transfer means that the procedure used to solve the problem-solving task did not follow the exact same steps as the instructional tasks. For example, in the instructional tasks, participants had to use a while loop to find an average of a list, and in the problem-solving tasks, participants had to use a while loop to find an average of values that exceeded a threshold (examples can be found in the Supplementary – Worked Example

#1 and Assessment #2). The problem-solving task had extra steps but still used the same abstract procedure that was taught.

The results did not change when comparing groups within only the nearer or farther transfer tasks. In both cases, there was still a statistically significant interaction with the same pattern of scores, F (nearer; 2, 188) = 4.04, p = .02, est. ω^2 = .06, F (farther, 2, 188) = 2.99, p = .03, est. ω^2 = .05. These results suggest that the interventions had the same effect on problem-solving performance regardless of the type of transfer that was required to complete the problem-solving tasks.

Parsons problem score

The Parsons problem score was based on one Parsons problem and had a maximum score of 13 for putting each of the lines of code in the correct order. The overall mean score was 6.20, and the standard deviation was 4.27. Like for problem-solving performance, in the omnibus ANOVA analyses of these data, worked example format and transfer distance were treated as randomly assigned variables and the university was treated at a quasi-experimental variable.

Parsons problem score did not have a statistically significant main effect of worked example format, F (2, 188) = 1.11, p = .41, est. ω^2 = .03, transfer distance, F (2, 188) = 0.15, p = .73, est. ω^2 = .01, nor the interaction of the worked example format and transfer distance, F (2, 188) = 1.50, p = .31, est. ω^2 = .03. These results align with results in (*Study 3*) (Morrison et al., 2016) but not with (*Study 2*) (Margulieux et al., 2016), which found a main effect of worked example format and concluded that giving subgoal labels, regardless of transfer distance, improved Parsons problem score. This difference in results might be due to including only one Parsons problem in our protocol, possibly contributing to an unreliable measurement of Parsons problem performance. Based on the larger sample size of both the current analysis and that conducted in (*Study 3*) (Morrison et al., 2016), we would expect that the current result is more reliable. Therefore, we would not conclude that giving learners subgoals labels necessarily results in better performance on Parsons problems after receiving instructional materials similar to ours.

In the current analysis, we found a main effect of university, F (2, 188) = 10.16, p = .04, est. ω^2 = .06. There was no interaction of university and worked example format, p = .11, university and transfer, p = .51, nor university, worked example format, and transfer, p = .22. The difference between University 1 (M = 3.7) and University 2 (M = 4.6) was not statistically significant, t (116) = 1.35, p = .18, d = .25. In contrast, University 1 performed much worse than University 3 (M = 8.8), t (181) = 10.44, p < .001, d = 1.57. Similarly, University 2 performed much worse than University 3, t (133) = 5.53, p < .001, d = 1.04. These results are not unexpected, though, given that the participants from University 3 had already learned about solving problems with loops in

their programming courses. It is interesting that participants from University 3 performed statistically better than those in the other universities on the Parsons problem but not on the problem-solving tasks, which were writing code tasks. This supports the notion that students may demonstrate problem-solving knowledge in Parsons problems even if they cannot in traditional code-writing problems.

Post-test score

The post-test asked participants to complete, after instruction, the same five multiple-choice questions from the AP CS exam that they had completed prior to instruction. The maximum score was 5, and the mean was low, 2.40, with a standard deviation of 1.45. The post-test score did not have a statistically significant main effect of worked example format, F (2, 188) = 1.37, p = .34, est. ω^2 = .03, transfer distance, F (2, 188) = 0.24, p = .65, est. ω^2 = .01, nor the interaction of the worked example format and transfer distance, F (2, 188) = 1.39, p = .33, est. ω^2 = .02. These results align with individual experiment results from (*Study 1, Study 3*) (Morrison et al., 2016, 2015). In addition, there was no main effect of university, p = .76, interaction of university and worked example format, p = .50, university and transfer distance, p = .85, nor university, worked example format, and transfer distance, p = .27. We would expect, based on the results of the problem-solving tasks and Parsons problem, that participants would score higher on this post-test. Moreover, we would expect that participants from University 3 would perform better on this test than other participants because they were not excluded from analysis due to high pre-test scores and because they had learned about loops in their course already. Therefore, we conclude that this post-test, perhaps because it measured code-tracing skill more than problem-solving skill, did not effectively measure performance for any of the groups of participants, and we do not include this assessment when considering the conclusions of the study. These results support the idea that code tracing is a skill separate from code writing (Harrington & Cheng, 2018; Kumar, 2015).

Process data

To supplement our data about performance outcomes, we collected information about the learning process to explore differences among groups. These data include perceived cognitive load during instruction and time on task during instruction and assessment.

Cognitive load

The cognitive load survey asked participants questions about their cognitive load directly after instruction to measure their perceptions of cognitive load during instruction

(Morrison et al., 2014). Each of the 10 questions asked participants to rate their perceived cognitive load (e.g. “The topics covered in the activity were very complex”) on a scale from “0 – not at all the case” to “10 – completely the case,” making the maximum score 100. The mean was 40.9 with a standard deviation of 14.6. The cognitive load did not have a statistically significant main effect of worked example format, F (2, 188) = .51, p = .63, est. ω^2 = .01, transfer distance, F (2, 188) = 0.89, p = .43, est. ω^2 = .02, nor the interaction of the worked example format and transfer distance, F (2, 188) = .56, p = .60, est. ω^2 = .01. Furthermore, there was no main effect of university, p = .35, interaction of university and worked example format, p = .51, university and transfer distance, p = .61, nor university, worked example format, and transfer distance, p = .20, suggesting no differences among universities.

These results were not previously reported for individual experiments due to space constraints. In this case, though, finding no statistical difference is good as it suggests that students did not perceive a meaningful difference in mental workload even though the instructions asked them to engage in different tasks. One possible explanation of these null results is that participants in all conditions used the same amount of mental resources, whether they were engaging in our prescribed learning strategy or not. We have no supplemental evidence to make a strong argument for this possibility. We can say, however, that some participants performed better than others without perceiving differences in mental workload.

Time on task

The total amount of time that participants spent on the experiment was recorded. This includes time spent studying worked examples, solving practice problems, and completing the assessments. The amount of time that participants spent on the task depended on worked example format, F (2, 188) = 8.67, p < .001, est. ω^2 = .09. There was no effect of transfer distance, F (2, 188) = 0.55, p = .46, est. ω^2 = .003, nor was there an interaction, F (2, 188) = 1.20, p = .30, est. ω^2 = .01. Performance did not interact with university either, F (2, 188) = 0.63, p = .67, est. ω^2 = .002.

To explore the effect of worked example format on time on task, we used simple main effects analysis. Within the isomorphic transfer condition, *No Subgoal* participants completed the task faster (M = 52 min, SD = 21 min) than participants in the *Given* (M = 72, SD = 27) or *Generate* (M = 71, SD = 29) conditions, Mean Difference = 20.1 and 18.8 min, p = .003 and .007, d = .83 and .75, respectively.

The *Given* and *Generate* conditions did not differ on time on task, Mean Difference = 1.4 min, p = .85, d = .04. When considering the effect on time, it is important to remember that within the isomorphic transfer condition, participants who generated their own subgoal labels performed best, and participants without subgoal labels or who were given subgoal labels did not perform differently. This combination of results means that

participants who generated subgoal labels with isomorphic transfer took longer than those who did not receive subgoals, but they performed better. In contrast, participants who were given subgoal labels with isomorphic transfer took longer than those who did not receive subgoals but did not perform better. Therefore, taking longer to complete the task did not result in better performance for each group.

Following a similar pattern within the context transfer condition, the *No Subgoal* participants completed the task faster (M = 59 min, SD = 25 min) than participants in the *Given* (M = 67, SD = 25) or *Generate* (M = 79, SD = 35) conditions, Mean Difference = 12.2 and 20.1 min, p = .076 and .005, d = .32 and .66, respectively. Though the difference between the *No Subgoal* and *Given* groups is not statistically significant, we argue that it is meaningfully significant, albeit small, based on the mean difference and d value. The *Given* and *Generate* conditions did not meaningfully differ on time on task, Mean Difference = 7.9 min, p = .22, d = .21.

A piece of information to highlight from these results is that the standard deviation for the group who generated subgoals with the contextual transfer was 35 min, which is approximately 10 min more than the other groups. This means that participants in this condition had much more variance in the amount of time on task than those in other conditions. If we were to offer a *post hoc* explanation of this finding based on our observations as experimenters and exploring the data, we might argue that participants in this group were more likely to flounder and take an excessively long time to complete the experiment. This group had twice as many people as any other group who took 100 min or longer (6 participants compared to 1–3 participants in the other groups).

Similar to the isomorphic transfer condition, it is important to recognize that within the contextual transfer condition, participants who were given labels performed better than others. This combination of results means that participants who were given subgoal labels with context transfer took slightly longer than those who did not receive subgoals, but they performed better. Moreover, participants who generated subgoals with context transfer took substantially longer than those who did not receive subgoals, but they did not perform better. The combined results suggest that depending on the transfer distance and worked example format, better performance required more time on task, but more time on task did not guarantee better performance.

To explore the relationship between time on task and performance more deeply, we examine the correlation between these two dependent variables within each group. Overall, there was a strong, positive relationship between performance and time on task, r = 0.43, p < .001, as is typical in education research. However, this relationship was not consistent within each experimental group (see Table 5), suggesting that spending longer on the task did not necessarily coincide with higher performance. The relationship between time on task and performance was strongest when students learned subgoals with isomorphic transfer between examples and practice problems or

when students did not learn subgoals with contextual transfer. The relationship was weakest when students generated subgoals with contextual transfer or when students did not learn subgoals with isomorphic transfer. Therefore, despite the extra time that students learning subgoals spent on the task, their extra effort did not consistently result in higher performance. As such, we conclude that the benefit of learning subgoals (under particular circumstances) is due to more than coaxing students to spend more time on task.

Table 5. Correlation between time on task and performance within each experimental group. (Table view)

Experimental group	No subgoal	Given subgoal	Generate subgoal
Isomorphic Transfer	$r = 0.26$, $p = .140$	$r = 0.67$, $p < .001$	$r = 0.51$, $p = .008$
Contextual Transfer	$r = 0.53$, $p = .001$	$r = 0.41$, $p = .011$	$r = 0.27$, $p = .186$

Discussion

In the cumulative analysis of three studies that used the same experimental protocol across three groups of learners at different institutions, we found that the most effective instructional design interventions were those that (1) gave subgoal labeled worked examples with farther transfer between worked examples and practice problems or (2) asked students to generate subgoal labels for worked examples with nearer transfer between worked examples and practice problems. In our experiment, these two conditions performed equally, but in practice, there might be reasons to pick one over the other based on several factors, such as characteristics of students in the class, the teaching style of the instructor, or the instructional materials (e.g. curriculum or textbook) being used.

The students in the class might affect whether they will successfully generate subgoal labels. If the students are already engaging in self-explanation (e.g. they answer challenging questions in class), learn concepts quickly, or are highly motivated to learn the content, then promoting self-explanation through the generation of subgoal labels might be particularly effective. When we analysed the content of the subgoals generated by students, we found that students who learned with contextual transfer and generated more generalizable subgoals performed significantly better at problem-solving than any other group (Margulieux et al., 2016). If the students tend to be unable to self-explain, are otherwise struggling in the course (i.e. exhibit signs of already having high cognitive load), or seem unmotivated to learn, then giving subgoal labels would likely be more effective than asking them to generate subgoal labels. Based on whether

students generate or are given labels, the transfer distance between worked examples and practice problems can be adjusted to match the most effective conditions.

The instructor's teaching style could also affect how students should engage with subgoal labels. Based on (Margulieux & Catrambone, 2019), students who generated subgoal labels and received feedback on those labels performed better than students who generated subgoal labels without feedback. Therefore, if the instructor's teaching style includes providing feedback or class discussion during which students can refine their generated labels, then generating labels might be more effective than given labels. In contrast, if the course includes a lot of independent learning without many opportunities for feedback or too many students for the instructor to provide individual feedback, then generating labels might be no different than given labels, as was the case in these studies.

The last factor that might determine which type of subgoal learning best suits a course is the curricular materials being used in the course. If the curricular materials, including worked examples and practice problems, were designed by someone else, then the transfer distance between the worked examples and practice problems should determine the type of subgoal learning used. Isomorphic transfer would be best complemented by generating labels, and contextual transfer would be best complemented by given labels (Margulieux, Morrison, & Decker, 2019).

If isomorphic transfer between worked examples and practice problems is an option and the instructor does not have the time or resources to identify subgoal labels for the procedure, then allowing learners to generate subgoal labels for themselves is a good option. To do this, the instructor could use subgoal label training, add a prompt at the end of each problem-solving step, and ask students to generate their own labels. This option would likely be most effective if the instructor matched features between worked examples (e.g. step 2 of the first example is like step 3 of the second example). Margulieux and Catrambone (2019) found that providing hints about which features are similar between worked examples helped students to perform better when they generated their own subgoal labels. Like the feedback described in the paragraph about teaching style, providing hints could further improve the problem-solving skill of learners who are generating their own subgoal labels. It is important to clarify that (Margulieux & Catrambone, 2019) found that providing both hints and feedback did not improve performance; therefore, if feedback or hints are provided, the instructional materials should provide only one or the other.

However, if pre-defined given subgoal labels are used, the worked example – practice problem pairs should utilize contextual transfer to ensure maximum learning. As mentioned earlier, this is contradictory to what would be predicted by cognitive load theory. This is certainly one phenomenon that needs further research. It may be that with given subgoal labels and isomorphic problems students do not adequately self-

explain the process associated with each subgoal as the steps are identical within both the worked example and practice problem. To ensure that students in the given subgoal labels with isomorphic practice problems were adequately studying the worked example and attempting the practice problem, we examined the student code submissions. We reviewed student code submissions to ensure that they were not copied from the worked example. We did a visual inspection and a character by character comparison from the student code submission to the worked example presented. We found no instances of an exact copy of worked example code or any student submissions more than 10% identical to the worked example. Also, the time spent in the instructional period indicates that participants spent similar amounts of time regardless if they received isomorphic or contextual transfer worked example – practice problem pairs.

Conclusion

In this paper, we have aggregated the data from three previous studies to take a more holistic view and to examine the results for generalization across populations to provide the most nuanced and accurate information for using subgoal-labeled instructional materials in the classroom. By combining the data for maximum statistical power, we can view effect sizes to determine which treatments are likely to yield similar results in the future.

Our research into subgoal-labeled instruction in computing represents the first attempt in any discipline (that we are aware of) to test the generation of subgoal labels by participants and its effect on learning performance. We are also the first (to our knowledge) to vary transfer between worked examples and practice problems. By introducing these additional conditions into our research, we have found combinations which provide the most beneficial experience for the learner:

1. Given subgoal labeled worked examples with farther transfer between worked examples and practice problems or
2. Student-generated subgoal labels for worked examples with nearer transfer between worked examples and practice problems.

Either condition should yield the highest performance from students. Which you choose to implement may depend on the conditions discussed above.

Limitations

Our results are limited to having student performance data for only a single lab during an introductory programming course. From this, we cannot speculate or generalize to what the long-term impacts are from a learning trajectory perspective. Additionally, the

tests were conducted during a single lab session with no delayed test for knowledge over time. Thus, our results only speak to the immediate learning outcomes.

Another potential limitation of this work is the necessary solitary work required of the participants. We asked students in a lab to work alone at their computer for 1–2 h without assistance from peers or instructors or teaching assistants. This condition was necessary for experimental integrity but is not ecologically valid for many classroom lab environments. While we do not expect that collaboration would negate the learning effects of subgoal labels, it may affect them in unpredictable ways. For example, if some students found similarities between the worked example and practice problem and then helped others in the lab, then the farther transfer intervention might become universally more effective than the nearer transfer intervention. In the study condition where students were asked to generate subgoal labels, if students were working together then the condition would transform from a self-explanation activity to a peer-explanation activity which may or may not benefit each individual student (Chi, 2009).

Future work

Current research has examined student performance in learning with only a single construct within an entire introductory programming course (while loops). Research has moved from a laboratory environment (Margulieux et al., 2012) to a single lab instance (Morrison, Marguliuex, & Guzdial, 2015; Marguliuex, et al., 2016; Morrison, Marguliuex, & Decker, 2016) *Study 1, Study 2, Study 3*) (Morrison et al., 2015). The next logical step would be to use subgoal labels throughout an entire course and measure student learning. This could be implemented using either of the most beneficial subgoal conditions. Given subgoal labels could be used as long as the worked example and practice problems represented further transfer. Or students could be trained to generate their own subgoal labels and provided with worked example-practice problems with near transfer, and if students receive feedback on their generated subgoal labels to ensure generality.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Science Foundation 1712231 and 1712025.

Notes on contributors

Briana B. Morrison is an Assistant Professor of Computer Science at the University of Nebraska Omaha. She has over 20 years' experience in teaching computer science.

She has served on the ACM SIGCSE Board and ACM Education Committee. Her research area is CS Education where she explores cognitive load theory within learning programming, broadening participation in computing and expanding and preparing computing high school teachers.

Lauren E. Margulieux is an Assistant Professor of Learning Sciences at Georgia State University. Her research interests are in educational technology and online learning, particularly for computing education. She focuses on designing instructions in a way that supports online students who do not necessarily have immediate access to a teacher or instructor to ask questions or overcome problem solving impasses.

Adrienne Decker is an Assistant Professor in the Department of Engineering Education at University at Buffalo. Her research interests are in computing education, particularly at the introductory level. She is interested in techniques that support learning of introductory programming material at the university level and the impact that exposure to computing prior to university has on learners in the introductory courses.

Supplementary material

Supplemental data for this article can be accessed here.

References

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, 94(2), 416. Crossref.

Atkinson, R. K., Catrambone, R., & Merrill, M. M. (2003). Aiding transfer in statistics: Examining the use of conceptually oriented equations and elaborations during subgoal learning. *Journal of Educational Psychology*, 95(4), 762. Crossref.

Atkinson, R. K., & Derry, S. J. (2000). Computer-based examples designed to encourage optimal example processing: A study examining the impact of sequentially presented, subgoal-oriented worked examples. *Fourth International Conference of the Learning Sciences. Presented at the ICLS*. Ann Arbor, Michigan.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, 70(2), 181–214. Crossref.

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction*, 13(2), 221–252. Crossref.

Bransford, J. (2000). *How people learn: Brain, mind, experience, and school*. Washington, D.C.: National Academies Press.

Catrambone, R. (1994). Improving examples to improve transfer to novel problems. *Memory & Cognition*, 22(5), 606–615. Crossref. PubMed.

Catrambone, R. (1995). Aiding subgoal learning: Effects on transfer. *Journal of Educational Psychology*, 87(1), 5. Crossref.

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(4), 1020. Crossref.

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*, 127(4), 355. Crossref.
Chi, M. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1), 73–105. Crossref. PubMed.
Chi, M., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145–182. Crossref.
Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah: Erlbaum.
Eiriksdottir, E., & Catrambone, R. (2011). Procedural instructions, principles, and examples how to structure instructions for procedural tasks to enhance performance, learning, and transfer. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, 53(6), 749–770. Crossref. PubMed.
Gravetter, F. J., & Wallnau, L. B. (2016). *Statistics for the behavioral sciences*. Boston, MA: Cengage Learning.
Harrington, B., & Cheng, N. (2018). Tracing vs. writing code: Beyond the learning hierarchy. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (pp. 423–428). 10.1145/3159450.3159530 Crossref.
Joentausta, J., & Hellas, A. (2018). Subgoal labeled worked examples in K-3 education. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (pp. 616–621). ACM, Baltimore, MD. Crossref.
Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, 23(1), 1–19. Crossref.
Kim, J., Miller, R. C., & Gajos, K. Z. (2013). Learnersourcing subgoal labeling to support learning from how-to videos. In *CHI'13 extended abstracts on human factors in computing systems* (pp. 685–690). ACM, Paris, France. Crossref.
Kumar, A. N. (2015). Solving code-tracing problems and its effect on code-writing skills pertaining to program semantics. *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 314–319). 10.1145/2729094.2742587 Crossref.
Margulieux, L. E., & Catrambone, R. (2014). Improving problem solving performance in computer-based learning environments through subgoal labels. *Proceedings of the First ACM Conference on Learning@ Scale Conference* (pp. 149–150). ACM, Atlanta, GA. Crossref.
Margulieux, L. E., & Catrambone, R. (2019). Finding the best types of guidance for constructing self-explanations of subgoals in programming. *Journal of the Learning Sciences*, 28(1), 108–151. Crossref.
Margulieux, L. E., Catrambone, R., & Guzdial, M. (2016). Employing subgoals in computer programming education. *Computer Science Education*, 26, 1–24. Crossref.
Margulieux, L. E., Guzdial, M., & Catrambone, R. (2012). Subgoal-labeled instructional material improves performance and transfer in learning to develop mobile applications. *Proceedings of the Ninth Annual International Conference on International Computing Education Research* (pp. 71–78). ACM, Auckland, New Zealand. Crossref.

Margulieux, L. E., Morrison, B. B., Catrambone, R., & Guzdial, M. (2016). Training learners to self -explain: Designing instructions and examples to improve problem solving. *Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) 2016*(pp. 1). Singapore.

Margulieux, L. E., Morrison, B. B., & Decker, A. (2019). Design and pilot testing of subgoal labeled worked examples for five core concepts in CS1. *ITICSE'19: Innovation and Technology in Computer Science Education Proceedings* (pp. 7). 10.1145/3304221.3319756 Crossref.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2ne ed.). New York, NY: Psychology Press.

Morrison, B. B., Decker, A., & Margulieux, L. E. (2016). Learning loops: A replication study illuminates impact of HS courses. *Proceedings of the 2016 ACM Conference on International Computing Education Research* (pp. 221–230). 10.1145/2960310.2960330 Crossref.

Morrison, B. B., Dorn, B., & Guzdial, M. (2014). Measuring cognitive load in introductory CS: Adaptation of an instrument. *Proceedings of the Tenth Annual Conference on International Computing Education Research* (pp. 131–138). Crossref.

Morrison, B. B., Margulieux, L. E., Ericson, B., & Guzdial, M. (2016). Subgoals help students solve parsons problems. *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (pp. 42–47). 10.1145/2839509.2844617 Crossref.

Morrison, B. B., Margulieux, L. E., & Guzdial, M. (2015). Subgoals, context, and worked examples in learning computing problem solving. *Proceedings of the Eleventh Annual International Conference on International Computing Education Research* (pp. 21–29). Crossref.

Parsons, D., & Haden, P. (2006). Parson's programming puzzles: A fun and effective learning tool for first programming courses. *Proceedings of the 8th Australasian Conference on Computing Education - Volume 52* (pp. 157–163). Darlinghurst, Australia, Australia: Australian Computer Society, Inc.

Plass, J. L., Moreno, R., & Brünken, R. (2010). *Cognitive load theory*. Cambridge, UK: Cambridge University Press. Crossref.

Renkl, A., & Atkinson, R. K. (2002). Learning from examples: Fostering self-explanations in computer-based learning environments. *Interactive Learning Environments*, 10(2), 105–119. Crossref.

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., … Silverman, B. (2009). Scratch: Programming for all. *Commun. Acm*, 52(11), 60–67. Crossref.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22(2), 123–138. Crossref.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory* (Vol. 1). New York: Springer-Verlag. Crossref.

Sweller, J., van Merriënboer, J. J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. Crossref.

Tew, A. E., & Guzdial, M. (2011). The FCS1: A language independent assessment of CS1 knowledge. *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education* (pp. 111–116). ACM, Dallas, TX. Crossref.

van Gog, T., & Paas, F. (2012). Cognitive load measurement. In Norbert M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 599–601). New York: Springer. Crossref.
van Merriënboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17(2), 147–177. Crossref.
Wylie, R., & Chi, M. T. (2014). 17 the self-explanation principle in multimedia learning. In Mayer, R. (Ed.), *The cambridge handbook of multimedia learning (Cambridge Handbooks in Psychology)* (p. 413) Cambridge: Cambridge University Press. . Crossref.