Journal of Experimental Psychology: General
1998, Vol. 127, No. 4, 355–376

0096-3445/98/$3.00

The Subgoal Learning Model: Creating Better Examples So That Students Can Solve Novel Problems

Richard Catrambone
Georgia Institute of Technology

Learners have great difficulty solving problems requiring changes to solutions demonstrated in examples. However, if the solution procedures learners form are organized by subgoals, then they are more successful. Subgoal learning is hypothesized to be aided by cues in example solutions that indicate that certain steps go together. These cues may induce a learner to attempt to self-explain the purpose of the steps, resulting in the formation of a subgoal. Across 4 experiments it was found that a label for a group of steps in examples helped participants form subgoals as assessed by measures such as problem-solving performance and talk aloud protocols. Abstract labels were more likely than superficial labels to lead participants to form subgoals with fewer ties to surface features. Subgoals guide problem solving by helping learners focus on the steps to modify in novel problems that involve the same subgoals but require new steps to achieve them.

Learners have difficulty solving problems that involve more than minor changes to the procedure demonstrated by training problems or examples (e.g., Bassok, Wu, & Olseth, 1995; Catrambone, 1994, 1995, 1996; Novick & Holyoak, 1991; Reed, Dempster, & Ettinger, 1985; Ross, 1987, 1989). People tend to form solution procedures that consist of a long series of steps—which are frequently tied to incidental features of the problems—rather than more meaningful representations that would enable them to successfully tackle new problems (Singley & Anderson, 1989).

Consider a student who studies the following worked example (adapted from Reed, Ackinclose, & Voss, 1990):

> Tom can mow his lawn in 1.5 hours. How long will it take him to finish mowing his lawn if his son mowed 1/4 of it?
>
> Solution:
>
> $(1/1.5 * h) + .25 = 1 \rightarrow (.67 * h) + .25 = 1$
>
> $\rightarrow .67 * h = .75 \rightarrow h = 1.13$ hrs,

where h is the number of hours worked.

A student might learn from this that the way such "work" problems are solved is to take one person's time and divide 1 by it, multiply it by the unknown, add the amount that was already done, and set it all equal to 1. Such an approach would not be successful for the following problem (also adapted from Reed et al., 1990):

> Bill can paint a room in 3 hours and Fred can paint it in 5 hours. How long will it take them if they both work together?
>
> Solution:
>
> $(1/3 * h) + (1/5 * h) = 1 \rightarrow (.33 * h) + (.20 * h) = 1$
>
> $\rightarrow .53 * h = 1 \rightarrow h = 1.89$ hrs.

Although the same conceptual approach is used in both cases—represent the amount of work done by each worker and set it equal to the total amount of work to do—the step-by-step approach described after the first problem's solution does not capture it. A learner who memorized such a step-by-step approach would have difficulty solving the second problem because he or she would have little guidance for adapting the solution. This is the type of result often found in studies of problem solving (Novick & Holyoak, 1991; Reed et al., 1990).

Such findings are a cause for concern. Presumably one of the jobs of education is to equip people to deal with novel problems and situations, not just a small recognizable set. Yet it appears that this job does not get done. Learners seem to be predisposed, or the environment shapes them to develop the disposition, to have their problem solving guided by sets of memorized steps and by surface features of problems (Chi, Feltovich, & Glaser, 1981; Larkin, McDermott, Simon, & Simon, 1980; Ross, 1987, 1989). Surface or superficial features are those aspects of problems that, when changed, do not affect the solution procedure; that is, they have no necessary relevance to the solution to the problem. For example, it does not necessarily matter whether a mechanics problem in physics involves a block sliding down an inclined plane or two objects suspended over a pulley. If the problem asks for the velocity of a particular object, the

This research was supported by Office of Naval Research Grant N00014–91-J-1137. Experiment 1 was reported at the 17th Annual Meeting of the Cognitive Science Society, Pittsburgh, Pennsylvania, July 1995. Experiment 2 was reported at the 35th Annual Meeting of the Psychonomic Society, St. Louis, Missouri, November 1994. Experiment 3 was reported at the Office of Naval Research Grantees Meeting on Learning and Training, Chicago, September 1994.

Correspondence concerning this article should be addressed to Richard Catrambone, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332–0170. Electronic mail may be sent to rc7@prism.gatech.edu.

same set of general equations involving Newton's Laws and linear motion can be used even though the details of how the equations are filled out and related to one another will vary depending on information provided in the problem. Learners often do not realize that seemingly different sets of steps across problems might be calculating the same thing such as the force acting on a particular object.

Students tend to memorize the details of how the equations are filled out rather than learning the deeper, conceptual knowledge that is implicit in the details. Thus, if they are given a new problem that seems similar to an old one—at a surface level—they will try to apply a set of steps from the old problem. These steps are invoked when the learner recognizes certain features to be present in a problem (e.g., a problem involving a block sliding down an inclined plane). If the steps can not be used, the learner will frequently not know what to do.

A more fruitful approach to problem solving would be to organize one's problem-solving knowledge in some way that ties the steps to a meaningful hierarchical structure (e.g., Anzai & Simon, 1979; Brown, Kane, & Echols, 1986; Catrambone & Holyoak, 1990; Newell & Simon, 1972). This organization is more consistent with the way that experts tend to go about solving problems (e.g., Larkin et al., 1980). Although an expert's conception of how to solve problems in a domain does not necessarily tell us how to instruct novices, it does provide guidance on the types of organizing elements that might be useful for the novice to form.

Forming Hierarchical Organizations for Solving Problems

Increasingly studies are showing that teaching and training materials can be developed that *will* help learners form more useful representations of problem-solving knowledge. One such useful structure is a hierarchical one in which the higher level or conceptual aspects of a solution procedure form the skeleton of the solution approach. The higher level pieces are perhaps the ones the experienced solver initially accesses from memory. These pieces in turn can guide the search for lower level pieces that expand or instantiate the higher level ones. The various pieces of such a hierarchical structure are related typically through links that might indicate causal or other types of relationships (Gentner, 1983). For instance, a goal to find the amount of work a worker has done would have a connection to a lower part of the hierarchy for finding the worker's rate and how long he or she has been working.

Consider a hierarchical approach to solving a physics mechanics problem. Such an approach might include high-level goals such as to identify the relevant "systems" in the problem, to describe the systems, and to check the description qualitatively (Heller & Reif, 1984). The description of the system can be broken down into subgoals to describe the motion of the system and to describe the forces acting on the system. The description of the forces can be broken down into subgoals to identify each object touching the system of interest as well to identify long-range forces such as gravitational effects. Eventually, as one moves lower in the hierarchy, the subgoals may call upon certain concrete methods or steps to achieve them rather than calling on more subgoals.

Various studies have found that learners who form a hierarchical representation are typically able to solve novel problems more successfully than learners who were led to form a step by step organization of the problem-solving procedure (e.g., Dufresne, Gerace, Hardiman, & Mestre, 1992; Eylon & Reif, 1984). In these studies, researchers usually derived what they believed to be a useful hierarchical approach to problem decomposition and attempted to induce learners to internalize this approach by having them follow a prescribed method for solving or processing training problems. For instance, Heller and Reif (1984) formulated a model specifying the underlying knowledge and procedures needed to successfully solve mechanics problems. The authors required participants to solve three problems by adhering to a hierarchical approach for redescribing each problem in terms of relevant forces. The model was contrasted with a "flat" solution approach that intentionally omitted certain levels of the redescription process hierarchy such as checking for consistency between the direction of forces and the resulting acceleration (learners were not prevented from doing these checks, but they were not reminded to do them). Participants who were required to follow the more hierarchical approach were significantly better at redescribing novel problems and solving them.[1]

Subgoal Learning

The results of the studies mentioned previously are consistent with the claim that more hierarchically organized solution procedures may lead to better performance on novel problems compared to procedures that are essentially a set of memorized linear steps. As implicitly demonstrated earlier, one type of knowledge structure that would typically be associated with the higher levels of a hierarchy are goals and subgoals.

The term *subgoal* has been used in at least two ways in the problem-solving and transfer literatures. One use has been to consider a subgoal something generated by a learner when he or she reaches an impasse during problem solving (e.g., Newell, 1990, chap. 4; VanLehn, 1988). A second use is to consider a subgoal to be a feature of a task structure that can

[1] Heller and Reif (1984) were quick to point out that the hierarchical approach was only prescriptive and did not necessarily have any direct relationship to internal representations learners might have formed by following the model. It is also the case that besides containing certain high-level subgoals such as describing motion that the flatter approach lacked, the hierarchical approach included a number of low-level "hints," such as reminding learners to include properties such as mass in their drawings, that the flatter approach did not include. Thus, it is not entirely clear whether requiring participants to focus on a certain prescribed set of subgoals was most responsible for superior performance on new problems or whether the lower level procedural details were also crucial. This is not an issue in the present work because all learners saw the same procedural details in the training examples.

be taught to a learner (e.g., Catrambone & Holyoak, 1990; Dixon, 1987). In both cases it is usually predicted that a learner possessing an appropriate subgoal, or set of subgoals, will be in a better position to solve new problems compared to learners who memorized only a set of steps.

As used in the present article, a subgoal represents a meaningful conceptual piece of an overall solution procedure. To return to the problems given at the beginning of this article, a subgoal might be to find the amount of work done by each worker. To achieve such a subgoal might involve calling on lower level subgoals to find a worker's rate and to find how long the worker worked. These subgoals represent mini-problems in the context of solving the overall problem of determining how long it will take a worker to accomplish some task.

Relatively little research has been conducted on how learners form subgoals; most efforts have been directed towards predicting transfer by learners assumed to already possess the subgoals. One attempt to explain how subgoals might be learned was made by Anzai and Simon (1979). They recorded the moves and verbal protocol of a learner as she solved the Tower of Hanoi problem multiple times. Over trials the learner began to chunk groups of moves. That is, she would make a set of moves in quick succession followed by a pause before the next set of moves. Anzai and Simon argued that each burst of moves represented a chunk. Each chunk may have reflected a subgoal the learner was achieving by the particular burst of steps.

Anzai and Simon (1979) suggested that subgoal acquisition is greatly aided when the search space for operators (e.g., possible moves in the Tower of Hanoi problem) is constrained. When the search space is constrained, working memory load is reduced. One hypothesized advantage of a working memory load reduction is that a learner is better able to notice and remember a sequence of steps that led to a particular outcome (see also Sweller, 1988). In Anzai and Simon's model this working memory load reduction aids subgoal formation because a *sub*goal is formed when a learner is working towards a certain goal (perhaps derived from task instructions) and notices that a set of steps places him or her in a situation to carry out additional steps that ultimately achieve the goal. The learner will be better able to notice the result of the first set of steps, and be able to chunk that sequence of steps, if working memory load has been reduced.

With respect to the current study, it is hypothesized that features of example solutions can cue a learner to chunk or group a set of steps. As a result, the learner's chances of discovering that a particular outcome, the subgoal, can be achieved by executing that series of steps is increased. One such cue that could encourage grouping, and thus subgoal learning, is a label.

Results from the categorization literature are consistent with the previous speculation on the effect of a cue on grouping. Medin and his colleagues (e.g., Medin, Wattenmaker, & Hampson, 1987; Wattenmaker, Dewey, Murphy, & Medin, 1986) found that when learners were provided with a theme during a training session—e.g., think of objects in one category as being or not being reasonable substitutes for a hammer—they were quicker to learn categories. This was particularly true if causal or explanatory links to the category could be made for the features of the objects (see also Cabrera & Billman, 1996; Homa & Cultice, 1984). With respect to the proposed effect of a label on grouping, the categorization results suggest that features of example solutions that help learners form causal or explanatory links among solution steps will help learners form a category (i.e., a subgoal) that captures a useful relationship among the steps.

Why Subgoals Aid Problem Solving

Subgoals can be used by a learner to help him or her solve novel problems because problems within a domain typically share the same set of subgoals, although the steps for achieving the subgoals might vary from problem to problem. For instance, in algebra problems dealing with work, subgoals for determining each worker's rate and time are typically present even though these rates and times will be found in different ways depending on the givens in the problem.

Suppose the learner is attempting to achieve a particular subgoal in a novel problem and discovers that the old set of steps, perhaps learned from an example, will not work. The learner will have a reduced search space to consider when trying to adapt the method because he or she knows on which steps to focus for changing the procedure: the steps associated with the current subgoal. In contrast, a learner who has learned a solution procedure consisting of a single goal reached by a long series of steps will have a larger space to search for possible steps to change and thus, be less likely to determine successfully which steps need to be modified. If a particular subgoal needs to be achieved in a very different way than was demonstrated in the example (i.e., new steps are required rather than a modification of old steps), a learner possessing a representation with subgoals will have some guidance about what prior knowledge might be relevant for achieving that subgoal. A learner who memorized only a series of steps will be less likely to identify what prior knowledge he or she possesses that might be useful.

Reed et al.'s (1990) findings are consistent with the previous interpretation. They found that learners tend to memorize a particular set of steps for solving problems, and changes to that procedure, even if the change results in a simpler procedure, tend to cause decrements in performance. This observation echoes the findings of Luchins (1942) who observed that when a person learned a procedure for solving a problem, he or she would frequently follow that procedure for subsequent problems even though a simpler procedure could be used. Learners appeared to be blocked from discovering the simpler procedure because the older, more complicated one could be used. However, in many problem-solving situations, old procedures will produce incorrect results or can not even be carried out because some of the information needed to carry out the old procedure might be missing.

Subgoal Learning Model

The subgoal learning model (Catrambone, 1995, 1996) assumes that if a learner is cued that a set of solution steps belong together, he or she will be more likely to try to self-explain why those steps belong together, that is, to determine their purpose. This is similar to one of the types of self-explanations that Chi and her colleagues (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994) have observed as students studied various texts in physics and biology. The subgoal learning model can be summarized as follows:

1. A cue leads learners to group a set of steps.
2. After grouping the steps, learners are likely to try to self-explain why those steps go together.
3. The result of the self-explanation process is the formation of a goal that represents the purpose of that set of steps.

Although most learners can presumably engage in a self-explanation process with varying degrees of success, good students seem better at determining the appropriate boundaries between meaningful groups of steps in a solution procedure (Chi et al., 1989). The use of a label in examples may serve as a cue to the boundaries (cf. Ausubel, 1968). Ben-Zeev (1995) suggested that the same induction mechanism is capable of forming correct and incorrect rules as a function of the information from which it works. Thus, a learner will have a better chance of forming the appropriate subgoals and methods for solving problems in a domain if the training examples are constructed in a way to aid the induction process. The subgoal learning model predicts that cues for grouping help this process.

Although various indirect measures (e.g., problem-solving performance, learners' posttest descriptions of how to solve problems) have provided evidence supporting the claim that a cue can lead a learner to self-explain the purpose of a set of steps (Catrambone, 1994, 1995), one study also provided more direct evidence for this connection (Catrambone, 1996, Experiment 2). In that experiment learners were asked to explain the solutions to the examples as they studied them. Participants studying examples that labeled a group of steps were more likely than other participants to make a statement about those steps being a unit and were also more likely to offer an explanation for what the steps accomplished.

Testing Implications of the Subgoal Learning Model

The present research tests implications of the subgoal learning model by examining whether learning subgoals helps students overcome various difficulties in transfer to novel problems. A novel problem is defined as a problem that involves the same subgoal structure as the training examples but requires new or modified steps to achieve one or more of the problem's subgoals.

The subgoal learning model will be tested through manipulations that, according to the model, should aid subgoal learning and thus improve transfer. For instance, the model assumes that a learner will self-explain a set of grouped steps in order to determine their purpose. The success of this process though presumably depends on, among other things, the relevant background knowledge the learner can bring to bear (cf. Chi et al., 1994). Thus, a manipulation involving background knowledge should affect subgoal learning in predictable ways.

As a second manipulation, consider the semantics of a label used as a grouping cue. A label with ties to surface features of an example could lead a learner to form a subgoal that is linked to the surface features of the example rather than being more general. This link could affect subsequent transfer to problems that manipulate the relationship of the surface features and the solution procedure.

In addition, the semantics of the label and a learner's background knowledge may interact to affect subgoal learning. For example, a learner with a stronger background is expected to be more likely than a learner with a weaker background to form an appropriate subgoal from a label that is more abstract. The learner with the weaker background might require a label that has connections to the surface features of the problem in order to form a subgoal; such connections might provide additional guidance as to the purpose of the steps. However, the resulting subgoal could be inappropriately tied to surface features of the problem and thus, could be misleading for future problems. The learner with a stronger background would be better served with an abstract label because he or she is more likely to have the background knowledge needed to form the subgoal using an abstract cue and will not be exposed to a label that could potentially mislead him or her.

A reasonable question to ask at this point is: why not directly state the subgoals to learners rather than embedding them in examples? There are two problems with this approach. First, learners exhibit a clear preference for learning from and referring to examples when faced with new problems (e.g., LeFevre & Dixon, 1986; Pirolli & Anderson, 1985). Fong, Krantz, and Nisbett (1986) and Cheng, Holyoak, Nisbett, and Oliver (1986) found that the addition of examples to their training materials aided learning. Fong and Nisbett (1991) suggested that supplementary examples to a textual description of a solution procedure may provide learners with application or coding rules to help the learner map a solution onto a test problem. Second, although there have been a small number of successes teaching solution procedures directly (Fong et al., 1986), most attempts have been unsuccessful (e.g., Reed & Bolstad, 1991). One possible reason for the tendency to fail at teaching procedures directly may be that the materials did not encourage learners to form subgoals that organized the steps but rather to focus primarily on memorizing the steps.

Forming and Adapting Solution Procedures

Various researchers proposed that the following events occur when a learner is attempting to use a previously encountered example to help him or her solve a new problem: encoding of the original example and the test

problem, retrieval of that example when working on the test problem, mapping the solution from the original example to the test problem (that is, lining up features of the problem to features of the example), and finally perhaps forming a generalization or schema that covers the solution approaches used in the original example and in the test problem (see Gick & Holyoak, 1983; Keane, 1987; Reeves & Weisberg, 1994; Ross 1987, 1989).

Two features of the previous scenario need to be addressed with respect to the current article. First, the mapping phase may require the learner to adapt the solution procedure from the example. That is, besides bringing into correspondence the relevant parts of the example and the problem, the learner might need to alter the solution from the example if the same steps from the example can not be used in the problem (Novick & Holyoak, 1991). Thus, a failure to solve the test problem may not be a failure in mapping (i.e., the learner imports the relevant example's solution procedure to the test problem) but rather a failure in adapting the solution to make it work for the current problem. For instance, Chen (1995) created diagrams that suggested a particular implementation of a solution. He found that if a test riddle involved the same solution suggested by a diagram but required a change in implementation details, performance would drop. This drop was in addition to any effects due to surface features or conceptual similarity between the diagram and the riddle. Novick and Holyoak (1991) found a similar result with algebra word problems. In their study, learners would recognize the relevance of a particular procedure learned from an example for a new problem, but had difficulty figuring out how to modify the procedure for the problem. Subgoals can aid this adaptation process.

The second feature from the previous scenario that is relevant to the current article is that a learner might form a generalization *before* attempting a test problem (Ahn, Brewer, & Mooney, 1992). If examples are constructed in such a way as to encourage the learner to discover the higher level features, then the solution procedure might be organized around these features before the learner encounters a new problem (Catrambone, 1996; Clement, Mawby, & Giles, 1994; cf. Ross & Kennedy, 1990).

The present article takes the stand that appropriately designed examples can lead learners to form more generalized solution procedures such as ones organized around subgoals. The methods learned to achieve those subgoals will possibly need to be adapted when the learner encounters new problems. A learner will have a better chance of adapting a solution procedure that is organized by subgoals, and methods for achieving those subgoals, compared to a solution procedure that consists only of a long series of steps.

Overview of Experiments

In the materials used in the current study, the ultimate goal of each problem is to calculate a probability. The solution procedure for achieving this goal involves a number of steps, a subset of which constitutes a sequence of multiplication and addition operations that can be grouped under the subgoal: "find the total frequency of the event."

Consider the *No Label* solution to the probability example in Table 1 involving the Poisson distribution.[2] A learner could study this example and memorize the steps for solving a problem that involves the same set of steps even if the new problem involved farmers and tractors instead of lawyers and briefcases. For example, after studying the No Label solution, the learner's knowledge for the part of the solution procedure that involves finding λ, the average, might be represented as

Goal: Find λ
Method:
1. Multiply each category (e.g., owning exactly zero briefcases, owning exactly one briefcase, etc.) by its observed frequency.
2. Sum the results.
3. Divide the sum by the total number of lawyers to obtain the average number of briefcases per lawyer.

This representation would be effective for problems that involve calculating the average in the same way as the example. However, this representation fails to capture the fact that the first line of the No Label solution in Table 1 also involves calculating a total frequency. A novel problem that requires finding total frequency in a different way than in the example might cause problems for the learner using the previous representation. For instance, consider the problem in Table 2. In this problem the total frequency is calculated by adding a set of simple frequencies. This is a less complex method than was used in the example, but the learner might not be able to construct it because the subgoal for finding the total frequency, and an instance of a method for achieving it, were never isolated. If the learner had formed the following representation, then his or her chance of solving the first problem in Table 2 might be better.

Goal: Find λ
Method:
1. *Goal: Find total number of briefcases*
   Method:
   a. Multiply each category by its observed frequency.
   b. Sum the results to obtain the total number of briefcases.
2. Divide the total number of briefcases by the total number of lawyers to obtain the average number of briefcases per lawyer.

This representation identifies the subgoal of finding the total and isolates the steps involved in achieving that subgoal, thus the learner can focus on modifying that subset of steps.

Catrambone (1995, 1996) found that learners studying the *Superficial* or *Abstract Label* solutions (see the second and third solutions in Table 1) to the example were more likely than No Label learners (who studied the first solution in Table 1) to find the total frequency as measured by their

[2] The Poisson distribution is often used to approximate binomial probabilities for events occurring with some small probability. The Poisson equation is $P(X = x) = [(e^{-\lambda})(\lambda^{X})]/x!$, where λ is the average (the expected value) of the random variable X.

Table 1
Training Example With Solutions Using Superficial, Abstract, or No Labels

A judge noticed that some of the 219 lawyers at City Hall owned more than one briefcase. She counted the number of briefcases each lawyer owned and found that 180 of the lawyers owned exactly one briefcase, 17 owned two briefcases, 13 owned three briefcases, and 9 owned four briefcases. Use the Poisson distribution to determine the probability of a randomly chosen lawyer at City Hall owning exactly two briefcases.

No label

$$E(X) = \frac{1(180) + 2(17) + 3(13)+4(9)}{219} = \frac{289}{219}$$

$= 1.32 = \lambda =$ average number of briefcases owned per lawyer

$$P(X = x) = \frac{[(e^{-\lambda})(\lambda^x)]}{x!}$$

$$P(X = 2) = \frac{[(2.718^{-1.32})(1.32^2)]}{2!} = \frac{(.27)(1.74)}{2} = .235$$

Superficial label

$$E(X) = \frac{1(180) + 2(17) + 3(13) + 4(9)}{219} = \frac{\text{total number of briefcases owned}}{219} = \frac{289}{219}$$

$= 1.32 = \lambda =$ average number of briefcases owned per lawyer

$$P(X = x) = \frac{[(e^{-\lambda})(\lambda^x)]}{x!}$$

$$P(X = 2) = \frac{[(2.718^{-1.32})(1.32^2)]}{2!} = \frac{(.27)(1.74)}{2} = .235$$

Abstract label

$$E(X) = \frac{1(180) + 2(17) + 3(13) + 4(9)}{219} = \frac{\Omega}{219} = \frac{289}{219}$$

$= 1.32 = \lambda =$ average number of briefcases owned per lawyer

$$P(X = x) = \frac{[(e^{-\lambda})(\lambda^x)]}{x!}$$

$$P(X = 2) = \frac{[(2.718^{-1.32})(1.32^2)]}{2!} = \frac{(.27)(1.74)}{2} = .235$$

success at solving problems such as the first one in Table 2.[3] This was taken as initial evidence that the former groups tended to learn the subgoal to find the total frequency, whereas the latter group did not.

The present set of experiments extend the prior work by testing implications of the subgoal learning model. Experiment 1 tests the prediction that background knowledge should influence how successfully a learner can self-explain the purpose for a set of grouped steps and thus, how likely the learner is to form the subgoal for those steps. Experiment 2 seeks to test whether subgoal learning affects transfer after a delay between training and testing because prior work has found that learners tend to be less successful solving novel problems after a delay between training and testing (e.g., Nisbett, Fong, Lehman, & Cheng, 1987). Experiment 3 explores the effect of subgoal learning on learners' reliance

[3] In prior studies (e.g., Catrambone, 1995, 1996) the Superficial Label and Abstract Label solutions were referred to as *Meaningful Label* and *Less Meaningful Label* solutions, respectively. However, the terms *Superficial* and *Abstract* seem to better capture essential characteristics of the labels as they are used in these studies.

Table 2
Sample Test Problems

Total frequency calculated by adding simple frequencies

Over the course of the summer, a group of five kids used to walk along the beach each day collecting seashells. We know that on Day 1 Joe found four shells, on Day 2 Sue found two shells, on Day 3 Mary found five shells, on Day 4 Roger found three shells, and on Day 5 Bill found six shells. Use the Poisson distribution to determine the probability of a randomly chosen kid finding three shells on a particular day.

Solution (Not Seen by Participants)

$$E(X) = \frac{4 + 2 + 5 + 3 + 6}{5} = \frac{20}{5} = 4.0 = \lambda = \text{average number of shells per kid}$$

$$P(X = 3) = \frac{[(2.718^{-4.0})(4.0^3)]}{3!} = \frac{(.018)(64)}{6} = .195$$

Total frequency provided directly

A number of celebrities were asked how many commercials they made over the last year. The 20 celebrities made a total of 71 commercials. Use the Poisson distribution to determine the probability that a randomly chosen celebrity made exactly 5 commercials.

Solution (Not Seen by Participants)

$$E(X) = \frac{71}{20} = 3.55 = \lambda = \text{average number of commercials per celebrity}$$

$$P(X = 5) = \frac{[(2.718^{-3.55})(3.55^5)]}{5!} = \frac{(.029)(563.8)}{120} = .135$$

on surface features when they attempt to transfer example solution procedures to test problems. Finally, Experiment 4 provides a replication of Experiment 3 and tests whether generalizations occur primarily during the study of examples or when the examples are applied to an initial test problem, or both. It also uses a talk aloud procedure in order to provide converging evidence for subgoal learning.

Experiment 1

The aim of Experiment 1 was to test an implication of the subgoal learning model that if a learner receives a label as a cue to group a set of steps, the likelihood of the learner forming an appropriate subgoal is dependent upon the semantic content of the label and the learner's background knowledge. Although prior studies (e.g., Catrambone, 1996) have manipulated semantic content as a way of influencing subgoal learning, the effects of learners' backgrounds have not been examined.

Eylon and Reif (1984) found that learners, except for those of low academic achievement, benefited from material that was hierarchically well organized. On the other hand, Dufresne et al. (1992) found that learners who tended to offer surface feature-based explanations for similarity judgments about physics problems—and therefore were presumably students of lower achievement—were the ones who benefited most from training that focused on teaching a hierarchical approach to classifying problems.

It may be the case that learners with weaker math backgrounds need labels with connections to surface features of examples because they provide additional guidance for the self-explanation process. Learners with a weak math background might look at a series of addition and multiplication steps for the No Label solution in Table 1, that is

$$1(180) + 2(17) + 3(13) + 4(9),$$

and fail to group them. Because of this they might not perform the self-explanation that leads to the realization that those steps calculate a total. A superficial label such as *total number of briefcases owned* should cue the grouping and aid self-explanation whereas an abstract label such as Ω might not provide sufficient guidance for self-explanation. Conversely, for learners with a stronger background, an abstract label such as Ω should cue the grouping and the learner will be more likely to have the ability to successfully self-explain the purpose of the steps.

The previous argument leads to the prediction that participants with a stronger background should be equally likely to learn the subgoal to find a total in the Superficial and Abstract Label conditions. The Superficial Label and Abstract Label solutions (see Table 1) label the steps for finding the total frequency rather than merging them with the overall set of steps for finding the average. Learners with a strong background who study example solutions with either

superficial or abstract labels are predicted to find λ correctly on the novel test problems about equally often and should outperform the No Label group. However, for participants with a weaker math background, the semantic content of the label is predicted to play a larger role in helping them to understand the purpose of the labeled steps. For these learners, those receiving the superficial label should be more likely to form the subgoal to find a total than those receiving the abstract label. Therefore, participants with a weaker background in the Superficial Label condition should be more successful than similar students in the Abstract and No Label conditions at finding λ on the novel problems.

Subgoal learning in Experiment 1 was assessed through problem-solving performance as well as learners' descriptions of how to solve problems in the domain. It was hypothesized that if a person learns a subgoal, such as find a total, then he or she would be more likely to solve novel problems as well as to mention the subgoal in a description compared to a learner who did not learn that subgoal. These results would be found presumably because the solution procedure would be more hierarchically organized compared to one that was simply a list of steps.

Method

Participants. Participants were 150 students recruited from several Atlanta-area colleges who received course credit or payment for their participation. In order to participate in the experiment, a student either had to have taken no college-level calculus courses or to have had between two and four college-level mathematics courses beyond introductory calculus.[4] Additionally, students who had taken a prior probability course were excluded from the experiment.

Materials and procedure. All participants initially studied a cover sheet that briefly described the Poisson distribution and how it could be used as a replacement for more cumbersome techniques for calculating probabilities involving events that could be categorized as successes and failures. The Poisson equation was presented along with a simplified notion of a random variable.

Participants were randomly assigned to one of three conditions with the constraint that of the 50 participants per condition, 25 had a stronger (calculus) math background and 25 had a weaker (no calculus) math background. The Superficial Label group studied three examples demonstrating the weighted average method for finding λ in which the steps for finding the total frequency were given a label that was assumed to have meaning to the participants and made mathematical sense given the steps that preceded it (see the Superficial Label solution in Table 1 for an example). The Abstract Label group studied examples in which the steps for finding the total frequency were labeled with Ω, which was assumed to have little meaning for the participants in the context of the examples (see the Abstract Label solution in Table 1). The No Label group studied examples in which the steps for finding the total frequency were not labeled (see the No Label solution in Table 1).

After studying the examples, participants were asked to describe how to solve problems in the domain. The instructions were

> Suppose you were going to teach someone how to solve Poisson distribution problems of the types you have just studied. Please describe the procedure or procedures you would give someone to solve these problems. Please be as complete as possible.

After writing their descriptions, participants solved six problems. The first two were isomorphic to the example in Table 1; that is, they required the use of the same step-by-step procedure (weighted average method) for finding λ. These problems were given first so that participants would be able to immediately see that the prior examples were relevant for solving the test problems. The next four problems required new ways of finding the total frequency: either by recognizing that the value was given directly in the problem (see the second problem in Table 2 for an example) or by adding simple frequencies (see the first problem in Table 2). Participants were told not to look back at the examples when writing their descriptions or solving the test problems.

Two raters independently scored the explanations and agreed on scoring 90% of the time. Disagreements were resolved by discussion.

Design. The independent variables were math background (stronger [calculus] or weaker [no-calculus]) and type of example solutions studied (Superficial Label, Abstract Label, No Label), resulting in six experimental groups. The dependent measures were descriptions for how to solve the problems and transfer performance on the six test problems.

Results

Descriptions of how to solve problems. Because it was hypothesized that there would be an interaction between label type and math background, the following predictions were made. First, for learners with stronger math backgrounds, those receiving a label would be more likely to mention the notion of finding a total compared to No Label participants. The Abstract Label group might mention the notion of finding "Ω" rather than finding the "total number" of things if there was some tendency by learners to repeat the wording from examples. Second, for learners with weaker math backgrounds, it was predicted that Superficial Label participants would be more likely than the other two groups to mention the notion of finding a total.

Participants in the Abstract Label condition were the only ones to mention Ω. This makes sense given that this term is arbitrary and did not appear in the solutions seen by the other two groups. However, a frequency analysis measuring the association between condition and whether a participant mentioned the notion of finding a total, Ω, or neither, would be inflated because two of the groups would have zero frequency for mentioning Ω. To compensate for this, participants in the Abstract Label condition who mentioned the notion of finding Ω, even if they did not also explicitly call this value a total, were counted in Table 3 as having mentioned a total.

On the basis of the scheme described previously, all participants were categorized into one of two groups: those mentioning a total (and/or Ω) in their descriptions versus those mentioning neither. If participants are collapsed across label condition, those with a stronger math background mentioned total/Ω more often than those with a weaker background, $\chi^2(2, N = 150) = 7.11, p = .008$ (see Table 3). If only participants with a stronger math background are

[4] There is no theoretical justification for using calculus experience as a discriminator of math background. However, discussions with mathematics professors at local colleges suggest that it might be appropriate.

Table 3
Percentage of Participants Mentioning Finding a Total in Their Descriptions as a Function of Label and Math Background (Experiment 1)

Math background	Group: Superficial label	Group: Abstract label	Group: No label	*M*
Stronger	76	56	20	51
Weaker	44	32	12	29
M	60	44	16	40

Note. $N = 25$ for each cell.

considered, there was a significant difference among the groups in the frequency of mentioning total/Ω, $\chi^2(2, N = 75) = 16.11$, $p = .0003$. Pairwise comparisons indicate that both label groups mentioned total/Ω more often than the No Label group (vs. Superficial: $\chi^2(1, N = 50) = 15.70$, $p = .0001$; vs. Abstract: $\chi^2(1, N = 50) = 6.88$, $p < .009$), whereas there was no reliable difference between the two label groups, $\chi^2(1, N = 50) = 2.23$, $p = .14$. If only participants with a weaker math background are considered, there was again a significant difference among the groups in the frequency of mentioning total/Ω, $\chi^2(2, N = 75) = 6.30$, $p = .04$. Pairwise comparisons indicate that the Superficial Label group mentioned total/Ω more often than the No Label group, $\chi^2(1, N = 50) = 6.35$, $p < .02$, whereas there was a marginal difference between the Abstract Label and No Label groups, $\chi^2(1, N = 50) = 2.91$, $p = .083$. There was no reliable difference between the two label groups, $\chi^2(1, N = 50) = 0.76$, $p = .38$.

Transfer. Participants were given a score of 1 for a given problem if they found λ correctly and a score of 0 otherwise. The scores for the two problems that were isomorphic to the training examples, Problems 1–2, were summed, creating a score from 0 to 2 for performance on those problems. Similarly, the scores for the four novel problems, Problems 3–6, were summed, creating a score from 0 to 4 for performance on those problems.

All participants except six in the Superficial Label group, six in the Abstract Label group, and four in the No Label group solved both isomorphic problems correctly.

A two-way analysis of variance on performance across the four novel test problems was carried out with label condition and math background as the factors. Table 4 presents the average scores on these problems as a function of group. There was a significant effect of label type, $F(2, 144) = 13.60$, $p < .0001$, $MSE = 3.13$, and math background $F(1, 144) = 13.64$, $p = .0003$. There was also a significant interaction between these factors, $F(2, 144) = 3.31$, $p = .039$. The most typical mistake that students made on these problems was to write in the solution area that not enough information was given to solve the problem.

Table 4
Number of Novel Test Problems Solved Correctly as a Function of Label and Math Background (Experiment 1)

Math background	Group: Superficial label	Group: Abstract label	Group: No label	*M*
Stronger	3.04	2.88	1.44	2.45
Weaker	2.72	0.80	0.64	1.39
M	2.88	1.84	1.04	1.92

Note. $N = 25$ for each cell. Maximum possible score for any cell = 4.

If only participants with a stronger background are considered, pairwise comparisons indicated that the two label groups did not perform differently ($p > .7$). The Superficial Label group outperformed the No Label group ($p = .004$; required $p = .008$ using Shaffer (1986) sequential Bonferroni pairwise comparisons for providing a familywise α of .05 for multiple comparisons; see also Seaman, Levin, & Serlin, 1991). The Abstract Label group showed a tendency to outperform the No Label group, but the difference was not reliable ($p = .01$; required $p = .008$). However, given the directional prediction that the Abstract Label group would outperform the No Label group, it would be appropriate to conduct a one-tailed test. Doing this, the difference between the groups produces a p value of .005 which is less than the required value of .008. These results are consistent with those obtained in prior studies (Catrambone, 1995, 1996).

If only participants with a weaker background are considered, the Superficial Label group outperformed the other two groups (both $ps < .0005$ with required $p = .005$). There was no significant difference between the Abstract and No Label groups ($p > .7$).

Transfer to novel problems as a function of descriptions. In order to produce a strict test of the relationship between descriptions and transfer in novel problems, No Label participants were excluded from the analysis because it was demonstrated that these participants performed differently than the label groups in transfer. Considering only label participants then, those who mentioned total/Ω ($n = 52$) transferred more successfully than those who mentioned neither ($n = 48$), $F(1, 98) = 97.30$, $p < .0001$, $MSE = 2.96$, $Ms = 3.30$ and 1.33, respectively. If this analysis is repeated separately for participants with stronger and weaker math backgrounds, similar results are obtained. For participants with stronger math backgrounds, those mentioning total/Ω ($n = 33$) outperformed those who did not ($n = 17$), $F(1, 48) = 44.40$, $p = .0001$, $MSE = 2.28$, $Ms = 3.64$ and 1.65, respectively. For participants with weaker math backgrounds, it was also found that those mentioning total/Ω ($n = 19$) outperformed those who did not ($n = 31$), $F(1, 48) = 29.24$, $p = .006$, $MSE = 3.50$, $Ms = 2.74$ and 1.16, respectively.

Discussion

The purpose of Experiment 1 was to test an implication of the subgoal learning model that a learner's ability to form an appropriate subgoal—or to adapt a method for achieving that subgoal in a novel problem—is dependent upon the semantic content of the label and the learner's background

knowledge. The present experiment found that the semantic content of the label appeared to have less effect on subgoal formation for learners with a stronger math background compared to learners with a weaker background. Although learners of both backgrounds showed a decline in the likelihood of mentioning the subgoal of finding a total as a function of whether the label was superficial versus abstract versus not present, the statistical effects were different. For learners with a stronger math background, both label groups outperformed the No Label group as measured by the description and transfer tasks. This result is consistent with prior studies that have used learners with strong math backgrounds (Catrambone, 1995, 1996). However, for learners with a weaker math background, the results changed. Learners in the Superficial Label condition mentioned the subgoal of finding a total in their descriptions more often than learners in the No Label condition, whereas the frequency difference was less striking for Abstract versus No Label participants. In addition, while Superficial Label participants showed superior transfer compared to the No Label participants on the novel problems, there was no difference between the Abstract and No Label groups.

Overall the results suggest that the presence of a label, and not necessarily its semantic content, can make a learner with an appropriate background more likely to successfully attempt to self-explain the purpose of a set of steps. In addition, the likelihood of a learner being able to adapt a set of steps for achieving the subgoal appears to be affected by the learner's background. This is consistent with Novick and Holyoak's (1991) finding that learners' mathematical background affected their success at adapting procedures. For learners with a weaker background, the semantic content of the label appears to play a role in subgoal formation, possibly by constraining the self-explanation process. For stronger learners, such a "crutch" is less important.

Although the results of this experiment indicate that a learner, at least one with appropriate background knowledge, *can* form an appropriate subgoal, and adapt a set of steps, on the basis of an abstract cue in an example solution, it is reasonable to ask why a teacher or textbook writer would choose to use such a cue. That is, would it not be sounder educational practice to provide the learner with a label tied to problem features that helps learners form a subgoal more easily?

There are at least two reasons why the use of an abstract label might aid performance on novel problems more than a label tied to surface features, at least under certain circumstances. First, consider the self-explanation process postulated by the subgoal-learning model as the learner attempts to determine the purpose of a set of steps. The cue is assumed by the model to help the learner realize that the steps go together. The learner must then determine the steps' collective purpose. The learner presumably uses at least two sources of information in order to do this. One source is background knowledge as suggested by the Experiment 1. The second source is the semantic information in the cue itself. Consider the cue in the Superficial Label solution in Table 1. This label states that a set of steps calculates the total number of *briefcases* owned. Given this information, the learner might form the subgoal to calculate the total number of "things" (that is, inanimate objects as opposed to people). This is a too restrictive subgoal because, as will be shown later, one can easily construct problems that require calculating other totals such as the number of people. A learner who had seen the more abstract label might form the subgoal to find a total without necessarily tying it to objects. This subgoal is more general and closer to being formally correct (the most formal view would be "total frequency of the event").

The second reason why an abstract label could aid performance more than a superficially connected label is the assumption that an abstract label would require the learner to work harder to form a subgoal to explain the purpose of the steps than if the label were tied to the surface features of the problems. The extra effort required might help the learner integrate the subgoal and steps with prior knowledge, thus making the use of that information more flexible (Chi et al., 1994). It might also help learners to better remember and adapt this information when tested on novel problems after a delay (cf. McDaniel & Schlager, 1990).

Experiments 2 and 3 tested these implications of the subgoal learning model. The focus of Experiment 2 was to examine whether the hypothesized extra effort required to form subgoals from abstract labels will lead to superior transfer performance after a delay. The focus of Experiment 3 was to examine the notion that abstract labels can aid problem solving transfer when the roles of surface features in test problems are changed from their roles in training examples.

Experiment 2

Experiment 1 suggested one reason why labels might be used in example solutions: to help learners form solution procedures organized by subgoals. A reason to use abstract labels might be to induce learners to work harder to determine the purpose of a set of steps, thereby increasing the likelihood they will access background knowledge to form the subgoal. This interpretation is consistent with the findings of Catrambone and Holyoak (1989), who observed that participants who answered a set of fairly directive comparison questions about a set of stories—questions designed to help people focus on the crucial structural features—were much more likely to access and apply/adapt the solution from those stories to a new problem even after a week's delay compared to people who did not answer the questions.

The effort required to self-explain the purpose of a set of steps is assumed to be greater when a learner receives an example solution using abstract labels. Labels tied to surface features might encourage the learner to take a short-cut to explain the steps' purpose (i.e., the label almost becomes the explanation), leading the learner to form a subgoal that is potentially misleading. In addition, self-explanations can fill gaps in a text (McNamara & Kintsch, 1996) as well as help a learner to integrate the new knowledge with preexisting knowledge, thereby making the new information more memorable and accessible (Chi et al., 1994). Abstract labels

presumably encourage or require more gap filling than superficial labels and thus, can lead to more integration and memorableness of the subgoal and steps (at least for learners with the background knowledge to enable the gap filling). The result should be superior performance on novel problems after a delay between training and testing. Experiment 2 was designed to test this possibility. Although effort per se is not directly measured, it is assumed, other things being equal, that learners who expend more effort on self-explanation will be more likely to remember the information after a delay compared to those who expend less effort.

Learners with "stronger" math backgrounds were used in this and subsequent experiments. The rationale for this choice was based on the following considerations: (a) because students are typically required to possess the appropriate background in order to take a particular course, it seems reasonable to focus on students with stronger math backgrounds in experiments using probability materials which are often considered quite challenging; (b) this and subsequent experiments use a variety of manipulations and the number of participants required would be doubled if the background factor was included each time; (c) it was easier to find students with stronger math backgrounds at Georgia Tech.

Method

Participants. Participants were 180 students recruited from introductory psychology classes at the Georgia Institute of Technology who received course credit for their participation. None of the students had taken a probability course prior to participating in the experiment, but all had at least one college-level calculus course.

Materials and procedure. Participants studied the same cover sheet and examples as in the prior experiment and were randomly and evenly assigned to one of three label conditions: Superficial Label (n = 60), Abstract Label (n = 60), and No Label (n = 60). Within each condition half the participants were tested immediately after studying the examples, whereas the other half were tested one week later. Participants in the delay condition were told that one of the tasks they would be doing when they returned a week later would be to solve problems related to the ones they studied during the first session and because of this, they should do their best to remember what they learned.

Participants solved the same six test problems used in Experiment 1 and were scored in the same way.

Design. The independent variables were type of example solution studied (Superficial Label, Abstract Label, No Label) and timing of the posttest (delay, no delay), thus there were six groups in the experiment. The dependent measure was performance on the six test problems.

Results and Discussion

Test problems were scored as in the prior experiment. Relatively few participants failed to solve the isomorphic problems. A total of 11 participants failed to solve either one or both of the isomorphs; seven of these participants were in the delay condition.

There were significant effects due to type of label and delay on performance on the novel problems (see Table 5). Participants receiving a label solved the novel problems more successfully than No Label participants, $F(2, 174) = 5.41, p = .005, MSE = 2.98$. Participants tested immediately performed better than those tested after a delay, $F(1, 174) = 9.67, p = .002$. The interaction of label and delay was also significant, $F(2, 174) = 9.05, p = .05$. The interaction, coupled with the results in Table 5, suggest that participants receiving superficial labels in the examples showed a larger drop in performance going from no delay to delay whereas the decrement was less for learners in the Abstract Label condition.

Table 5
Number of Novel Test Problems Solved Correctly as a Function of Label and Delay (Experiment 2)

	Group			
Delay	Superficial label	Abstract label	No label	*M*
No	3.47	3.17	2.33	3.02
Yes	1.80	3.00	1.77	2.22
M	2.63	3.18	2.05	2.62

Note. N = 30 for each cell. Maximum possible score for any cell = 4.

In order to investigate the previous possibilities more closely, Shaffer (1986) sequential Bonferroni pairwise comparisons (familywise $\alpha = .05$) were conducted. For participants in the No Delay condition, both Abstract and Superficial Label participants outperformed No Label participants (both ps less than the required .008, one-tailed). There was no significant difference between the Abstract and Superficial Label conditions ($p = .5$). For participants in the Delay condition, the Abstract Label participants outperformed the Superficial and No Label participants (both ps less than the required .005, one-tailed). There was no significant difference between the Superficial and No Label conditions ($p = .94$).

Another way of examining the relative effects of delay on the different label conditions is to compare performance across delay separately for each label condition. For participants in the Superficial Label condition, there was a significant difference on performance on novel test problems as a function of delay, $F(1, 58) = 16.30, p = .0002, MSE = 2.56$. Conversely, for participants in the Abstract Label condition, there was not a significant effect due to delay, $F(1, 58) = 0.16, p = .69$.

These results indicate that learners who presumably had to work harder to form a subgoal to represent the purpose of a set of steps were more likely to remember the subgoal—as measured by problem-solving performance—when they were tested after a delay compared to those who were assumed to not have had to work as hard. Learners who were less likely to be induced to form a subgoal transferred less successfully both before and after a delay. These results are consistent with the findings of Catrambone and Holyoak (1989) who noted that those learners who produced better written descriptions of the commonalities of example analogs were also the ones who were more likely to import the solution from those analogs to a target problem. In both the

present study and in Catrambone and Holyoak's study, learners who extracted the common structure of the training examples (in the present case the common structure would be the subgoals) were more likely to successfully apply them to target problems in which the details differed.

Experiment 3

Most participants in Experiment 1 who mentioned finding a total in their descriptions described the total in terms of objects or things: 87% in the Superficial Label condition and 82% in the Abstract Label condition. Although the percentages are similar for the two label conditions—keeping in mind that more Superficial Label than Abstract Label participants mentioned the idea of a total at all—it is possible that this surface feature connection was strongest for participants in the Superficial Label condition because the label in this condition explicitly mentioned objects from the problems. One implication of forming a subgoal that is tied to surface features is that the learner is confusing surface and structural features of the domain. A way to test this possibility is to construct test problems that systematically manipulate the relationship between surface and structural features and observe the degree to which the features guide learners' performance.

For instance, Ross (1987, 1989) provided students with various types of probability examples to study such as problems dealing with permutations and combinations. The permutation examples involved people picking objects in a certain order (e.g., scientists, in order of seniority, picking from a pool of computers at random). Because the examples involved *people* picking *objects*, the number of objects provided the starting value for the denominator in the permutation equation. Some of the test problems involved people being assigned to objects (e.g., as a particular computer is unpacked, a randomly chosen scientist is assigned to use it). In these cases the number of people in the problem provided the starting value for the denominator. However, students typically placed the number of objects in the denominator. Students appeared to confuse the surface features of objects and humans with domain-relevant features such as the set of possible choices (see also Bassok et al., 1995).

With respect to the experimental materials used in the present study, most learners, at least at the college level, were assumed to be sufficiently sophisticated to generalize "total number of briefcases" (and tractors and guitars; the objects in the other two training examples). The generalization that might be formed though was unclear. One possibility was that the generalization would be "total number of objects" if all the examples involved humans using objects. Learners forming this generalization would be predicted to be more successful solving novel problems that require the total number of objects to be calculated in new ways compared to learners not forming this generalization. However, given that this generalization is still tied to a surface feature, objects, these learners might fail to solve correctly a novel problem that required the number of humans, rather than objects, for the total. Learners studying examples with the abstract label who form the subgoal for finding a total might be less likely to have this subgoal tied to a surface feature. As a result, these learners would be less likely to make mistakes on novel problems that switch the roles of humans and objects from their roles in the training examples.

In Experiment 3 all participants studied examples demonstrating the weighted average method for finding λ (see Table 1 for an example) and saw either the Superficial Label, Abstract Label, or No Label solution. Performance predictions varied as a function of training condition and type of test problem.

The first four test problems were isomorphic to the training examples. The first pair involved calculating the total number of objects in order to find λ and the second pair involved calculating the number of people. Participants in all conditions were expected to solve the first pair with little difficulty. Conversely, the second pair of isomorphs were expected to give participants difficulty because they involved a reversal in the roles of humans and objects (see the first problem in Table 6 for an example). These problems provided the numbers needed to calculate a total based on number of people. In order to make it possible for participants to solve this problem incorrectly, a second set of numbers was also present in these two problem statements that allowed one to calculate the total number of objects as was done in the training examples. In the first problem in Table 6, the correct way to calculate the total is to multiply each category of cab (e.g., those driven by just one driver, those driven by just two drivers, etc.) by the number of cabs that fell into each category. This approach provides the total number of drivers (who had driven cabs) and thus allows the calculation of average number of drivers per cab. This ultimately allows one to calculate the probability of a randomly chosen cab being driven by a certain number of drivers.

The incorrect approach to the first problem in Table 6 would be to take each category of driver (e.g., those that had driven just one cab, those that had driven two cabs, etc.) and multiply it by the number of drivers that fell into each category. This would allow the calculation of the average number of cabs per driver which would ultimately allow one to calculate the probability of a randomly chosen driver having driven a certain number of cabs.

It was expected that when confronted with these two sets of numbers in the second pair of isomorphs, many participants would choose the set (which always came first) that allowed one to multiply categories of people by the number of people in each category. This was the approach used in the training examples for finding the total number of objects. It was predicted that on the second pair of isomorphs the Superficial Label and No Label participants would be more likely than Abstract Label participants to calculate, inappropriately, the total number of objects in order to find λ. This prediction was made because Superficial Label participants were expected to be more likely than Abstract Label participants to associate objects with finding a total and the No Label participants would simply be attempting to repeat the steps from the examples.

Table 6
Additional Sample Test Problems (Experiments 3 and 4)

Weighted average with humans providing total frequency

The manager of a large taxicab company took a survey of some of the cab drivers and found that each driver had one or two specific cabs that they preferred to drive when possible. The manager found that, of the 29 drivers surveyed, 7 of them had managed to drive a single cab each time they worked, whereas 12 of the drivers had driven 2 different cabs and 10 of the drivers had driven 3 different cabs. The manager also examined some of the cabs and found that of the 29 cabs examined, 4 of them had been driven by only 1 driver, 9 of the cabs had been driven by 2 different drivers, 5 of the cabs had been driven by 3 different drivers, and 11 of the cabs had been driven by 4 different drivers. Use the Poisson distribution to determine the probability that a randomly chosen cab had been driven by exactly 2 different drivers.

Solution (Not Seen by Participants)

$$E(X) = \frac{1(4) + 2(9) + 3(5) + 4(11)}{29} = \frac{81}{29} = 2.79 = \text{average number of drivers per cab}$$

$$P(X = 2) = \frac{[(2.718^{-2.79})(2.79^2)]}{2!} = \frac{(.061)(7.78)}{2} = .237$$

Total frequency given directly with humans providing total frequency

Over a period of time at a certain video store, 243 people rented 104 different videos. Use the Poisson distribution to determine the probability that a randomly chosen video was rented exactly four times.

Solution (Not Seen by Participants)

$$E(X) = \frac{243}{104} = 2.34 = \lambda = \text{average number of renters per video}$$

$$P(X = 4) = \frac{[(2.718^{-2.34})(2.34^4)]}{4!} = \frac{(.096)(29.98)}{24} = .12$$

Total frequency calculated by adding simple frequencies of humans

An accounting firm employing many accountants worked on a large number of tax returns and used many types of tax forms. Four of the accountants were interviewed, and it was found that 1 worked on three types of tax forms that day, another worked on nine, a third worked on five, and the 4th worked on six. In addition, of the many different types of tax forms used, it was found that one type of tax form was used by 12 accountants at the firm, another type was used by 8 accountants, a third type was used by 6 accountants, and a fourth type was used by 9 accountants. Use the Poisson distribution to determine the probability of a randomly chosen type of tax form being worked on by 7 different accountants.

Solution (Not Seen by Participants)

$$E(X) = \frac{12 + 8 + 6 + 9}{4} = \frac{35}{4} = 8.75 = \lambda = \text{average number of accountants per form}$$

$$P(X = 7) = \frac{[(2.718^{-8.75})(8.75^7)]}{7!} = \frac{(.00016)(3926960)}{5040} = .125$$

The role reversal described previously was also expected to affect performance on the transfer problems that involve a change in steps. Predicted performance on the novel problems varied as a function of group and role correspondence. Four of the novel problems provided the total frequency directly. Two of the these problems involved objects providing the total event frequency (such as the second problem in Table 2), whereas the other two involved humans providing the total event frequency (see the second problem in Table 6 for an instance of the latter type). It was predicted that on the first pair of problems both label groups would outperform the No Label group. It was predicted that on the second pair of problems Superficial Label participants would be more likely than Abstract Label participants to incorrectly place the number of objects in the numerator of the fraction in order to find λ. This prediction was made because of the

hypothesis that Superficial Label participants would be more likely than Abstract Label participants to associate objects with finding a total. The value placed in the numerator of the fraction presumably represents the value that the participant believes is the total.

The last four problems involved adding simple frequencies in order to find a total frequency. Two of the problems involved objects being used to calculate the total event frequency (such as the first problem in Table 2) and two involved humans being used to calculate the total event frequency (see the third problem in Table 6 for an example of the latter type). It was predicted that on the first pair of problems the label groups would outperform the No Label group. The second pair of problems provided two sets of numbers. The first set of numbers could be added to produce a total number of objects and the second set could be added to produce a total number of people. If only one set was provided, then participants would be more procedurally constrained and there would be less of a chance of finding a performance difference between Superficial and Abstract Label participants. It was predicted that on the second pair of problems the Superficial Label participants would be more likely than Abstract Label participants to make the mistake of calculating a total using objects rather than humans.

Thus, for reversed correspondence (relative to the examples) novel test problems, as well as the reversed correspondence isomorphs, it was predicted that Abstract Label participants would show less of a decrement in performance than Superficial Label participants, relative to the groups' performance on the same correspondence problems. No Label participants were expected to do poorly on all novel test problems although performance was predicted to be worse on the problems featuring reversed role correspondence unless there was a floor effect.

Method

Participants. Participants were 90 students recruited from an introductory psychology class at the Georgia Institute of Technology who received course credit for their participation. None of the participants had taken a probability course prior to participating in the experiment, but all had at least one college-level calculus course and therefore could be considered to have a "stronger" math background.

Materials and procedure. Participants studied the same cover sheet as in Experiment 1.

Participants were randomly assigned to one of three groups (n = 30 per group). The Superficial Label, Abstract Label, and No Label groups studied the same three examples as the corresponding groups in Experiments 1 and 2.

After studying the examples, participants solved 12 test problems. The first four test problems were isomorphic to the training examples, that is, they required the weighted average method for finding λ. The first two involved objects in the total frequency and the next two involved humans in the total frequency (see the first problem Table 6 for an example of the latter set). The next four test problems involved the total frequency being given directly in the problem; the average could be found by simply dividing the given total frequency by the total number of trials. The first pair of the set involved objects in the total frequency (see the second problem in Table 2 for an example), whereas the second pair involved humans in the total frequency (see the second problem in Table 6 for an example). The next four test problems involved calculating the total frequency by adding a set of simple frequencies. The first pair of the set involved objects in the total frequency (see the first problem in Table 2 for an example) whereas the second pair involved humans in the total frequency (see the third problem in Table 6 for an example). Participants were told not to look back at the examples when solving the test problems.

Participants' written solutions were scored for whether they found λ correctly.

Design. The between-subjects variable was type of example solutions studied (Superficial Label, Abstract Label, No Label). The within-subjects variable was correspondence of the roles of humans and objects in the test problems to their roles in the examples. The dependent measure was performance on the 12 test problems.

Results and Discussion

As in Experiments 1 and 2, participants were given a score of 1 for a given problem if they found λ correctly and a score of 0 otherwise. The scores for Problems 1 and 2, the two problems that were isomorphic to the training examples and had the same role correspondence of humans and objects as the examples, were summed, creating a score from 0 to 2 for performance on those problems. Similarly, a score from 0 to 2 was calculated for the isomorphs that had a reversed role correspondence of humans and objects (Problems 3 and 4). Finally, a score from 0 to 4 was calculated for the novel test problems with the same role correspondence as the examples (Problems 5, 6, 9, and 10) and a score from 0 to 4 was calculated for the novel test problems with a reversed role correspondence (Problems 7, 8, 11, and 12).

As expected, all groups did quite well at finding λ on test problems that were isomorphs to the training examples and had objects and humans in the same role as in the examples. In fact, all participants solved each of these problems correctly. Performance on the isomorphs with the role reversals produced a different outcome. The average number of problems (out of two) solved by participants in the Superficial Label, Abstract Label, and No Label conditions was 0.80, 1.33, and 0.53, respectively, which was a significant difference, $F(2, 87) = 5.49$, $p = .0057$, $MSE = 0.91$. Pairwise comparisons showed that the Abstract Label group outperformed the Superficial Label group and the No Label group (both $ps < .04$), whereas the latter two groups did not differ ($p > .28$).

Table 7 presents the groups' performance on the novel test problems as a function of whether the problems involved the same or reversed role correspondence of humans and objects compared to the training examples. An analysis of variance was carried out on the performance on the novel test problems with group as the between-subjects variable and role correspondence (same as examples vs. reversed from the examples) as the within-subject variable.

There was a significant difference among the three groups with respect to finding λ on the novel test problems, $F(2, 87) = 6.21$, $p = .003$, $MSE = 5.70$. There was also an effect of role correspondence indicating that problems with reversed role correspondence were solved less successfully

Table 7
Number of Novel Test Problems Solved Correctly as a Function of Label and Role Correspondence (Experiment 3)

Correspondence	Group: Superficial label	Group: Abstract label	Group: No label	*M*
Same	3.07	3.00	1.53	2.53
Reversed	1.67	2.67	1.13	1.82
M	2.37	2.83	1.33	2.18

Note. $N = 30$ for each cell. Maximum possible score for any cell = 4.

than those with the same role correspondence as the examples, $F(1, 87) = 26.56$, $p < .0001$, $MSE = 0.86$. Finally, there was a significant interaction between group and role correspondence, $F(2, 87) = 6.25$, $p = .003$, $MSE = 0.86$, suggesting that the correspondence manipulation affected the groups differently.

Separate analyses were carried out for each group comparing performance on same and reversed role correspondence problems. The Superficial Label group showed a significant decrease in performance on the reversed role correspondence problems compared to the same role correspondence problems, $F(1, 29) = 19.12$, $p = .0001$, $MSE = 1.54$, whereas the Abstract Label and No Label groups did not show significant differences in performance on the problem types.

Consistent with the previous analysis, if performance on only the reversed role correspondence problems is considered, a significant effect of group is found, $F(2, 87) = 5.33$, $p = .007$, $MSE = 3.41$, with pairwise comparisons indicating that the Abstract Label group outperformed the other groups (both $ps < .04$) but the Superficial Label group did not outperform the No Label group ($p > .26$). Conversely, if performance on only the same role correspondence problems is considered, a significant effect of group is again found, $F(2, 87) = 7.17$, $p = .001$, $MSE = 3.14$, but now the pairwise comparisons indicate that the two label groups outperformed the No Label group (both $ps < .002$), whereas there was no difference between the two label groups ($p > .88$).

It was predicted that a typical mistake made by the Superficial Label participants in solving the novel reversed role correspondence problems would be to put or calculate a value for total number of objects in the numerator. One way of examining the likelihood of making this mistake is to examine performance on reversed role correspondence problems by Superficial Label participants who solved the same role correspondence problems correctly. This approach would therefore consider only participants who demonstrated the ability to transfer to problems that involved a change in procedure relative to the training examples. As a result, mistakes on the reversed role correspondence problems would presumably be due to confusion about roles rather than difficulty adapting steps.

Of the eight Superficial Label participants who found λ correctly in all four novel same role correspondence problems, five of them put objects in the numerator for finding λ in at least three of the reversed role correspondence problems. Conversely, of the seven Abstract Label participants who found λ correctly in all four same role correspondence problems, only two of them put objects in the numerator for finding λ in at least three of the reversed role correspondence problems. Although these numbers are too small to achieve statistical significance, their pattern is consistent with the interpretation that the Superficial Label participants who were able to adapt the solution procedure from the examples were more likely to be misled by surface features compared to the analogous Abstract Label participants.

Experiment 3 suggests that the subgoal formed in response to a label may be less closely tied to surface details of examples when the label does not reference those details. Learners receiving examples using abstract labels transferred more successfully than other learners to novel problems that altered, with respect to the examples, the correspondence between surface features and the solution procedure. This result suggests that the generality of the procedure formed from examples can be increased through the use of labels that do not contain references to surface features of examples, at least for learners with appropriate background knowledge.

Experiment 4

One concern from the prior experiments is whether any generalizations formed by learners were actually due to learners' attempts to use the examples to solve the first test problem rather than—or in addition to—being due to generalizations caused by features (i.e., labels) of the examples themselves. Such a concern is motivated by the findings of Ross and Kennedy (1990). In a typical experiment they had learners study four probability principles (e.g., permutations, combinations) that were each illustrated through a worked example. After studying the principles and examples learners attempted to solve eight problems, two for each principle. The first test problem for each principle either did or did not contain a cue indicating which prior training example was relevant for solving the problem. The second test problem for each principle never contained a cue.

Ross and Kennedy (1990) found that when learners received a cue on the first test problem for a particular principle, they were more likely to solve the second test problem for that principle compared to cases in which the first test problem was uncued. More specifically, this benefit manifested itself in terms of an increased likelihood in using the correct principle for the second test problem as well as instantiating the variables correctly. The improved instantiation of variables on the second test problem for each principle was seen in cases in which the roles played by humans and objects in the example were reversed in the first and second test problems relative to the example.

For instance, if the example involved humans picking objects (e.g., scientists choosing computers), the problems would involve objects "picking" humans (e.g., as a particu-

lar computer is unpacked, a randomly chosen scientist is assigned to use it). Learners who were cued to the relevant example when working on the first test problem for a particular principle were more likely to get the roles for humans and objects correct when working on the second test problem for that principle (there was no difference between cued and uncued performance on the *first* test problem with respect to getting the roles correct). Ross and Kennedy (1990) argued that differences between the problem and the cued example led learners to form a generalization as they attempted to apply the example to the problem. This generalization affected performance on the second test problem.

In order to investigate the previous possibility with respect to the materials and procedures used in the present study, the roles of humans and objects in the first two test problems were either the same or reversed from their roles in the training examples in Experiment 4. If the purported subgoal learning effect—as demonstrated by how well learners could handle procedure modification and role reversal—observed across the first three experiments was primarily a function of the labels in the training examples, then performance should be largely a function of type of labels regardless of the roles in the initial test problems. If, however, generalization is substantially affected by the application of an example to the first test problem, then if the roles in the first test problem are reversed from the examples, this might lead learners to form a generalization (at least about roles) that will help subsequent performance on other test problems with role reversals. If the roles in the first test problem are the same as in the examples, then learners might be more likely to form a mental set concerning roles and therefore be less likely to handle role reversals successfully in subsequent problems.

A second concern addressed in Experiment 4 was that the first three experiments provided only indirect support for the claim that a cue such as a label prompts learners to self-explain the purpose of a set of steps. More direct evidence for this claim might be provided by having learners talk aloud while studying the training examples in order to test whether they are in fact spontaneously attempting to explain the purpose of groups of steps. A prior study (Catrambone, 1996) used a talk aloud approach during the training phase and found support for the self-explanation process. However, it would be useful to replicate this finding. In addition, the prior study did not manipulate the role correspondences. Thus, the talk aloud protocols in the present study might shed some light as to whether learners explicitly generalize over the roles played by humans and objects as well as form subgoals while studying examples.

In the present experiment learners were asked to explain the solutions to the examples they studied. It was predicted that participants studying examples that labeled the steps for finding the total frequency would be more likely than No Label participants to make a statement about certain steps being a unit and would also be more likely to offer an explanation for what the steps accomplished. Participants were also asked to explain aloud how they were approaching the first four test problems in order to examine whether they explicitly note the role reversals. If a learner is less likely to form a mental set concerning roles when the initial test problem has humans and objects playing different roles from those in the examples, then participants should be more likely to note the role reversal if it occurs on the initial test problem that if it does not occur until the third test problem. Such an observation by the learner should also be associated with an increased likelihood of getting the role assignments correct on the reversed role correspondence problems.

The No Label, Superficial Label, and Abstract Label groups' studied the same examples and solved the same problems as their counterparts in Experiment 3.

Method

Participants. Participants were 180 students recruited from an introductory psychology class at the Georgia Institute of Technology who received course credit for their participation. None of the students had taken a probability course prior to participating in the experiment, but all had at least one college-level calculus course.

Materials and procedure. Participants studied the same cover sheet as in the prior experiments and then studied three examples that used either the Superficial Label, Abstract Label, or No Label solution.

Participants were tested individually. After studying the cover sheet, participants were given a brief description of the process of talking aloud and then played the experimenter in a game of tic-tac-toe. Participants were asked to "describe what is going through your mind and what your strategy is each time you write an 'X.' " After this warm-up task, participants were given the three examples. They were told that they would be asked to solve problems after they studied the examples and therefore they should make sure they understood the examples well enough to solve similar and novel problems. They were asked to talk aloud in order to show the experimenter what they were doing to understand the solutions to the examples. If a participant was silent for 7 s[5] while studying, he or she was prompted to "please say out loud what is going through your mind." After they finished the first example, participants were no longer prompted because the examples were isomorphic and continual prompting could have made participants feel that they were being asked to repeat themselves arbitrarily. After studying the examples, participants solved the same 12 test problems used in Experiment 3. Participants were asked to talk aloud while working on the first four problems.

Participants solved the test problems in one of two orders. In the same-roles-first order, the order of the problems was the same as in Experiment 3. In particular, the first two problems were isomorphic to the examples and humans and objects played the same roles as in the examples, whereas the next two problems were isomorphs with reversed roles. In the reversed-roles-first order, the first two problems were isomorphic to the examples but had humans and objects playing reversed roles relative to the roles they played in the examples (see the first problem in Table 6 for a sample problem), whereas the next two isomorphs had humans and objects playing the same roles as in the examples. The remaining eight problems were in the same order for both conditions and matched the order used in Experiment 3. The reason for this is that the generalizations that Ross and Kennedy's (1990) participants produced were presumably caused by applying an example to a single problem, thus in the present experiment the role generalization, if

[5] This value was chosen after pilot testing and does not have any particular theoretical motivation.

affected by problem order, should occur before participants would have reached the nonisomorphic problems.

Participants' written solutions were scored for whether they found the average correctly. The talk aloud protocols produced during the training phase and during the first four test problems were tape recorded and transcribed. The transcriptions were annotated to indicate gestures by the participants such as when they pointed to a particular part of an example while talking aloud. Two raters independently coded the talk aloud protocols while one rater scored the solutions to the test problems. The raters agreed on coding 91% of the time (see the following discussion of the coding system). Disagreements were resolved by discussion.

Design. The independent variables were type of example solution studied (Superficial Label, Abstract Label, No Label) and order of the first four test problems (same-roles-first vs. reversed-roles-first), thus there were six groups in the experiment. There were 30 participants per condition, thus requiring 180 participants. The dependent measures were the presence of certain elements in the talk aloud protocols (see below) and performance on the 12 test problems.

Results and Discussion

Because of the large number of participants, the coding of the protocols so far has been limited to an examination of the presence of a small set of features. Specifically, the self-explanation features that were coded from the training phase were: (a) whether or not a participant made an observation about the steps for finding the total frequency being a group (he or she did not have to mention the fact that the steps calculated a total), and (b) whether a participant mentioned that those steps were in fact calculating a total. Participants' self-explanations were combined across the three examples. That is, if a participant mentioned either of the previous features for any of the examples, then he or she was scored as having mentioned the feature. The feature coded from the talk aloud protocols while participants worked on the first four test problems was whether or not a participant commented that the roles of humans and objects were different in either of the reversed role correspondence problems compared to the examples.

Self-explanations and role-reversal observations. There was a significant difference among the three label conditions in the frequency with which they observed that the steps for finding total frequency were a group, $\chi^2(2, N = 180) = 17.51, p = .0002$, with percentages of 57%, 52%, and 22% for the Superficial Label, Abstract Label, and No Label groups, respectively (the additional division of participants as a function of order of test problems is not relevant here because these self-explanations were generated during the study phase). Both label conditions mentioned the grouping more often than the No Label condition (vs. Abstract Label: $\chi^2(1, N = 120) = 11.63, p = .0007$; vs. Superficial Label: $\chi^2(1, N = 120) = 15.42, p = .0001$), whereas there was no significant difference between the two label conditions; $\chi^2(1, N = 120) = 0.30, p = .58$.

There was a significant difference among the three conditions in the frequency with which they mentioned that the steps calculated a total, $\chi^2(2, N = 180) = 14.33, p = .0008$, with percentages of 43%, 38%, and 13% for the Superficial Label, Abstract Label, and No Label conditions, respectively. Both label groups mentioned that the steps calculated a total more often than the No Label group (vs. Abstract Label: $\chi^2(1, N = 120) = 9.79, p = .002$; vs. Superficial Label: $\chi^2(1, N = 120) = 13.30, p = .0003$), whereas there was no significant difference between the two label groups; $\chi^2(1, N = 120) = 0.31, p = .58$. These relative differences among the conditions are very similar to those found in Catrambone (1996).

Neither training condition or the order of the test problems had an effect on the likelihood of participants mentioning anything about the roles of humans and objects in the first four test problems. Thirty percent of the participants in each of the two label conditions mentioned that the roles were reversed in at least one of the reversed role correspondence problems, whereas 23% of the No Label participants made that observation, $\chi^2(2, N = 180) = 0.64, p = .89$. For both the reversed-roles-first condition and the same-roles-first condition, 28% of the participants mentioned the role reversal.

Transfer. The test problems were scored exactly as in Experiment 3. Table 8 presents the results for the various problem types.

As in Experiment 3, all groups did quite well at finding λ on test problems that were isomorphs to the training examples and had objects and humans in the same role as in the examples (these were the first pair of problems for participants in the same-roles-first condition and the second pair of problems for participants in the reversed-roles-first condition). Performance on the isomorphs with the role reversals was not as good (see Table 8).

An analysis of variance was carried out on the performance on the isomorphic test problems with label and test order as the between-subjects variables and role correspondence (same as examples vs. reversed from the examples) as the within-subject variable. There were no significant effects due to label, $F(2, 174) = 1.54, p = .22, MSE = 0.74$, problem order, $F(1, 174) = 0.18, p = .67$, or their interaction, $F(2, 174) = 0.01, p = .99$. There were significant effects due to role correspondence, $F(1, 174) = 131.89, p < .0001, MSE = 0.44$, and the interaction of role correspondence and label, $F(2, 174) = 4.54, p = .012$. No other interaction was significant. The lack of an effect of test order, or an interaction between test order and role correspondence, suggests that, at least for isomorphs, test order did not appear to play a role on whether learners induced a mental set concerning roles.

Although all groups did well on the same role correspondence problem isomorphs, the interaction between label and role correspondence indicates that the Abstract Label group had less difficulty solving reversed role correspondence isomorphs relative to the other conditions. Pairwise comparisons on the reversed role correspondence isomorphs showed that the Abstract Label group outperformed the No Label group ($p < .02$) and marginally outperformed the Superficial Label group ($p < .06$), whereas the Superficial and No Label groups did not differ ($p > .57$).

An analysis of variance was carried out on the performance on the novel test problems with label and problem order as the between-subjects variable and role correspon-

Table 8
Number of Test Problems Solved Correctly as a Function of Label, Problem Order, and Role Correspondence (Experiment 4)

	Group						
	Superficial label		Abstract label		No label		
Problem type/ correspondence	Same roles first	Reversed roles first	Same roles first	Reversed roles first	Same roles first	Reversed roles first	*M*
Isomorphic							
Same	1.90	1.83	1.87	1.80	1.90	1.90	1.87
Reversed	1.00	0.97	1.33	1.30	0.90	0.87	1.06
M	1.45	1.40	1.60	1.55	1.40	1.38	1.46
Novel							
Same	2.90	2.83	2.73	2.70	1.90	1.83	2.48
Reversed	1.76	1.73	2.37	2.33	1.23	1.27	1.78
M	2.38	2.33	2.55	2.52	1.52	1.50	2.13

Note. $N = 30$ for each cell. Maximum possible score for any cell for isomorphic problems = 2. Maximum possible score for any cell for novel problems = 4.

dence as the within-subject variable. There was a significant difference due to label with respect to finding λ on the novel test problems (see Table 8), $F(1, 174) = 5.02$, $p = .008$, $MSE = 6.23$. There was no effect due to problem order of the four initial isomorphs, $F(1, 174) = 0.02$, $p = .90$. There was an effect of role correspondence indicating that problems with reversed role correspondences were solved less successfully than those with the same role correspondence as the examples, $F(1, 174) = 46.77$, $p < .0001$, $MSE = 0.94$. Finally, there was a significant interaction between label and role correspondence, $F(2, 174) = 4.64$, $p = .01$, suggesting that the correspondence manipulation affected the groups differently. No other interactions were significant. Once again, the lack of an effect of test order, or an interaction between test order and role correspondence, suggests that learners who had solved same role correspondence isomorphic test items first were no more likely to form a mental set than other participants.

Collapsing across problem order, pairwise comparisons among the three label conditions on the same role correspondence novel problems shows that both label groups outperformed the No Label group (both $ps < .02$), whereas the two label groups did not differ ($p > .65$). Pairwise comparisons among the three label conditions on the reversed role correspondence novel problems showed that the Abstract Label group outperformed the No Label group ($p = .002$) and showed a marginally significant advantage over the Superficial Label group ($p < .085$; although $p < .043$ for a one-tailed test which would be appropriate here because a directional difference was predicted), whereas the Superficial Label and No Label groups did not differ ($p > .14$).

Although participants in the Abstract Label group seemed to have a more general representation for the solution procedure compared to the other groups, even Abstract Label participants still had trouble adapting the procedure to problems with reversed role correspondences. That is, both label groups appeared to be conservative in their induction of the subgoals (i.e., the subgoals still tended to have some ties to superficial features). Still, the Abstract Label group was better able to adapt their procedures on reversed role correspondence problems compared to the Superficial Label group.

Transfer as a function of self-explanations. Because the label groups showed better transfer than the No Label group, the relationship between self-explanations and transfer was analyzed using only label participants. Participants who mentioned that the steps for finding the total frequency were a group performed better on the novel test problems than those who did not make that observation (see Table 9), $F(1, 118) = 14.03$, $p = .0003$, $MSE = 5.42$. There was also an effect of role reversal, $F(1, 118) = 33.46$, $p < .0001$, $MSE = 0.86$, as well as an interaction between these two factors, $F(1, 118) = 22.32$, $p < .0001$. Interestingly, this interaction seems to have been driven primarily by the large drop in performance from same role to reversed role correspondence problems by participants who mentioned the grouping. Nevertheless, the performance on the reversed role correspondence problems by those who mentioned the grouping was still better than the performance on the same role correspondence problems by those who did not mention the grouping.

Similar results were found with respect to the relationship between mentioning a total and transfer. Participants who mentioned that the steps found a total performed better on the novel test problems than those who did not make that observation (see Table 9), $F(1, 118) = 18.96$, $p < .0001$, $MSE = 5.22$. There was also an effect of role reversal, $F(1, 118) = 47.84$, $p < .0001$, $MSE = 0.87$, as well as an interaction between these two factors, $F(1, 118) = 21.26$, $p < .0001$.

Transfer as a function of comments made when solving the first four test problems. As with the previous analyses, the relationship between transfer and comments about role reversals was analyzed using only label participants. Participants were scored for whether or not they made an explicit comment about the roles of humans and objects being different than they were in the examples for any of the reversed role correspondence isomorphs.

Participants who commented on the role reversal did not

Table 9
Number of Novel Test Problems Solved Correctly as a Function of Self-Explanations, Observations of Role Reversals, and Role Correspondence (Experiment 4)

Role correspondence	Response: Yes	Response: No
Mentioned that steps formed a group		
	(n = 65)	(n = 55)
Same	3.57	1.87
Reversed	2.31	1.75
Mentioned that steps calculated a total		
	(n = 49)	(n = 71)
Same	3.90	2.03
Reversed	2.49	1.75
Commented on role reversal		
	(n = 36)	(n = 84)
Same	2.97	2.71
Reversed	2.03	2.06

Note. Maximum possible score for any cell = 4.

perform significantly differently on the novel test problems compared to those who did not make that observation (see Table 9), $F(1, 118) = 0.11, p = .74, MSE = 6.05$. There was an effect of role reversal, $F(1, 118) = 31.84, p < .0001$, $MSE = 1.01$. Of most interest is the interaction between role reversal and observation because such an interaction would suggest that those who commented on the role reversal would be less likely to be misled by it when solving the novel test problems. No interaction was found between these two factors, $F(1, 118) = 1.04, p = .31$.

Given the predictive association between the other self-explanation features and transfer, it is odd that no such relationship is found here. One possible explanation for a lack of relationship in this case might be that learners are more likely to explicitly talk aloud about procedural features of solution procedures than they are to talk about changes in roles. As a result, a number of the participants categorized as "no" in Table 9 might have noticed the role reversal but simply did not comment on it. Consistent with this possibility, Novick and Holyoak (1991) found that learners were more likely to talk about procedural correspondences between examples and problems (e.g., how numbers from an example line up with numbers from a problem) compared to conceptual correspondences (e.g., whether humans or objects provide the pool of choices). Another possibility is that although one might note the roles, it could be relatively difficult to adapt this knowledge (Ross, 1987, 1989).

The talk aloud results from Experiment 4 provide more direct evidence that a label can lead a learner to group a set of steps and to attempt to determine the goal or function of those steps through a self-explanation process. This self-explanation process also seems to help a learner to form a solution approach that is less likely to contain surface features in the representation, thus reducing the likelihood of these learners being misled on problems that involve changes in surface features.

Still, even learners who mentioned the idea of finding a total or reported noticing role reversals had trouble on reversed role correspondence problems. This suggests a relatively conservative induction; that is, even though some participants did mention finding a total in their protocols, their conception of a total still may have been linked to objects. This would be on top of difficulties that they still may have had adapting the solution steps (Novick & Holyoak, 1991). The superior performance by the Abstract Label participants on the reversed role correspondence problems suggests that they may have been at least a bit less conservative in their induction than other participants. Conversely, Superficial Label participants may have been too conservative and thus, when working on reversed role problems, had difficulties because of their conservatism concerning roles in addition to difficulties adapting the solution steps.

The results also suggest that the nature of the initial test problems does not necessarily affect solution generalization. This result would seem to contradict the findings of Ross and Kennedy (1990) who reported that learners were more likely to form an appropriate generalization if they were cued to apply the relevant example to a test problem compared to learners who were not cued. However, there are two differences between Experiment 4 and the experiments conducted by Ross and Kennedy that may account for the conflicting results. First, participants in Experiment 4 studied three examples, whereas Ross and Kennedy's participants studied one per principle. The greater number of examples in the present study could have aided generalization prior to learners working on the first test problem (although a pilot study has shown a benefit of a label with just one example). Second, the examples studied in the two label conditions were designed to teach subgoals to learners; this was not the case with Ross and Kennedy's materials. It could be that when a person learns subgoals from examples for solving a problem, any additional generalizations that might occur because of applying the example to an initial test problem are obscured or are relatively minor.[6] Clearly this issue needs to be examined systematically.[7]

The superior transfer to novel problems by the label

[6] A third difference between the studies is that problems in the present study never included cues to prior examples. However, this is not an issue because the examples studied were isomorphs (i.e., they involved the exact same solution steps) to one another and therefore there was no one example that was better than the others for a learner to recall when he or she worked on the test problems.

[7] For instance, it may be the case that although the subgoal structure is learned through examples, modifications or generalizations to the methods for achieving those subgoals are made primarily on an as-needed basis such as when the learner is confronted with a novel problem. Such a possibility could be explored through experiments in which learners are asked to describe their problem-solving procedures after studying (labeled or unlabeled) examples versus after studying examples and solving some problems that either do or do not require modifications to previously learned steps.

groups, as well as the superior transfer by those participants who appeared to form the subgoal to find the total—as assessed through the talk aloud protocols—is consistent with the interpretation that subgoal learning helps a person to solve novel problems that involve the same subgoals but require a change in the steps for achieving them. The smaller effect on role reversals suggests that this type of adaptation is qualitatively different than adaptations to solution steps and may respond better to other sorts of instructional interventions.

General Discussion

Learners frequently have difficulty solving problems that involve more than trivial changes to the steps demonstrated in training examples. This appears to occur because learners focus on memorizing steps rather than understanding what groups of steps accomplish. The present study was based on the assumption that transfer can be improved if learners form a solution procedure that is structured by subgoals and a method for achieving each subgoal rather than just a single linear set of steps for the entire procedure. Learners seem to be more likely to form such a structured solution procedure if they receive cues—such as labels—in example solutions that suggest that a set of steps form a group. If a learner notices the grouping then he or she seems to be more likely to attempt to explain to himself or herself the purpose of the steps and thus, to form a subgoal representing that purpose. The experiments in this study provide converging evidence for the subgoal-learning model and the benefits of subgoal learning on transfer.

Learners with a stronger math background are more likely to form an appropriate subgoal from a label cue compared to learners with a weaker background. Learners with a weaker background appear to rely more heavily on the semantic content of the label in order to form a subgoal. Unfortunately, a label that is related to surface features of a problem will be more likely to lead a learner to form a solution procedure that is tied to those features. An abstract label is less likely to lead a learner to make this mistake, although the learner must have relevant background knowledge in order to take advantage of an abstract label. These results suggest that cues such as labels can play a strong role in the formation of solution procedures. Because of this, care must be taken to construct cues in a way to aid the formation of structured solution procedures. For learners with weaker backgrounds these cues might need to be tied at least partially to example features despite the danger that this may lead the learner to form representations that have erroneous surface ties. However, for learners with stronger backgrounds, the cues can be constructed more abstractly, thus helping them to form appropriate subgoals.

Experiments 3 and 4 provide evidence that subgoals, particularly when formed from cues that do not lead to misleading connections to surface features of problems, can also help learners deal with changes in problems involving the mapping of entities in the problems to their mathematical roles. In addition, the subgoal learning that occurs during training appears to make additional generalizations during an initial problem-solving episode less likely to be needed. Experiment 2 suggests that the effort required to self-explain a set of steps connected by an abstract label produces better retention of a subgoal.

Although the present set of experiments used materials involving the Poisson distribution, related experiments using algebra word problems and problems involving permutations and combinations have produced analogous results (e.g., Catrambone, 1994). These studies have found that the use of labels and short phrases designed to encourage learners to group sets of steps have helped learners to be more successful solving novel problems.

Subgoals, Prior Knowledge, and Procedure Modification

Chi et al. (1994) argued that the chances for integrating new knowledge with old knowledge increase when a learner has the opportunity to do the integration in a "*minute* and *ongoing* fashion" (p. 474; emphasis the original authors'). That is, a learner is more likely to self-explain effectively and integrate or assimilate new knowledge with prior knowledge if the learner can perform these tasks in relatively small pieces. The use of grouping cues in solution procedures provides the learner with the opportunity to self-explain each part of the procedure, thus increasing the likelihood of integrating the resulting new knowledge with prior knowledge.

Wattenmaker et al. (1986) suggested that the degree of difficulty in learning a particular category structure is at least partly a function of the type of knowledge that learners bring to the task. With respect to the present work, one could argue that a label leads learners to bring relevant prior knowledge to bear in order to self-explain the purpose of a set of steps. In a subsequent transfer situation, a learner possessing a solution procedure organized by subgoals will have an increased chance of accessing appropriate prior knowledge in order to solve a novel problem. This will occur because the subgoals can help constrain the search space the learner explores when trying to modify the old solution procedure.

Subgoal learning may be particularly useful for some types of transfer but less so for others. Learners who were predicted to learn the subgoal to find the total frequency were more successful than other learners in terms of solving novel problems that required a change in the steps for finding the total. However, both types of learners still had difficulty when the superficial mapping of objects and humans was reversed, albeit the group predicted to have formed subgoals seemed to suffer less. This suggests that the various types of generalizations, modifications, and inferences needed to adapt a solution procedure to novel problems are aided to different degrees by possibly different factors. For instance, perhaps the role-reversal adaptation would be aided by a direct statement to a learner to pay attention to the roles of humans and objects in the examples and test problems.

Use of Labels Revisited

Learners with stronger math backgrounds in Experiment 1 who received abstract labels were more likely to form representations free of misleading surface features. It is interesting to consider though what the nature of the representations might be if the labels were less abstract, but still unconnected to surface details of the examples. For instance, instead of containing the label *total number of briefcases owned*, suppose an example contained the label *total frequency of the event*. This label is formally correct and not related to surface details of the example. Perhaps this sort of label would produce the best transfer because it would presumably serve as a grouping cue and aid self-explanation (relative to a more abstract label such as Ω) but not provide a misleading tie-in to surface features. The effects of such a label on the representations formed by learners with a weaker background are unclear and worth investigating. It is possible that such a label would provide enough guidance to help weaker background learners to form the necessary subgoal without inappropriate surface ties or it may be the case that such a label remains too abstract to help these learners form the subgoal.

Alternatively, given that the label would be presented in the context of examples, it is also possible that learners, regardless of background, would tend to do a conservative induction (cf. Medin & Ross, 1989) and instantiate or represent "event" as objects if the worked examples were like the ones used in the present experiments. Thus, an additional manipulation would be to vary what constitutes an "event" in the examples in order to see if that leads to a better generalization for learners. Although this approach might be successful for the types of problems used in the present research, it might not scale up. That is, given the large number of possible variations of surface and procedural details in potential examples and problems in various domains, it could take a huge number of examples to guide learners to the appropriate generalizations. Therefore, manipulations—such as labeling—that are expected to induce an active self-explanation process might be the most fruitful and efficient approach for designing effective examples that lead to useful generalizations and therefore, successful problem solving.

References

Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 391–412.

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review, 86,* 124–140.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view.* New York: Holt, Rinehart & Winston.

Bassok, M., Wu, L. L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretive effects of content on problem-solving transfer. *Memory & Cognition, 23,* 354–367.

Ben-Zeev, T. (1995). The nature and origin of *rational errors* in arithmetic thinking: Induction from examples and prior examples. *Cognitive Science, 19,* 341–376.

Brown, A. L., Kane, M. J., & Echols, C. H. (1986). Young children's mental models determine analogical transfer across problems with a common goal structure. *Cognitive Development, 1,* 103–121.

Cabrera, A., & Billman, D. (1996). Language-driven concept learning: Deciphering *Jabberwocky. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 539–555.

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

Catrambone, R. (1995). Aiding subgoal learning: Effects on transfer. *Journal of Educational Psychology, 87,* 5–17.

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1020–1031.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1147–1156.

Catrambone, R., & Holyoak, K. J. (1990). Learning subgoals and methods for solving probability problems. *Memory & Cognition, 18,* 593–603.

Chen, Z. (1995). Analogical transfer: From schematic pictures to problem solving. *Memory & Cognition, 23,* 255–269.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18,* 293–328.

Chi, M. T. H., Bassok, M., Lewis, R., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M, & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439–477.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Clement, C. A., Mawby, R., & Giles, D. E. (1994). The effects of manifest relational similarity on analog retrieval. *Journal of Memory and Language, 33,* 396–420.

Dixon, P. (1987). The processing of organizational and component step information in written directions. *Journal of Memory and Language, 26,* 24–35.

Dufresne, R. J., Gerace, W. J., Hardiman, P. T., & Mestre, J. P. (1992). Constraining novices to perform expertlike problem analyses: Effects on schema acquisition. *The Journal of the Learning Sciences, 2,* 307–331.

Eylon, B., & Reif, F. (1984). Effects of knowledge organization on task performance. *Cognition and Instruction, 1,* 5–44.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253–292.

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120,* 34–45.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38.

Heller, J. I., & Reif, F. (1984). Prescribing effective human problem-solving processes: Problem description in physics. *Cognition and Instruction, 1,* 177–216.

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 83–94.

Keane, M. (1987). On retrieving analogues when solving problems. *The Quarterly Journal of Experimental Psychology, 39A,* 29–41.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980, June). Expert and novice performance in solving physics problems. *Science, 208,* 1335–1342.

LeFevre, J., & Dixon, P. (1986). Do written instructions need examples? *Cognition and Instruction, 3,* 1–30.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs, 54*(Whole No. 248).

McDaniel, M. A., & Schlager, M. S. (1990). Discovery learning and transfer of problem solving skills. *Cognition and Instruction, 7,* 129–159.

McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes, 22,* 247–287.

Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 5.* (pp. 189–223). Hillsdale, NJ: Erlbaum.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242–279.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987, October). Teaching reasoning. *Science, 238,* 625–631.

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 398–415.

Pirolli, P. L., & Anderson, J. R. (1985). The role of learning from examples in the acquisition of recursive programming skill. *Canadian Journal of Psychology, 39,* 240–272.

Reed, S. K., Ackinclose, C. C., & Voss, A. A. (1990). Selecting analogous problems: Similarity versus inclusiveness. *Memory & Cognition, 18,* 83–98.

Reed, S. K., & Bolstad, C. A. (1991). Use of examples and procedures in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 753–766.

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 106–125.

Reeves, L. M., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin, 115,* 381–400.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629–639.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 456–468.

Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 42–55.

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practical procedures. *Psychological Bulletin, 110,* 577–586.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association, 81,* 826–831.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257–285.

VanLehn, K. (1988). Toward a theory of impasse-driven learning. In H. Mandl & A. Lesgold (Eds.), *Learning issues for intelligent tutoring systems* (pp. 19–41). New York: Springer.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L., (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology, 18,* 158–194.

Received October 29, 1996
Revision received April 22, 1997
Accepted December 17, 1997 ■