The Journal of Experimental Education, *2006*, *74*(2), 109–136

Teaching and Assessing Deductive Reasoning Skills

JACQUELINE P. LEIGHTON
University of Alberta

ABSTRACT. The author examined the effectiveness of training in symbolic logic for improving students' deductive reasoning. A total of 116 undergraduate students (approximately equal numbers of men and women) enrolled in 1st-year university philosophy courses in symbolic logic participated in 2 studies. In both studies, students completed booklets of categorical and conditional syllogisms at the beginning of the course and again at the end of the course. In Study 2, students also specified their reasoning strategies. Results indicated that students' strategies changed with training (students increased their use of mental models and mental rules with categorical and conditional syllogisms, respectively), but their reasoning performance improved only moderately. The educational implications of these results are explored.

Key words: assessment, cognitive processes, reasoning, strategies, transfer

IMAGINE THAT YOU ARE TOLD, "All instances of school violence are preceded by verbal warnings." How do you conclude whether this statement is true or false? One way to evaluate the truthfulness of the statement is to think systematically of counterexamples—instances in which no verbal warning or threat was issued and violence ensued. If you can think of at least one counterexample, then the statement under consideration can be deemed false. The value of applying deductive reasoning methods, such as considering counterexamples, to our everyday thinking is that our thinking becomes more systematic and less prone to deception. Because of its truth-preserving quality, deductive reasoning is a cornerstone of scientific research in which claims to new knowledge are evaluat-

Address correspondence to: Jacqueline P. Leighton, Centre for Research in Applied Measurement and Evaluation (CRAME), Department of Educational Psychology, 6-110 Education North. University of Alberta, Edmonton, Alberta, Canada T6G 2G5. E-mail: jacqueline.leighton@ualberta.ca

ed rigorously. Scientists, however, are not the only ones who should profit from systematic thinking. Students stand to become better decision makers, problem solvers, and thinkers when they learn to reason systematically (Byrnes, 2001; Chen & Klahr, 1999; Klaczynski & Narasimham, 1998; Kuhn, 1992, 2001; Kuhn, Garcia-Mila, Zohar, & Anderson, 1995).

Numerous studies have revealed that adolescents and adults exhibit regular biases in their deductive reasoning (e.g., Johnson-Laird, 1999; Klaczynski, Byrnes, & Jacobs, 2001; Markovits, Doyon, & Simoneau, in press; Mynatt, Doherty, & Dragan, 1993; Torrens, Thompson, & Cramer, 1998; Tversky & Kahneman, 1983; Ward & Overton, 1990; Wason, 1960). These biases, which include failing to search for counterexamples and generating conclusions that corroborate prior beliefs, are observed when participants solve both familiar and unfamiliar problems (for a review, see Baron, 2000; Evans & Over, 1996; Stanovich, 1999). Little is known about the source of these biases because little is known about the conditions under which students learn to reason deductively and the strategies that mediate their deductive reasoning (Roberts, 2004; Roberts & Newton, in press).

Deductive reasoning typically has been viewed as a skill that emerges successfully without training (Cosmides, 1989; Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Rips, 1994; Schaeken, De Vooght, Vandierendonck, & d'Ydewalle, 2000). However, this view increasingly lacks empirical support, and more researchers are calling for studies of how individuals learn to reason formally (Evans & Over, 1996; Roberts & Newton, in press). Deductive skills are not taught in secondary grades when students would be expected to increasingly exercise their reasoning skills (Byrnes, 2001; Piaget, 1972). Instead, deductive skills are assumed to be learned *indirectly* by students in traditional content areas such as science and mathematics (National Center for Education Statistics, 2003); as students learn content knowledge and procedural skills in science and math, they are assumed to "pick up" higher order thinking skills such as deductive reasoning. The problem is that students do not seem to be picking up these skills given the predominance of biased reasoning noted previously. Although deductive skills are formally taught at the college level, few studies at the college level have focused on (a) the effectiveness of training deductive reasoning in a classroom setting and (b) the changes in strategies that accompany improvements in deductive reasoning (Leighton, 2001; Leighton & Sternberg, 2003; Roberts, 2004). Determining the effectiveness of training deductive reasoning in ecologically valid settings such as classroom environments is needed to determine the suitability of integrating this focus of instruction at earlier grades.

One of the goals of educational reform (e.g., the No Child Left Behind Act of 2001) is to give all students an opportunity to be successful learners in increasingly complex and technical environments. Learning how to reason deductively, to evaluate the soundness of everyday arguments, or to generate key and necessary inferences from masses of information is mandatory for success in scientif-

ic and other professional and scholarly activities (Kuhn, 2001; Nickerson, 2003; Overton, 1990; Piaget, 1972; Siegler, 1996; Stanovich, 1999). In fact, deductive reasoning is considered fundamental to many scholarly disciplines as evidenced by the analytical sections of the Law School Admissions Test and the Graduate Records Exam—the latter being the entrance exam taken by candidates applying for postgraduate study (Powers & Dwyer, 2003). Given the relative dearth of studies devoted to determining whether deductive reasoning skills can be improved through formal training and the kinds of strategies that underlie performance, my goals in the present study were to (a) investigate the effectiveness of training deductive reasoning—in particular, categorical and conditional reasoning[1]—in a classroom setting, (b) assess the training of deductive reasoning using standard tasks in different formats, and (c) explore the changes in reasoning strategies that accompany any improved performance.

Training Deductive Reasoning

Roberts and Newton (in press) recently asked, "How can we understand human reasoning without understanding how people learn to improve their reasoning?" (p. 245). The novice–expert literature provides a compelling background for exploring the training of deductive reasoning skills because deductive reasoning—just like mathematics or creative writing or any disciplined domain—can be viewed as a specific *domain of knowledge* (Leighton & Sternberg, 2003). The knowledge and skills of deductive logic are precise but are sometimes mistakenly viewed as general only because they have wide applicability. However, the applicability of knowledge is separate from its substance. Mathematics and creative writing are also disciplines with wide applicability, yet few educators would argue that these disciplines fail to constitute specific domains of instruction. This is essential to recognize because viewing deductive reasoning as domain-general knowledge and, therefore, not open to direct instruction has been an unfortunate assumption that has stymied study into its training.

The study of expertise, at the broadest level, focuses on identifying what distinguishes outstanding performers within a domain from less outstanding performers (Anderson & Leinhardt, 2002; Ericsson & Charness, 1994; Morris, 2002). According to the literature on novice–expert performance differences, problem solving within any content area—even deductive logic—can be improved with the acquisition of relevant domain-specific knowledge and practice skills (Anderson & Leinhardt; Ericsson & Charness; Morris). That is, acquiring a vast amount of well-organized, domain-specific knowledge and processing strategies can improve performance within a domain by facilitating the ability to

[1]Categorical reasoning and conditional reasoning are two commonly measured forms of deductive reasoning (Evans & Over, 1996; Johnson-Laird, 1999; Johnson-Laird & Bara, 1984; Leighton & Sternberg, 2004).

recognize important problem features quickly, to access chunks of relevant problem-solving strategies, and to solve problems efficiently and correctly (Chase & Simon, 1973; de Groot, 1965; Gobet & Simon, 1996). For example, Gobet and Simon found that champion (expert) chess players could recall more chess positions than novice chess players. Likewise, Allard and Starkes (1991) found that elite athletes were able to recall more information about game situations after brief exposure than nonelite athletes. In sum, experts recognize meaningful relations or patterns in their domains of expertise because they can draw on specific content knowledge (Gobet & Simon).

Despite the recognized superiority of expert problem solving, research on novice–expert differences has not translated well into effective educational programs to boost student performance (Alexander, Sperl, Buehl, Fives, & Chiu, 2004). One difficulty with translating research on expertise into educational programs is that becoming an expert within a content area takes many years to achieve, much longer than the duration of any one educational program. Nonetheless, a valuable lesson to be taken from the novice–expert literature is that with deliberate study and practice, improved performance is within the grasp of anyone (Alexander et al.). By increasing content knowledge via direct training, students can improve their ability to distill important patterns of information and apply appropriate strategies (Ericsson, 1996; Ericsson & Kintsch, 1995). Alexander's model of domain learning (see Alexander et al.) in particular underscores this goal of training—that is, to gradually help students progress toward expert performance by facilitating their content knowledge, interest, and problem-solving strategies.

How training is dispensed and assessed is important to consider, however. To illustrate this point, some investigators attempted to train deductive reasoning with mixed results (Cheng, Holyoak, Nisbett, & Oliver, 1986; Lehman, Lempert, & Nisbett, 1988; Morris & Nisbett, 1993). For example, Cheng et al. found that undergraduate students who participated in a short training study as well as a semester-long course in standard deductive logic failed to avoid biases in their conditional reasoning. In other words, training students' domain-specific knowledge of deductive logic did not improve their performance on conditional reasoning tasks. However, there were two problems with this study: First, Cheng et al. failed to assess a range of deductive skills because students' categorical reasoning was not assessed along with their conditional reasoning. The logic training students underwent in the Cheng et al. study might have facilitated their categorical reasoning but not their conditional reasoning. Students' categorical reasoning should have been assessed alongside their conditional reasoning in order to have a more complete picture of students' posttraining performance gains.

A second problem was the method used to assess conditional reasoning. Cheng et al. (1986) assessed students' conditional reasoning with variations of the Wason selection task (Wason, 1966; see Figure 1). Although describing the

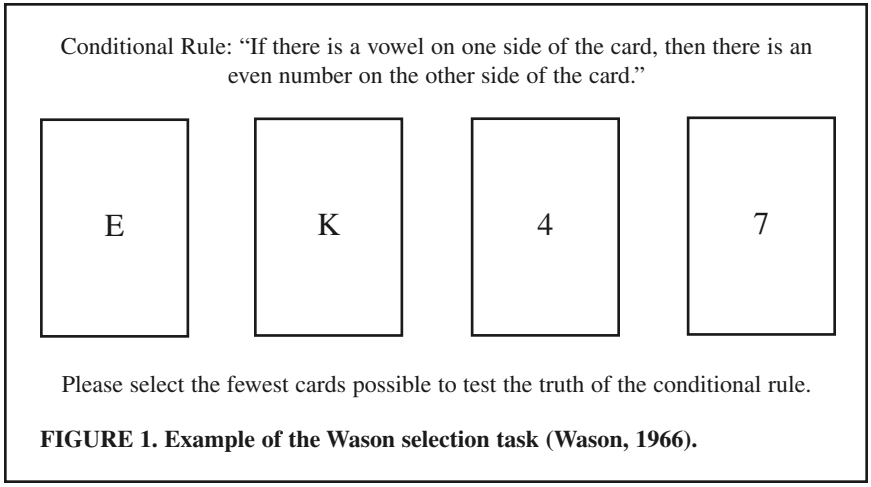

FIGURE 1. Example of the Wason selection task (Wason, 1966).

vast literature associated with this task is outside the scope of this article, the task is considered controversial by researchers because of the ambiguity in whether it measures conditional reasoning, hypothesis testing, or meta-reasoning skills (e.g., Hardman, 1998; Kirby, 1994; Sperber, Cara, & Girotto, 1995; see also Sternberg & Ben-Zeev, 2001, for a discussion on whether the Wason selection task qualifies as a deductive task). Using the Wason selection task to assess conditional reasoning specifically, then, is not appropriate given the ambiguity of what it really measures (e.g., see Sternberg & Ben-Zeev). In defense of the Wason task, Leighton (2001) did find that a sizeable portion of college students could learn to solve the Wason task correctly after they were given direct feedback about their performance. However, Evans and Over (1996) suggested leaving the Wason task aside because "evidence for deductive competence is much easier to locate" with other, more standard tasks (p. 11).

A similar criticism can be made of two other studies examining the training of deductive reasoning skills. Lehman et al. (1988) and Morris and Nisbett (1993) also included variations of the Wason selection task to assess conditional reasoning. First, Lehman et al. included variations of the Wason selection task to assess graduate students in psychology and chemistry and professional students in law and medicine. They found that psychology graduate students and students in law and medicine improved their conditional reasoning after 2 years of study within their disciplines. However, the implications of those results with regard to training are difficult to interpret because deductive reasoning skills were not directly taught to students through formal means (i.e., a course in logic), and the sample of participants in the Lehman et al. study was highly restricted—graduate students and professional school students. These advanced students may be considered highly motivated and increasingly expert in a number of domains

(e.g., research methodology, hypothesis testing, critical thinking). It is, therefore, unclear whether any gains in conditional reasoning can be attributed to a single training program or whether the gains originated from the value these students place in critical thinking generally.

Second, Morris and Nisbett (1993) assessed a cross-section of first- and third-year philosophy and psychology graduate students in categorical and conditional reasoning. Morris and Nisbett found that third-year philosophy students exhibited better categorical reasoning (but not conditional reasoning) than first-year philosophy students. Furthermore, third-year psychology students exhibited better conditional reasoning (but not categorical reasoning) than first-year psychology students. As with the Lehman et al. (1988) study, conditional reasoning performance was measured using variations of the Wason selection task. The previously mentioned ambiguity of this task aside, using the Wason selection task to assess psychology graduate students is especially problematic because this task is well known in the psychological literature; in fact, it is the most studied task in the reasoning literature (Evans & Over, 1996).

Given the limitations with the assessment methods used in these studies (i.e., Cheng et al., 1986, Lehman et al., 1988, and Morris & Nisbett, 1993), my goals in the present research were to (a) investigate the effectiveness of training categorical and conditional reasoning in an ecologically valid setting where students are motivated to learn and learning gains are expected; (b) assess the effectiveness of training students using standard deductive tasks in different formats (standard deductive tasks are well matched to the tasks used during training and have been used in previous studies, e.g., Johnson-Laird & Bara, 1984); and (c) explore the changes in strategies that accompany improved performance so as to contribute to a small but key literature focused on exploring the strategies that mediate deductive reasoning (Bucciarelli & Johnson-Laird, 1999; Evans, 2000; Newton & Roberts, 2000; Roberts, 2000a, 2000b).

Reasoning strategies in deductive reasoning. Reasoning strategies are defined in this article using Siegler and Jenkin's (1989) broad definition as "any procedure that is nonobligatory and goal directed" (p. 11). This broad definition is chosen only because narrower definitions (e.g., Evans, 2000) are typically associated with additional requirements (e.g., conscious awareness) that are still untested (see Roberts & Newton, in press, for a discussion of this issue). Unlike domains such as mathematics (e.g., Siegler, 1996; Siegler & Jenkins), the role of strategies in deductive reasoning has only recently received adequate attention (Bucciarelli & Johnson-Laird, 1999; Evans; Newton & Roberts, 2000; Roberts, 2000a, 2000b). Deductive reasoning strategies were neglected in the past because psychologists focused on identifying a fundamental, general-purpose reasoning mechanism that operated consistently in all people and in all contexts (Roberts, 1993, 2004).

Roberts (1993, 1997, 2000a) identified problems with this fundamental reasoning mechanism. To begin, he raised doubts as to how a researcher could know whether any observed reasoning procedure could legitimately be considered fundamental and not simply a manifestation of voluntary strategic behavior (see also Leighton & Sternberg, 2003, and Stenning & Oberlander, 1995, for a discussion of how domain-specific knowledge might mediate strategy selection and use). In other words, the procedures individuals report as they reason through a problem might reflect only strategy selection and use and not anything more fundamental. Of course, if it turned out that everyone followed the same procedures to solve all deductive tasks and there were no individual differences, then it might be plausible to argue that something fundamental was shared by people's reasoning procedures. However, Roberts (1997, 2000) noted that individual differences on deductive tasks are found constantly (e.g., Ford, 1995; Galotti, Baron, & Sabini, 1986). Consequently, he (2004) suggested "that it only makes sense to discuss reasoning in terms of strategies that people apply, where a strategy is *a coherent set of goal-directed procedures whose application is not compulsory* (see also Siegler, 1996; Siegler & Jenkins, 1989)" (p. 242).

First, I conducted a study to examine whether training students in symbolic logic resulted in improved deductive reasoning. Following preliminary trends observed in the first study, I conducted a second and larger study to give students the opportunity to report their strategy for solving each set of deductive reasoning tasks. It is important to note that the training the students received in Study 1 and Study 2 was domain-specific in the knowledge and strategies of deductive logic. As logicians would surely agree, deductive logic is as much a specific content domain as is, for example, mathematics. As such, deductive logic should be viewed as compliant to direct training and learning (see Leighton & Sternberg, 2003; Roberts & Newton, in press). As previously mentioned, the knowledge and skills of deductive logic are sometimes mistakenly viewed as general only because they have wide applicability. However, the broad applicability of deductive logic does not undermine its status as a specific domain of knowledge.

STUDY 1

Method

Participants

The study took place in the fall/winter of 2001 and involved a class of 49 undergraduate students (mean age = 21.34 years; $SD = 1.92$) enrolled in a first-year introductory philosophy course in symbolic logic at a large research-intensive university. The textbook *A Concise Introduction to Logic* by Patrick J. Hurley (2002) was used in the course. The appropriate book chapters were used to teach

the various kinds of reasoning targeted in the present study.[2] Book chapters were also supplemented by other material from the instructor, who was an experienced, tenured professor in the area. Of the 49 students, 27 students were men and 22 students were women. The gender distribution is reported here for completeness but was not a variable of interest in this article.[3] Students were given the option of payment or research credit for their participation, and 85% of students selected research credit. Approximately 76% of students declared that their major area of study was in the faculty of arts, and approximately 24% declared their major in the faculties of science, business, education, or undecided.

Procedure

Participating students were assessed twice on categorical and conditional syllogisms—once at the beginning of their philosophy course (logic training Trial 1) and once again at the end of their course (logic training Trial 2). The assessment method included categorical and conditional syllogisms because these tasks were well matched to the training problems students experienced during the course and were sufficiently novel to measure students' adaptive skills (see Bar-

[2]The course began with a discussion of ordinary-language reasoning and, in particular, focused on common fallacies or errors in reasoning. The instructor focused on parts of Chapter 1, most of Chapter 3, and also on various other sources. For this section of the course, the instructor also had students look at newspaper columns, editorials, and letters to the editor as sources of arguments to evaluate. The second topic focused on conditional/sentential logic. *Sentential logic* is the logic of compound sentences, such as "If Kwasi already took the test, then he is back in his dorm room; Kwasi already took the test; what can you conclude?" For this topic, the instructor had students read Chapter 6 of the textbook as well as selected parts of Chapter 7. The next topic focused on reasoning with categorical syllogisms. *Categorical syllogisms* involve quantified premise sets, such as "All chickens have gizzards. Nothing with a gizzard is suitable for framing. So no chickens are suitable for framing." The instructor had students read Chapters 4 and 5 from the textbook, and students completed the exercises from Chapter 5. The second and third topics, conditional and categorical logic, respectively, were specifically relevant to the assessment of students' performance on conditional and categorical syllogisms in the present study. The final topic was probability and statistical reasoning, and the instructor had the class read Chapter 9, along with various supplemental readings.

[3]I did not consider gender a variable of interest because the focus of the current study was on the efficacy of training and assessment of deductive reasoning and not on demographic characteristics that cannot be modified to promote more effective thinking. Studies of deductive reasoning (e.g., Cosmides, 1989; Johnson-Laird & Bara, 1984; Schaeken et al., 2000) often do not include gender as a variable of interest because reasoning is viewed as a skill that does not favor one gender over another. In recent support of this view, Morse and Morse (1995) examined the influence of training and strategy differences on reasoning and failed to find gender differences in performance of divergent and convergent thinking.

nett & Koslowski, 2002). Moreover, categorical and conditional syllogisms are standard tasks used in deductive reasoning studies (Evans & Over, 1996; Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Leighton & Sternberg, 2003; Rips, 1994).

I randomly assigned students to one of two assessment formats—selection of responses or construction of responses. I manipulated assessment format because different formats make different processing demands on students (Anderson, 1990; Noice & Noice, 2002; Tulving & Thompson, 1973; Watkins & Tulving, 1975). For instance, a selected-response format requires students to recognize and select the correct response, and a constructed-response format requires students to recall and construct the correct response. It is normally assumed that recognition will be superior to recall because the task of recognizing and selecting a response from a list of alternatives activates more memory sources than the task of recalling and constructing a response, but this is not always the case (Anderson; Noice & Noice; Tulving & Thompson; Watkins & Tulving). For example, Tulving and Thompson, Watkins and Tulving, and, more recently, Noice and Noice indicated that recall can be superior to recognition when the test context matches the context of original study and when the material is well learned. Therefore, I deemed it necessary to assess students using both formats to determine which format—either selected or constructed response—was the more sensitive measure of deductive reasoning and its training.

Within each assessment condition, students completed a test booklet containing a set of 16 randomly presented categorical syllogisms and a set of 16 randomly presented conditional syllogisms, counterbalanced for order and of varying difficulty (see Table 1 for syllogism types). Each test booklet contained seven pages: The first page requested student demographic information (i.e., name, major area of study, and birth date), the second page provided standardized instructions to the first task (either categorical or conditional syllogisms depending on the counterbalanced order), followed by two pages of syllogisms. The fifth page provided standardized instructions for the second task, followed by the remaining two pages of syllogisms.

I chose the 16 categorical syllogisms from Johnson-Laird and Bara's (1984) list of categorical syllogisms, which are divided between two levels of difficulty: one-model determinate syllogisms are considered easy to solve because they require the creation of a single mental representation[4] (see Figure 2); two-model

[4]One-model determinate syllogisms require a single mental representation to solve and so are not considered to tax working memory to the same extent as two- and three-model determinate syllogisms, which require two and three mental representations, respectively (Johnson-Laird, 1999; Johnson-Laird & Bara, 1984). A problem's load on working memory is considered an indicator of a problem's difficulty (Evans & Brooks, 1981; Klauer, Stegmaier, & Meiser, 1997; Meiser, Klauer, & Naumer, 2001; Toms, Morris, & Ward, 1993).

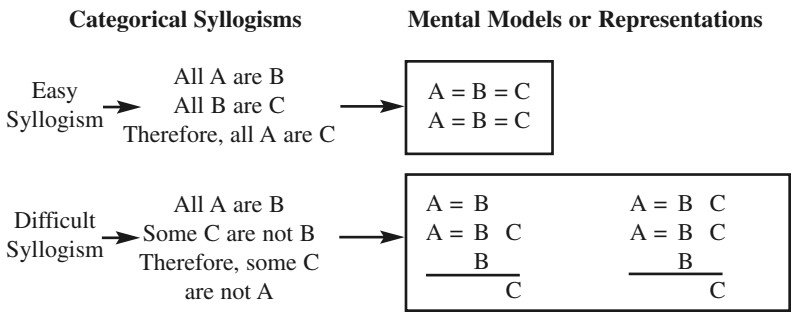

FIGURE 2. Examples of representations associated with tasks of varying difficulty (adapted from Johnson-Laird & Bara, 1984).

indeterminate syllogisms are considered difficult to solve because they require the creation of at least two mental representations; and two-model and three-model (multiple-model) determinate syllogisms are also considered difficult to solve because they require the creation of two and three mental representations, respectively. I chose the 16 conditional syllogisms from Evans and Lynch (1973), which are divided between two levels of difficulty: One-model determinate syllogisms (i.e., *modus ponens*) are easy to solve because they require the creation of a single mental representation; indeterminate syllogisms (i.e., denial of the antecedent and affirmation of the consequent) are difficult to solve because they require the creation of at least two mental representations; and three-model determinate syllogisms (i.e., *modus tollens*) are also difficult to solve because they require the creation of three mental representations. Two-model and three-model syllogisms are considered to be approximately equally difficult for students to solve because students who initiate the cognitive effort to generate two representations tend to also initiate the effort to generate three representations (Johnson-Laird & Bara). Conversely, students who cannot generate two representations to solve a problem will also not generate three representations when needed. In other words, two-model syllogisms seem to act as a threshold for separating easy from difficult syllogisms.

Although I asked students to solve both determinate and indeterminate syllogisms for completeness, I did not plan on including data originating from indeterminate syllogisms in the analysis. The problem with data stemming from indeterminate syllogisms is that they very likely misrepresent students' deductive reasoning skills. Students often generate the correct answer to these syllogisms for incorrect reasons. For example, Johnson-Laird and Bara (1984) have shown that students often guess the correct answer—*no valid conclusion*—to these syllogisms when they have not constructed the required mental representations.

I adapted the following task instructions for the first assessment condition (constructed-response) from Johnson-Laird and Bara (1984):

> In the following pages, you will see 16 pairs of statements about different people or groups of people, whom you should imagine as assembled in a room. After reading each pair of statements, please write down what, if anything, follows necessarily from these premises about the occupants of the room. If you consider that there is no conclusion that follows necessarily from the premises, please write down "Nothing." Your conclusions should be based solely on the information presented in the statements, and NOT on plausible suppositions or general knowledge.
>
> For example, please read the statements below:
>
> 1. All artists are beekeepers.
>   *All beekeepers are chefs.*
>   Then . . . all artists are chefs.
>
> From these statements, I can definitely conclude that *All artists are chefs,* so I write this down below the statements.
>
> 2. Some acrobats are bakers.
>   *Some bakers are canoeists.*
>   Then . . . nothing.
>
> From these statements, I can NOT conclude anything definite about acrobats and canoeists, so I write down "Nothing" below the statements.
>
> If you have any questions, please ask the experimenter now. If you don't have any questions, please begin Task 1 on the next page. Please, do NOT skip any questions.

After completing Task 1, students were given instructions for Task 2. The instructions for Task 2 were modeled after the instructions for Task 1 to maintain consistency:

> In the following pages, you will see 16 pairs of statements about people located in well-known places. After reading the pair of statements, please write down what, if anything, follows necessarily from these premises about the location of one of the persons. If you consider that there is no conclusion that follows necessarily from the premises, please write down "Nothing." Your conclusions should be based solely on the information presented in the statements and NOT on plausible suppositions or general knowledge.
>
> For example, please read the statements below:
>
> 1. If Hank is in Chicago, then Julia is in Baghdad.
>   *Hank is in Chicago.*
>   Then . . . Julia is in Baghdad.
>
> From these statements, I can definitely conclude that *Julia is in Baghdad*, so I write this down below the statements.
>
> 2. If Linda is in Madrid, then Robert is NOT in Philadelphia.
>   *Linda is NOT in Madrid.*
>   Then . . . nothing.
>
> From these statements, I can NOT definitely conclude that Robert is NOT in Philadelphia, so I write down "Nothing" below the statements.

If you have any questions, please ask the experimenter now. If you don't have any questions, please begin Task 2 on the next page. Please, do NOT skip any questions.

The students completed the 32 syllogisms in approximately 45 min.

Results

Given students' 12-week training in symbolic logic, with 6 weeks of this time being devoted to training categorical and conditional reasoning exclusively, I expected students to improve their performance on difficult categorical and conditional syllogisms. There was less of an expectation for students to improve their performance on easy categorical and conditional syllogisms because these syllogisms were assumed to elicit high levels of performance even without training.

To determine the effects of assessment format, logic training, and syllogism type on students' reasoning performance, I conducted mixed analyses of variance (ANOVAs; Glass & Hopkins, 1996; Keppel, 1991). I performed a mixed ANOVA combining a between-subject variable[5] (i.e., assessment format: selected response vs. constructed response) with two within-subject variables (i.e., logic training: Trial 1 vs. Trial 2; and syllogism type: categorical vs. conditional) to examine students' performance on easy syllogisms. The same analysis was performed to examine students' performance on difficult syllogisms.

Easy Categorical and Conditional Reasoning

I conducted a 2 (assessment format: selected response vs. constructed response) × 2 (logic training: Trial 1 vs. Trial 2) × 2 (syllogism type: categorical vs. conditional) mixed ANOVA to examine students' performance on easy problems. Alongside each statistically significant effect is its p value and an index of its effect size ($\hat{\omega}^2$), a measure of treatment magnitude or proportion of explained variance (Keppel, 1991; Myers & Well, 1995). According to Cohen (1977) and Keppel, $\hat{\omega}^2$ values of approximately .15, .06, and .01 are considered large, medium, and small, respectively.

The analysis of easy syllogisms indicated two significant effects: (a) a significant main effect of syllogism type, indicating that students' overall performance was better for easy conditional syllogisms (i.e., M = 7.84 at Trial 1 and 7.76 at Trial 2) than for easy categorical syllogisms (i.e., 5.68, 6.24), $F(1, 47) = 45.117$,

[5]The unbalanced sample sizes between the two format conditions (i.e., 23 students in the selected-response condition and 26 students in the constructed-response condition) made it necessary to confirm the equality of the error variances for the two conditions. I used Levene's test of equality of error variances (Glass & Hopkins, 1996) to identify unequal error variances. When necessary, I conducted Welch's t test to confirm any statistically significant group differences, thus controlling for the inequality in variances (Glass & Hopkins; Myers & Well, 1995).

$p = .000$, $\hat{\omega}^2 = .49$; and (b) a three-way interaction among assessment format, trial, and syllogism type, indicating that the effect of format was not consistent across the two time trials for categorical and conditional syllogisms, $F(1, 47) = 5.915$, $p = .019$, $\hat{\omega}^2 = .11$. As shown in Table 2, the selected-response format had a facilitative effect for both easy categorical (5.82) and easy conditional syllogisms (8.00) at Trial 1. Although this facilitative effect continued and proved stronger for easy categorical syllogisms (6.96) at Trial 2, it did not continue for easy conditional syllogisms (7.48) at Trial 2. In fact, the constructed-response format proved to be more facilitative of performance for conditional syllogisms at Trial 2.

Difficult Categorical and Conditional Reasoning

I conducted a 2 (assessment format: selected response vs. constructed response) × 2 (logic training: Trial 1 vs. Trial 2) × 2 (syllogism type: categorical vs. conditional) mixed ANOVA to examine students' performance on difficult problems. The results of the analysis indicated three significant effects: (a) a significant main effect of trial, indicating that students' overall performance on difficult syllogisms was better after training than before training, $F(1, 47) = 13.458$, $p = .001$, $\hat{\omega}^2 = .22$; (b) a two-way interaction between format and syllogism type, indicating that the selected-response format led to better performance on difficult categorical syllogisms but not to better performance on difficult conditional syllogisms, $F(1, 47) = 9.115$, $p = .004$, $\hat{\omega}^2 = .16$, where the constructed-response format led to better performance; and (c) a two-way interaction between trial and syllogism type, indicating that training facilitated performance more for difficult categorical syllogisms[6] than for difficult conditional syllogisms, $F(1, 47) = 7.042$, $p = .011$, $\hat{\omega}^2 = .13$ (see Table 2).

Discussion

These preliminary results suggest three conclusions: (a) Students performed better on easy conditional syllogisms than on easy categorical syllogisms, (b) training in symbolic logic was especially associated with improved performance on difficult categorical syllogisms and less so for difficult conditional syllogisms, and (c) the assessment format was important in measuring this improved performance—the selected-response format was better suited to measuring categorical

[6]The improvement in categorical reasoning performance is unlikely to be due to general statistical regression to the mean because participating students were not preselected based on extremely low (or high) Trial 1 scores (Cook & Campbell, 1979). In addition, if regression to the mean were responsible for these results, there also should have been a significant improvement on difficult conditional reasoning problems. However, this pattern was not observed (Cook & Campbell).

TABLE 1. Set of 16 Conditional Syllogism Tasks and 16 Categorical Syllogism Tasks

Conditional syllogism tasks			
MP Low difficulty (V)	DA High dfficulty (NVC)	AC High difficulty (NVC)	MT High difficulty (V)
If A then B A Not A ? ?	If A then B B Not B ? ?	If A then B	If A then B
If not A then B Not A A ? ?	If not A then B B Not B ? ?	If not A then B	If not A then B
If A then not B A Not A ? ?	If A then not B Not B B ? ?	If A then not B	If A then not B
If not A then not B Not A A ? ?	If not A then not B Not B B ? ?	If not A then not B	If not A then not B

Categorical syllogism tasks

One-model Low difficulty (V)	Two-model High dfficulty (NVC)	Two-model High difficulty (V)	Three-model High difficulty (V)
Some B are A All B are C ?	Some A are not B No B are C ?	Some A are not B All C are B ?	Some A are B No C are B ?
No A are B All C are B ?	No B are A Some C are not B ?	Some B are not A All B are C ?	Some B are A No B are C ?
Some A are B All B are C ?	Some B are A Some B are C ?	All A are B Some C are not B ?	No A are B Some B are C ?
All B are A Some C are B ?	Some A are not B Some C are B ?	All B are A Some B are not C ?	All B are A No C are B ?

Note. V = valid (determinate); NVC = invalid (indeterminate); MP = *modus ponens*; DA = denial of the antecedent; AC = affirmation of the consequent; MT = *modus tollens*.

TABLE 2. Study 1: Students' Trial 1 and Trial 2 Average Scores on Easy and Difficult Syllogisms, by Syllogism Type and Task Format

	Trial 1				Trial 2			
	Categorical		Conditional		Categorical		Conditional	
Response	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
				Easy syllogisms				
Selected	5.82	1.696	8.00	.000	6.96	1.581	7.48	1.729
Constructed	5.54	2.731	7.70	.928	5.62	1.878	8.00	.000
Total	5.68	2.286	7.84	.688	6.24	1.854	7.76	1.199
				Difficult syllogisms				
Selected	2.48	1.806	2.08	2.661	3.61	2.061	2.60	2.658
Constructed	1.15	1.120	3.30	3.147	2.69	2.379	3.46	3.373
Total	1.78	1.611	2.74	2.963	3.12	2.260	3.06	3.058

Note. $n = 23$ and 26 for the selected-response and the constructed-response conditions, respectively. All scores are out of maximum of 8.00.

reasoning than conditional reasoning, with the constructed-response format appearing to be better for measuring conditional reasoning. This interaction between format and syllogism type suggests that the skills associated with categorical reasoning (reasoning about category membership) might be more complex and might require more time to learn well than the skills associated with conditional reasoning (Noice & Noice, 2002). To this end, it might not be sufficient to rate the difficulty of categorical and conditional syllogisms according to the mental representations required. Although both difficult categorical and conditional syllogisms might require multiple representations for their solution, the qualitative nature of these representations might be different, suggesting subtle sources of difficulty that need to be explored to improve instruction.

These results support Morris and Nisbett (1993), who found that third-year undergraduate philosophy students improved in their responses to categorical syllogisms but not in their responses to conditional syllogisms compared with first-year undergraduate philosophy students. The Morris and Nisbett study and the present study illustrate that categorical reasoning, despite its complexity, might be more compliant to direct training than conditional reasoning. In addition, these results support Cheng et al. (1986), who found that students failed to improve their conditional reasoning by falling trap to conditional fallacies. Despite the support the present results provide to previous studies, students' overall per-

formance on difficult syllogisms was modest, suggesting that learning to reason systematically on challenging tasks requires sustained training. Alexander et al. (2004) made this point with their model of domain learning: Improved performance within a domain can be observed with specific training; however, genuine expertise within a domain requires continual effort.

Students' performance in Study 1 signaled the importance of examining students' reasoning strategies and whether their modest performance on difficult problems mirrored stagnancy in strategic thinking. In an effort to follow up on these preliminary results, I conducted a second study. This time, in addition to training another sample of undergraduate philosophy students, I asked students to report their strategy for solving each set of syllogisms. Students reported their strategy by answering a question that was added at the end of each set of syllogisms.

STUDY 2

Method

Participants

The second study took place in the fall/winter of 2003 and involved 75 undergraduate students (mean age = 21.06 years, *SD* = 2.166) enrolled in a first-year introductory philosophy course in symbolic logic at a large research-intensive university. In this introductory philosophy course, students used the textbook *Understanding Symbolic Logic* by Virginia Klenck (2001) and were taught by an experienced, tenured professor in the content area. As with Study 1, the appropriate book chapters were used to teach the various kinds of reasoning targeted in the present study. For example, the course included a discussion of sentential logic, monadic predicate logic, quantifiers, categorical propositions, proofs in predicate logic, and a discussion of relational predicate logic. Of the 75 students, 39 were men and 36 were women. As mentioned in the description of Study 1, the gender distribution is reported here for completeness but was not a variable of interest in the present article. As an indicator of their interests, 57% of students declared their major to be in the faculty of arts, and 42% declared their major to be in the faculty of science, with 1% declaring a major in the faculty of business, faculty of education, or undecided. Students were given research credit for their participation.

Procedure

I assessed the students within the first week of their philosophy course (logic training Trial 1) and again within the last week of their course (logic training Trial 2). Seventy-five philosophy students were assessed at Trial 1 and from

these, 67 students were tested again at Trial 2 (5 men and 3 women had dropped out by Trial 2). All analyses were, therefore, conducted with the 67 remaining students.

The design and materials for Study 2 were identical to the design and materials for Study 1, with the exception that I asked all students in Study 2 to report their strategy for solving the set of 16 categorical syllogisms and the set of 16 conditional syllogisms. A selected-response question was used to request students' strategies. This question was presented on a separate page, resulting in two additional pages being added to the test booklets (once after the categorical syllogisms and again after the conditional syllogisms). For example, after solving the set of categorical syllogisms, students were presented with the following question and list of alternatives:

> From the following descriptions, please choose the one that BEST describes the way or method you used or followed to answer the pairs of statements for TASK 1:
>
> (a) I used rules or statements to answer the pairs of statements.
> (b) I used pictures or models to answer the pairs of statements.
> (c) I guessed my answers to the pairs of statements.
> (d) I used a mixture of rules and pictures to answer the pairs of statements.
> (e) I don't know.
> (f) Other _____

Untrained individuals, some who are successful and some who are not, commonly report using these alternatives as strategies in solving categorical and conditional reasoning tasks (Galotti, 1989; Galotti et al., 1986; Johnson-Laird, 1999; Johnson-Laird & Bara, 1984; Rips, 1994; Roberts, 2004; Roberts & Newton, in press). For instance, rule-based strategies are symbolic in nature and normally take the form of production-system commands such as *IF <condition> THEN <action>*. Picture or model-based strategies are spatial in nature and normally take the form of tokens or Euler circles (see Figure 2). A mixture of rules and models suggests a combination of production commands with spatial tokens.

Given the predominance of these reported strategies in previous studies (Galotti; Johnson-Laird; Johnson-Laird & Bara; Rips; Roberts & Newton), I decided it was reasonable to assume that these alternatives are understandable to individuals and were, therefore, included as options for students to select in the current study (Johnson-Laird & Bara; Rips; Roberts & Newton). The other alternatives ("I guessed," "I don't know") were also included to allow students to self-report not knowing how they solved the syllogisms. Finally, for students who did not want to choose from one of the alternatives, I included the option for them to write in their strategy. However, very few students chose this alternative and, when they did, it was to elaborate on either a rule-based or model-based strategy.

Results

I present the results in two parts. First, students' performance on categorical and conditional syllogisms is presented. These analyses are analogous to those conducted for Study 1. Alongside each statistically significant effect is its *p* value and an index of its effect size noted in the text ($\hat{\omega}^2$), a measure of treatment magnitude or proportion of explained variance (Keppel, 1991). Second, students' strategies on categorical and conditional syllogisms are presented. Alongside each statistically significant effect is its *p* value and an index of its effect size ($\hat{\eta}^2$), a measure of proportion of explained variance noted in the text (Glass & Hopkins, 1996). According to Cohen (1977) and Keppel, proportions of explained variance associated with values of approximately .15, .06, and .01 are considered large, medium, and small, respectively.

Easy Categorical and Conditional Reasoning

I used a 2 (assessment format: selected response vs. constructed response) × 2 (logic training: Trial 1 vs. Trial 2) × 2 (syllogism type: categorical vs. conditional) mixed ANOVA to analyze student performance on easy problems. As with Study 1, assessment format was a between-subject variable, and logic training and syllogism type were within-subject variables. Of the 75 philosophy students

TABLE 3. Study 2: Students' Trial 1 and Trial 2 Average Scores on Easy and Difficult Syllogisms, by Syllogism Type and Task Format

	Trial 1				Trial 2			
	Categorical		Conditional		Categorical		Conditional	
Response	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
				Easy syllogisms				
Selected	6.187	2.402	7.187	2.086	6.126	2.685	7.250	2.200
Constructed	5.314	2.518	7.886	.471	5.542	2.571	7.828	.568
Total	5.731	2.484	7.552	1.510	5.821	2.622	7.552	1.589
				Difficult syllogisms				
Selected	3.125	1.897	3.750	3.162	4.031	2.221	4.626	3.230
Constructed	1.600	1.684	2.743	3.184	2.200	1.795	3.714	3.793
Total	2.328	1.934	3.224	3.190	3.075	2.197	4.149	3.539

Note. n = 32 and 35 for the selected-response and the constructed-response conditions, respectively. All scores are out of maximum of 8.00.

who participated in Trial 1, 67 remained in the class and participated in Trial 2. All analyses were, therefore, conducted with only the 67 remaining students. The results of the analysis indicated two significant effects (Table 3): (a) a significant main effect of syllogism type, indicating that students' performance was generally better on easy conditional syllogisms than on easy categorical syllogisms, $F(1, 65) = 44.671$, $p = .000$, $\hat{\omega}^2 = .41$; and (b) a significant two-way interaction between format and syllogism type, $F(1, 65) = 6.840$, $p = .011$, $\hat{\omega}^2 = .10$. As illustrated in Table 3, this interaction indicates that the selected-response format facilitated performance on easy categorical syllogisms but not on easy conditional syllogisms; in fact, performance on easy conditional syllogisms was facilitated by the constructed-response format. The limited facilitative effect of the selected-response format was also observed in Study 1.

Difficult Categorical and Conditional Reasoning

I used a 2 (assessment format: selected response vs. constructed response) × 2 (logic training: Trial 1 vs. Trial 2) × 2 (syllogism type: categorical vs. conditional) mixed ANOVA to analyze students' performance on difficult problems. The results of the analysis indicated three significant effects: (a) a significant main effect of assessment format, indicating that students performed better when they selected their responses to difficult syllogisms than when they constructed their responses, $F(1,65) = 7.469$, $p = .008$, $\hat{\omega}^2 = .10$; (b) a significant main effect of trial, indicating that philosophy students performed better after training than before training, $F(1, 65) = 9.424$, $p = .003$, $\hat{\omega}^2 = .13$; and (c) a significant main effect of syllogism type, indicating that students performed better on difficult conditional syllogisms than on difficult categorical syllogisms, $F(1, 65) = 12.819$, $p = .001$, $\hat{\omega}^2 = .17$.

Discussion

For easy syllogisms, the results suggest two general conclusions: First, students performed much better on easy conditional syllogisms than on easy categorical syllogisms, underscoring the greater complexity of the latter form of deductive reasoning. This result was also observed in Study 1. Second, students performed better when they selected their responses to easy categorical syllogisms than when they constructed their responses to the syllogisms. However, the facilitative effect of the selected-response format did not transfer well to easy conditional syllogisms. The selected-response format did not facilitate performance on easy conditional syllogisms, whereas the constructed-response format did (see Table 3). These results support research by Tulving and Thompson (1973), Watkins and Tulving (1975), and Noice and Noice (2002), suggesting that when material is well learned (e.g., easy conditional syllogisms), the constructed-response format may be the better assessment method to use.

For difficult syllogisms (see Table 3), logic training helped students improve their performance overall (as observed in Study 1). The selected-response format facilitated performance on both categorical syllogisms and conditional syllogisms, although performance was generally better on conditional syllogisms than categorical syllogisms. The facilitative effect of the selected-response format was also observed in Study 1, in which it favored performance on categorical syllogisms more than on conditional syllogisms. Even with the facilitation provided by the selected-response format, however, performance on difficult deductive tasks remained fairly weak in Study 2. That performance on difficult deductive tasks remained modest, despite a full course in logic, speaks to the need for more extensive training in complex domains such as logic (see Alexander et al., 2004). These results highlight the sensitivity of the selected-response format for assessing skills that are not yet fully mastered and may take more time to learn.

Changes in Strategies

The objective in asking students about their strategy in responding to categorical syllogisms and conditional syllogisms was to examine whether philosophy students' strategies changed after they received training in symbolic logic. If students indicated they had used (a) rules, (b) a mixture of rules and models, or (c) models, then I classified them as using a strategy. These strategies are documented in the reasoning literature as known methods for solving categorical and conditional syllogisms (Roberts & Newton, in press). I used McNemar's symmetry chi-square test to determine whether philosophy students changed the frequency of strategy use for solving categorical syllogisms and conditional syllogisms between Trial 1 and Trial 2 (see Table 4). McNemar's symmetry test is used to check the associations of proportions between paired samples (Glass & Hopkins, 1996). Results from this test indicated that philosophy students did not

TABLE 4. Study 2: Frequency and Percentage of Students Using a Strategy, by Syllogism Type and Trial (N = 67)

Trial 1				Trial 2			
Categorical		Conditional		Categorical		Conditional	
n	%	*n*	%	*n*	%	*n*	%
49	73	47	70	56	84	59	88

Note. Students who selected using rules, a mixture of rules and models, or models were coded as having used a strategy.

TABLE 5. Study 2: Frequency and Percentage of Students Using a Specific Strategy, by Syllogism Type and Trial (N = 67)

	Trial 1						Trial 2					
	Rules		Mix		Models		Rules		Mix		Models	
Syllogism type	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Categorical	22	33	22	33	5	7	12	18	28	42	16	24
Conditional	31	46	14	21	2	3	40	60	16	24	3	4

increase their frequency of strategy use on categorical syllogisms from Trial 1 to Trial 2 (73% vs. 84%), $\chi^2(1, N = 67) = 3.267$, $p = .071$. However, there were significant differences in the types of strategies used between Trial 1 and Trial 2. For example, there was a statistically significant decrease in rule use (33% vs. 18%), $\chi^2(1, N = 67) = 4.545$, $p = .033$, $\hat{\eta}^2 = .068$, and a significant increase in model use (7% vs. 24%), $\chi^2(1, N = 67) = 5.762$, $p = .016$, $\hat{\eta}^2 = .086$, on categorical syllogisms (see Table 5). There was no statistically significant change in mixed strategy use (33% vs. 42%) on categorical syllogisms. For conditional syllogisms, philosophy students did demonstrate a statistically significant increase in their frequency of strategy use from Trial 1 to Trial 2 (70% vs. 88%), $\chi^2(1, N = 67) = 9.000$, $p = .003$, $\hat{\eta}^2 = .134$. Specifically, there was a statistically significant increase in rule use (46% vs. 60%), $\chi^2(1, N = 67) = 4.263$, $p = .039$, $\hat{\eta}^2 = .064$, but no change in either mixed strategy (21% vs. 24%) or model use (3% vs. 4%).

GENERAL DISCUSSION

The goals of the present studies were to (a) investigate whether improvements in students' categorical and conditional reasoning were associated with training in symbolic logic, (b) assess the training using standard tasks in different formats, and (c) explore whether changes in reasoning strategies accompanied improved performance. Unlike previous investigations (i.e., Cheng et al., 1986; Lehman et al., 1988; Morris & Nisbett, 1993), a primary aim of the present studies was to assess the effectiveness of deductive training with standard deductive tasks in an ecologically valid setting (a classroom environment), where learning gains are assumed of students.

The present findings suggest that training students in the domain-specific knowledge of symbolic logic improves performance but by modest amounts. The lack of improvement on easy categorical syllogisms was surprising, given that students were not initially at ceiling on these syllogisms and had room to improve (see Table 3). Philosophy students did improve their performance on difficult syllogisms, especially when students had the opportunity to select their re-

sponses. The improvement in categorical reasoning was distinctively paralleled by a significant decrease in reported rule-based strategies and a significant increase in reported model-based strategies. The improvement in conditional reasoning was paralleled by a significant increase in rule-based strategies. Some philosophy professors have suggested that model-based strategies are better suited for solving categorical syllogisms, whereas rule-based strategies are better suited for solving conditional syllogisms (O. Simchen, personal communication, March 8, 2000; A. Stairs, personal communication, February 13, 2001). Models are easier to use with categorical syllogisms because there are no simple rule-based methods for dealing with quantifiers such as "all" or "some." Alternatively, rules are easier to use with conditional syllogisms because rules are straightforward to apply in this case.

The tailoring of strategies for different kinds of tasks occurs with gains in domain-specific knowledge (see Roberts & Newton, in press; Siegler, 1996). This tailoring occurs because students discover, through training and deliberate practice, that some strategies are more likely to lead to successful results than other strategies with specific classes of problems. With deductive problems, this tailoring appears to be related to training—more experience within a domain increases the likelihood of being exposed to and discovering better strategies for a given set of problems.

As in other domains, it is likely that initial domain-specific training in logic has more of an effect in the selection of strategies considered to solve deductive reasoning problems and less of an effect in the actual performance (Anderson, 1990; Siegler, 1996). For example, philosophy students seemed to have capitalized on their training by using a model-based strategy that is considered more appropriate than a rule-based strategy when solving categorical syllogisms. Trained philosophy students changed their strategy in the desired direction.

In terms of the assessment of categorical and conditional reasoning, most of the analyses revealed that students performed better when they were allowed to select their responses than construct their responses. The superiority of the selected-response format was not due to students' misunderstanding of how to frame their responses in the constructed-response condition—that is, failing to understand the appropriate quantifiers to use or propositional structure of the response. The instructions to the task illustrated examples of the responses. Furthermore, in the data analyses, I examined the test booklets in the constructed-response conditions and did not find that responses included irregular quantifiers or propositions but, rather, included regular quantifiers or propositions used incorrectly (e.g., the use of "some" instead of "some not").

So what can be said for the superiority of the selected-response format in the context of assessing categorical and conditional reasoning? In general, the selected-response format appears to be a more sensitive measure of students' deductive reasoning skills on difficult problems, suggesting the newness and per-

haps reticent nature of systematically generating multiple representations of information (Noice & Noice, 2002). Selected-response formats have the advantage of making the problem space transparent to students—all alternatives are made available and salient. The task for students to mentally represent all alternatives is lessened. Given the strain on working memory of having to consider multiple representations of information, having the alternatives made available and salient may make the reasoning process easier to handle (especially for difficult problems). Students can simply focus on one alternative at a time and decide whether is it a necessary conclusion or not.

Of course, given the limited (and predictable) set of possible conclusions that can be generated for these formal syllogisms, it would not be difficult for students to generate all conclusions quickly in the constructed-response condition. So why would they not generate the list of five alternatives and engage in the same procedure as in the selected-response format? It is possible that the constructed-response format elicits a response set, in which fewer alternatives than the limited set are considered (a form of satisficing; see Simon, 1955). Alternatively, instead of working backward from the alternatives, students in the constructed-response condition may choose to work forward, increasing the complexity of the task for themselves. More research is certainly needed into the metacognitive strategies invoked by selected-response versus constructed-response formats in students. It is important to note, however, that the selected-response format was not always a more sensitive measure of students' reasoning. It depended on the type of deductive problem—for example, the selected-response format was a more sensitive measure of easy categorical reasoning, but the constructed-response format appeared to be the more sensitive measure of easy conditional reasoning. The differential sensitivity of these formats suggests that categorical reasoning, even on easy tasks, is more challenging than conditional reasoning.

In sum, training deductive skills in a classroom over the course of a semester did not enhance performance as much as expected. However, training did alter the kinds of strategies students considered and used to solve certain classes of problems. In particular, trained students increased their use of models and decreased their use of rules on categorical syllogisms and increased their use of rules on conditional syllogisms. Although these changes in strategy use did not lead to impressive improvements in reasoning, changes in strategy use are considered indicators of emerging expertise (Alexander et al., 2004; Siegler, 1996); therefore, these changes may signal emerging knowledge or mastery nonetheless. This study, placed in an ecological valid setting—the classroom—underscores what other studies have found in the laboratory and under less ecologically valid settings: Learning deductive skills is not easy, and a single course or training workshop is insufficient to elevate performance substantially (see Alexander et al.). Teaching deductive skills might be initiated earlier in secondary school so

that the length of training needed to establish these forms of thinking is afforded to students. Furthermore, more research is needed into the cognitive processes that monitor systematic thinking and the conditions that promote these processes. Understanding how to avoid reasoning biases and develop systematic thinking in students of all backgrounds early in their educational experiences will increase their opportunities in careers that demand the application of this demanding skill.

REFERENCES

Alexander, P. A., Sperl, C. T., Buehl, M. M., Fives, H., & Chiu, S. (2004). Modeling domain learning: Profiles from the field of special education. *Journal of Educational Psychology, 96*, 545–557.

Allard, F., & Starkes, J. L. (1991). Motor-skill experts in sports, dance, and other domains. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 126–152). New York: Cambridge University Press.

Anderson, J. R. (1990). *Cognitive psychology and its implications* (3rd ed.). New York: Freeman.

Anderson, K. C., & Leinhardt, G. (2002). Maps as representations: Expert novice comparison of project understanding. *Cognition & Instruction, 20*, 283–322.

Barnett, S. M., & Koslowski, B. (2002). Adaptive expertise: Effects of type of experience and the level of theoretical understanding it generates. *Thinking and Reasoning, 8*, 237–267.

Baron, J. (2000). *Thinking and deciding* (3rd ed.). New York: Cambridge University Press.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science, 23*, 247–303.

Byrnes, J. P. (2001). *Cognitive development and learning* (2nd ed.). Boston: Allyn and Bacon.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*, 1098–1120.

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

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally College Publishing.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187–276.

de Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton.

Ericsson, K. A. (1996). The acquisition of expert performance. In K. A. Ericsson (Ed.), *The road to excellence* (pp. 1–50). Mahwah, NJ: Erlbaum.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*, 725–747.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*, 211–245.

Evans, J. St. B. T. (2000). What could and could not be a strategy in reasoning? In W. Schaeken, G. De Vooght, A. Vandierendonck, & G. d'Ydewalle (Eds.), *Deductive reasoning strategies* (pp. 1–22). Mahwah, NJ: Erlbaum.

Evans, J. St. B. T., & Brooks, P. G. (1981). Competing with reasoning: A test of the working memory hypothesis. *Current Psychological Research, 1,* 139–147.

Evans, J. St. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64*, 391–397.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.

Ford, M. (1995). Two modes of mental representation and problem solving in syllogistic reasoning. *Cognition, 54*, 1–71.

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin, 105*, 331–351.

Galotti, K. M., Baron, J., & Sabini, J. P. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology: General, 115*, 16–25.
Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn and Bacon.
Gobet, F., & Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology, 31*, 1–40.
Hardman, D. (1998). Does reasoning occur in the selection task? A comparison of relevance-based theories. *Thinking and Reasoning, 4*, 353–376.
Hurley, P. J. (2002). *A concise introduction to logic*. Belmont, CA: Wadsworth.
Johnson-Laird, P.N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, MA: Harvard University Press.
Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology, 50,* 109–135.
Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16*, 1–61.
Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
Kirby, K. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition, 51*, 1–28.
Klaczynski, P., Byrnes, J. P., & Jacobs, J. E. (2001). Introduction to the special issue. The development of decision making. *Journal of Applied Developmental Psychology, 22*, 225–236.
Klaczynski, P. A., & Narasimham, G. (1998). Development of scientific reasoning biases: Cognitive versus ego-protective explanations. *Developmental Psychology, 34*, 175–187.
Klauer, K. C., Stegmaier, R., & Meiser, T. (1997). Working memory involvement in propositional and spatial reasoning. *Thinking and Reasoning, 3*, 9–47.
Klenck, V. (2001). *Understanding symbolic logic* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.
Kuhn, D. (1992). Piaget's child as scientist. In H. Beilin & P. Pufall (Eds.), *Piaget's theory: Prospects and possibilities* (pp. 185–210). Hillsdale, NJ: Erlbaum.
Kuhn, D. (2001). Why development does (and does not) occur: Evidence from the domain of inductive reasoning. In J. L. McClelland & R. Siegler (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives* (pp. 221–249). Hillsdale, NJ: Erlbaum.
Kuhn, D., Garcia-Mila, M., Zohar, A., & Anderson, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, 60* (4, Serial No. 245).
Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday life events. *American Psychologist, 43*, 431–443.
Leighton, J. P. (2001, April). *An analysis of students' hypothesis testing skills.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
Leighton, J. P., & Sternberg, R. J. (2003). Reasoning and problem solving. In A. F. Healy & R. W. Proctor (Vol. Eds.), *Experimental psychology* (pp. 623–648). New York: Wiley.
Leighton, J. P., & Sternberg, R. J. (Eds.). (2004). *The nature of reasoning*. New York: Cambridge University Press.
Markovits, H., Doyon, C., & Simoneau, M. (in press). Individual differences in working memory and conditional reasoning with concrete and abstract content. *Thinking and Reasoning*.
Meiser, T., Klauer, K. C., & Naumer, B. (2001). Propositional reasoning and working memory: The role of prior training and pragmatic content. *Acta Psychologica, 106,* 303–327.
Morris, A. (2002). Mathematic reasoning: Adults' ability to make the inductive-deductive distinction. *Cognition & Instruction, 20*, 79–118.
Morris, M. W., & Nisbett, R. E. (1993). Tools of the trade: Deductive schemas taught in psychology and philosophy. In R. E. Nisbett (Ed.), *Rules for reasoning* (pp. 228–256). Hillsdale, NJ: Erlbaum.
Morse, L. W., & Morse, D. T. (1995). The influence of problem-solving strategies and previous training on performance of convergent and divergent thinking. *Journal of Instructional Psychology, 22*, 341–348.
Myers, J. L., & Well, A. D. (1995). *Research design and statistical analysis*. Hillsdale, NJ: Erlbaum.
Mynatt, C. R., Doherty, M. E., & Dragan, W. (1993). Information relevance, working memory, and the consideration of alternatives. *The Quarterly Journal of Experimental Psychology, 46A,* 759–778.

National Center for Education Statistics. (2003). *Overview and inventory of state education reforms: 1990–2000*. Retrieved August 5, 2002, from http://nces.ed.gov/pubs2003/2003020.pdf

Newton, E. J., & Roberts, M. J. (2000). An experimental study of strategy development. *Memory and Cognition, 28,* 565–573.

Nickerson, R. S. (2003). *Psychology and environmental change*. Mahwah, NJ: Erlbaum.

Noice, T., & Noice, H. (2002). Very long-term recall and recognition of well-learned material. *Applied Cognitive Psychology, 16,* 259–272.

Overton, W. F. (1990). Competence and procedures: Constraints on the development of deductive reasoning. In W. F. Overton (Ed.), *Reasoning, necessity, and logic: Developmental perspectives* (pp. 1–32). Hillsdale, NJ: Erlbaum.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development, 15*, 1–12.

Powers, D. E., & Dwyer, C. A. (2003). *Toward specifying a construct of reasoning*. Princeton, NJ: Educational Testing Service.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Roberts, M. J. (1993). Human reasoning: Deduction rules or mental models, or both? *The Quarterly Journal of Experimental Psychology, 46A*, 569–589.

Roberts, M. J. (1997). On dichotomies and deductive reasoning research. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16*, 196–204.

Roberts, M. J. (2000a). Individual differences in reasoning strategies: A problem to solve or an opportunity to seize? In W. Schaeken, G. De Vooght, A. Vandierendonck, & G. d'Ydewalle (Eds.), *Deductive reasoning strategies* (pp. 23–48). Mahwah, NJ: Erlbaum.

Roberts, M. J. (2000b). Strategies in relational inference. *Thinking and Reasoning, 6*, 1–26.

Roberts, M. J. (2004). Heuristics and reasoning I: Making deduction simple. In J. P. Leighton & R. J. Sternberg (Eds.), *Nature of reasoning* (pp. 234–272). New York: Cambridge University Press.

Roberts, M. J., & Newton, E. J. (in press). Strategy usage in a simple reasoning task: Overview and implications. In M. J. Roberts & E. J. Newton (Eds.), *Methods of thought: Individual differences in reasoning strategies.* New York: Psychology Press.

Schaeken, W., De Vooght, G., Vandierendonck, A., & d'Ydewalle, G. (Eds.). (2000). *Deductive reasoning strategies*. Mahwah, NJ: Erlbaum.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Siegler, R. S., & Jenkins, E. A. (1989). *How children discover new strategies*. Hillsdale, NJ: Erlbaum.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69,* 99–118.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 57*, 31–95.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science, 19*, 97–140.

Sternberg, R. J., & Ben-Zeev, T. (2001). *Complex cognition: The psychology of human thought*. New York: Oxford University Press.

Toms, M., Morris, N., & Ward, D. (1993). Working memory and conditional reasoning. *The Quarterly Journal of Experimental Psychology, 46A,* 679–699.

Torrens, D., Thompson, V. A., & Cramer, K. M. (1998). Individual differences and the belief bias effect: Mental models, logical necessity, and abstract reasoning. *Thinking and Reasoning, 5*, 1–28.

Tulving, E., & Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 359–380.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293–315.

Ward, S. L., & Overton, W. F. (1990). Semantic familiarity, relevance, and the development of deductive reasoning. *Developmental Psychology, 26*, 488–493.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology, 12,* 129–140.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Middlesex, England: Penguin.
Watkins, M., & Tulving, E. (1975). Episodic memory: When recognition fails. *Journal of Experimental Psychology: General, 104,* 5–29.