Evaluative feedback can improve deductive reasoning

Sangeet Khemlani[1] and Adam Moore[2]
{skhemlani, adam.moore457}@gmail.com

[1]Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC 20375 USA
[2]Center for Advanced Brain Imaging, Georgia Institute of Technology, Atlanta, GA 30318 USA

Abstract

We examine whether reasoning is improved by evaluative feedback, i.e., the information of whether a reasoner's answer was correct or incorrect, and report two studies that show that evaluative feedback increases the chances that participants will produce normatively correct responses for deductive reasoning problems. In Experiment 1, participants who were given feedback about their performance did better on problems based on disjunctions that were designed to elicit illusory inferences. In Experiment 2, participants answered difficult syllogisms with more accuracy when they were provided with feedback. We conclude by contrasting the rule-, heuristics-, and model-based accounts of deduction on their ability to explain the effects of evaluative feedback.

Keywords: feedback, reasoning, illusory inferences, syllogisms

Introduction

People often receive feedback after they have drawn an inference. Feedback can manifest in a contrarian's objection, a pat on the back, a heated argument, or a grunt of disapproval. In many cases, feedback can be *prescriptive,* i.e., it can be accompanied by further instructions and suggestions for improvement, such as what one might receive in a classroom environment. In other cases, feedback can be *evaluative* and devoid of any pedagogical value, such as a final grade in a course. Prescriptive feedback has been shown to improve participants' reasoning on a wide variety of tasks (Cheng, Holyoak, Nisbett, & Oliver, 1986; Khemlani & Johnson-Laird, 2009; Leevers & Harriss, 1999). The effect is robust but unsurprising: if prescriptive feedback could not make better reasoners out of humans, it would be difficult to explain the internalization of rules, heuristics, and insights. Our investigation focuses instead on evaluative feedback, i.e., feedback that neither explains nor characterizes performance as any more than a minimal description of whether performance was correct or incorrect on a particular trial (Neth, Khemlani, & Gray, 2008). We are interested in this impoverished form of feedback because it is unclear what effect, if any, it should have on a person's future performance on similar problems (Klayman & Ha, 1987; Wason, 1960).

Suppose reasoners were told that the deduction they drew from a set of premises was incorrect. Since they have no further information on why their answers were incorrect, it is not clear that the feedback could apply to different sets of premises. Reasoners may remember the structure of the premises so that if they encounter the same problem again, they can provide a correct answer, but there is little reason to think that the feedback should produce any systematic improvement in reasoning beyond correcting an answer to a particular problem unless reasoners directly search for an explanation of why they went wrong (Walsh & Johnson-Laird, 2009). Moreover, given multiple-choice problems in which the elimination of one answer does not identify the correct answer, evaluative feedback might produce no effect whatsoever. Since memories are susceptible to interference and decay, it is uncertain whether evaluative feedback will have any impact on the ability to solve related but syntactically different problems. Few studies have examined how immediate evaluative feedback informs reasoning (but cf. Wason, 1964), and few psychological theories of reasoning explicitly permit evaluative feedback to modulate the way individuals reason (Braine & O'Brien, 1998; Oaksford & Chater, 2007; Rips, 1994; Stenning & Van Lambalgen, 2008) though there is evidence that the first thing individuals do upon learning that their conclusion is incorrect is to check their reasoning (Johnson-Laird, Girotto, & Legrenzi, 2004).

If feedback influences the way people make deductions, theories of reasoning ought to accommodate such effects by showing how individuals make use of the additional information with which they are supplied. In the following experiments, we show that immediate, evaluative feedback improves the way individuals reason. We conclude by explaining how three prominent theories of reasoning might account for improvements in performance due to evaluative feedback.

Experiment 1: Sentential reasoning

Experiment 1 presented participants with a set of problems that were expected to yield "illusory" sentential inferences. Sentential inferences are those based on sentential connectives such as *and* (a conjunction) and *or* (a disjunction). Illusory inferences are systematic errors that are produced when people fail to consider all of the possibilities consistent with the premises (Khemlani & Johnson-Laird, 2009). Each problem was based on a set of premises in which one disjunction was embedded within another. The disjunctions were either exclusive or inclusive; for example, consider these premises based on two exclusive disjunctions:

Suppose one of the following assertions is true and one is false:
1. You have the blue candies and the red candies.
2. You have the red candies or else the orange candies, but not both.

Is it possible to have the blue candies and the orange candies only?

Table 1: The four types of problem in Experiment 1, their premises and corresponding questions, the predicted conclusions, and the correct conclusions to each question.

Problem			Conclusion	
Type	Premises	Question	Predicted	Correct
Exclusive-exclusive	One is true and one is false: 1. A and B. 2. B or else C	Is it possible to have A and B only?	Yes	No
Exclusive-inclusive	One is true and one is false: 1. A and B. 2. B or C or both.	Is it possible to have A and B?	Yes	No
Inclusive-exclusive	One or both are true: 1. A and B. 2. B or else C.	Is it possible to have A and C only?	No	Yes
Inclusive-inclusive	One or both are true: 1. A and B. 2. B or C or both.	Is it possible to have A and C only?	No	Yes

The rubric makes clear that there is an exclusive disjunction between assertions 1 and 2. In previous studies participants tended to respond "no", that it was not possible to have only the blue and orange candies (Khemlani & Johnson-Laird, 2009, Experiments 2 and 3). The answer is illusory, however; the premises allow for the possibility of having only blue and orange candies. The difficulty of problems that yield illusory inferences is robust; even when participants received remedial instructions that explained how to overcome the illusions, they made errors more often than not.

In the present study, participants were provided feedback about their responses. They were randomly assigned to two different feedback conditions: *feedback*, in which participants were informed about whether their answers were correct or incorrect; and *no feedback*, in which they received no information about their performance but rather continued to the next problem after a brief delay.

Method

Participants and design. 53 volunteers were recruited through a platform hosted by Amazon.com through which people participate in experiments over the Internet for monetary compensation (for a discussion on the validity of results from this platform, see Paolacci, Chandler, & Ipeirotis, 2010). None of the participants had received any training in logic. They received four sorts of problems based on disjunctive premises, and all of the problems were designed to elicit an illusory inference. Table 1 presents the four sorts of problems, each of which was presented twice using different materials. We tested two groups of participants; one group received feedback on their answers and the other did not.

Procedure and materials. On each trial, participants received a disjunctive set of premises and a question that was intended to elicit a fallacious response. Participants then selected buttons marked "Yes" or "No". Once the participant responded, there was a delay for 2 seconds during which feedback, if appropriate, was displayed on the screen. In the no feedback condition, participants received just a delay before moving on to the next problem. Whenever feedback was given to a participant, it replaced the text of the premises and conclusion so that participants did not have access to the problem itself, and could not re-evaluate the premises. The materials used in the study pertained to various combinations of colored candy, and participants received each set of materials only once.

Results and discussion

Table 2 presents the percentages of correct responses for each group of participants. Participants found the problems quite difficult, and produced correct responses 30% of the time. They made more correct responses when presented feedback than when not (Mann-Whitney test: 38% vs 21%, $z = 3.00$, $p < .0001$).

Table 2: The percentage of correct responses to the four types of problem in Experiment 1 as a function of the type of feedback received.

	Received feedback?	
Problem Type	Yes	No
Exclusive-exclusive	28	12
Exclusive-inclusive	28	8
Inclusive-exclusive	55	27
Inclusive-inclusive	41	35

Participants in Experiment 1 performed better when presented evaluative feedback about the correctness or incorrectness of their answers on problems designed to yield illusory inferences. Likewise, performance did not differ as a function of the order in which the problems were presented; participants in the feedback condition did not do better on the last three trials compared to the first three trials in the experiment (33% vs. 37%, Wilcoxon test, $z = .68$, $p = .49$).

Table 3: The premises of the fourteen types of syllogistic problems used in Experiment 2 and a set of candidate conclusions, which include: a correct conclusion that necessarily follows from the first and second premises; a consistent conclusion that does not necessarily follow from the premises; and the most common erroneous conclusion that reasoners generate.

Problem		Candidate conclusion		
First premise	Second premise	Correct	Consistent	Common conclusion
Some A are B	No C are B	Some A are not C	Some C are not A	No A are C
All A are B	Some C are not B	Some C are not A	Some A are not C	Some C are A
No A are B	Some B are C	Some C are not A	Some A are not C	No A are C
All B are A	No B are C	Some A are not C	Some C are not A	No A are C
Some A are not B	All C are B	Some A are not C	Some C are not A	Some A are C
No B are A	Some B are C	Some C are not A	Some A are not C	No A are C
No B are A	All B are C	Some C are not A	Some A are not C	No A are C
All B are A	Some B are not C	Some A are not C	Some C are not A	Some A are C
Some B are A	No B are C	Some A are not C	Some C are not A	No A are C
All B are A	All B are C	Some A are C	All C are A	All A are C
No A are B	Some C are B	Some C are not A	Some A are not C	No A are C
No A are B	All B are C	Some C are not A	Some A are not C	No A are C
Some B are not A	All B are C	Some C are not A	Some A are not C	Some A are C
Some A are B	No B are C	Some A are not C	Some C are not A	Some A are C

One alternative explanation of the results in Experiment 1 is that instead of making participants better, feedback might have slowed them down so they could read the premises more carefully. A portion of the participants might have initially sped through the study, and if the effect of feedback was to get them to pay attention and stop responding erratically, then the results could be explained without recourse to theoretical claims about performance increases. We are skeptical of such an explanation for two reasons. First, most participants did not respond randomly; they performed reliably worse than chance. Second, every participant received a 2-second delay between trials, and so at the outset they were unable to rush through the study.

Another explanation of the results in Experiment 1 is that instead of making participants perform better, feedback made participants more erratic. The percentage of correct responses was not reliably greater than what would be expected if participants chose responses at random, which could have been driven by a reduction in participants' confidence in their initial answers due to the feedback they received. Likewise, one limitation of the present study is that erroneous disjunctive inferences, while representative of sentential reasoning, come about as a result of a tendency to overlook possibilities (Khemlani & Johnson-Laird, 2009, p. 622). To overcome these limitations, we used a different task and a more diverse set of materials in Experiment 2. Instead of having participants choose between just two alternatives, we provided participants several putative conclusions for syllogistic reasoning problems, only one of which validly followed from the premises.

Experiment 2: Syllogistic reasoning

Experiment 2 examined whether feedback could help participants discover the correct response to a syllogism from a set of alternatives. Syllogistic reasoning is logically simple but psychologically complex, and many theories have been proposed to deal with how humans process syllogisms. Modern theories of syllogistic reasoning are based on mental models (Johnson-Laird & Bara, 1984; Polk & Newell, 1995), formal rules of inference (Braine & O'Brien, 1998; Rips, 1994), or the mood of the most informative premise (Oaksford & Chater, 2007).

Not all syllogisms are created equal; some are easy and can be solved in a matter of seconds, and others are so vexing that reasoners may spend many minutes considering their premises. Consider one such problem:

All of the brewers are accountants
All of the brewers are cashiers
What must be true?

Reasoners often conclude that all accountants are cashiers, or else that no valid conclusion follows from the premises. The former conclusion is false because not all accountants are necessarily brewers. The latter is false as well, because a valid conclusion exists: it follows that some accountants are cashiers. The moral of the story is that syllogisms are not always easy to solve, and in the present study, we chose those syllogisms that pose the most trouble for reasoners (see Khemlani & Johnson-Laird, in press, for a review).

Participants were once again randomly assigned to the two feedback conditions that were used in Experiment 1, i.e., they either received feedback or did not.

Method

Participants and design. 56 volunteers were recruited from the same participant pool that was used in Experiment 1. All of the participants were untrained in logic, and they completed the experiment using an interface written in Ajax. They received fourteen syllogistic reasoning problems; Table 3 presents the premises of the problems and their

corresponding alternative conclusions. As in the previous study, we tested a group of participants who received feedback against a control group that received no feedback.

Procedure and materials. Participants took the study over the Internet, and for each problem they received two quantified premises and four alternative conclusions. The problems were taken from syllogisms identified by previous research as being the most difficult for reasoners (Bucciarelli & Johnson-Laird, 1999; Chapman & Chapman, 1959; Oaksford & Chater, 1999). One of the four alternative conclusions was correct, and the other three were distractors. The distractors consisted of a) a conclusion that is consistent with, but does not follow necessarily from, the premises; b) the most common but incorrect response that participants had spontaneously generated in previous studies (see Buciarelli & Johnson-Laird, 1999); and c) a "null" response, i.e., "no valid conclusion". The order in which the alternative conclusions were displayed on the screen was randomized.

Participants were told that only one of the four responses was correct. They registered their response by selecting buttons assigned to one of the four conclusions. When the participant responded, there was a delay for 2 seconds during which feedback, if appropriate, was displayed on the screen. Whenever feedback was given to a participant, it replaced the text of the premises. The materials used in the study pertained to various combinations of occupations, e.g., "All of the brewers are accountants," and participants received each set of materials only once.

Results and discussion

Table 4 presents the proportion of agreement to the four different types of conclusions that were presented on each trial. The problems were difficult; across the study, participants agreed to correct conclusions 39% of the time. The feedback (44%) condition yielded reliably more correct responses than the no feedback condition (44% vs. 33%, Mann-Whitney test, $z = 2.15$, $p < .05$), and this pattern held for 10 of the 12 syllogisms (Binomial test, $p < .05$). As in Experiment 1, performance in the feedback condition did not increase steadily; accuracy on the first five trials was not reliably lower than on the last five trials (41% vs. 45%, Wilcoxon test, $z = .48$, $p = .63$).

Table 4: The proportion of agreement to the four types of conclusions in Experiment 2 as a function of the type of feedback received.

	Received feedback?	
Conclusion type	Yes	No
Correct	44	33
Consistent	21	20
Common	26	26
No valid conclusion	9	16

As in Experiment 1, we consider the alternative explanation that instead of making participants perform better, feedback slowed participants down and forced them to read the premises more carefully. The present results are not consistent with this account, because regardless of presence or absence of feedback participants chose the consistent conclusion about 20% of the time. If the feedback motivated them to be more careful, then they would have made fewer errors of interpretation, and we would see a difference between the extent to which they agreed with consistent answers. The uniformity of their answers suggests that in fact, participants were reading and comprehending the problems at the same level of competence regardless of the feedback they were given.

General Discussion

Across two different paradigms calling for deductive reasoning, evaluative feedback improved performance relative to no feedback. No psychological theory of deduction is constructed to explicitly make predictions about effects of feedback. However, we conclude by examining how the principles of various theories of reasoning might be used to account for the performance gains observed in our studies.

Psychological theories of deduction fall into three broad categories: those based on formal rules akin to those in the proof theory of logic (e.g., Rips, 1994; Stenning & Van Lambalgen, 2008), those based on the processing of subjective probabilities and probabilistic heuristics (Oaksford & Chater, 2007), and those based on models akin to those in the semantic theory of logic (e.g., Johnson-Laird, 1983; Polk & Newell, 1995). Each type of theory yields a different account of how feedback might be integrated into deductive processes to improve reasoning performance.

Theories of deduction based on formal rules

Theories based on the application of formal rules of inference propose that reasoning is a process of proof in which syntactic rules are used to derive conclusions from the premises. A precursor to reasoning is accordingly the recovery of the logical form of premises to allow the application of rules. Once the logical form has been recovered, rules are applied over the formal structure of the premises to yield conclusions. Theories based on formal rules posit only those rules that allow participants to draw valid deductions, but recognize that humans often make errors in reasoning. For instance, Rips (1994, p. 386) suggests that logical errors are made more often for problems that require more steps of proof or that require complex rules to be applied to the premises. To solve a particular problem, syntactic rules must be utilized to derive a proof of its conclusion, step by step. Thus improvement in reasoning on a given problem can be explained by a) an increased tendency to recognize that a particular rule is necessary, and b) the increased frequency with which the rule is applied. Rips (1994) reports studies of such improvements. If evaluative feedback improves the way

individuals reason, then it should affect the way particular rules are recognized and applied. Thus, rules theories predict that feedback makes it easier to recover the rules relevant to the problem at hand. But it is not clear how rules theories would account for generalized performance increases based on evaluative feedback, i.e., increases in reasoning that affect many rules at once. For complex reasoning problems that require several rules to be applied, a credit assignment problem exists: a reasoner does not know, based on evaluative feedback alone, which rule has been incorrectly applied. The reasoner's performance can increase only if credit is assigned to the rule that was incorrectly applied; otherwise, it is possible that the reasoner does worse on future trials. Rules theories offer no hint at how the credit assignment problem could be overcome, but one solution is to statistically abstract the conditional relationship between the use of each particular rule in all relevant contexts and the ultimate outcome. Such a solution relies on gathering and encoding massive amounts of data, and so it is incompatible with performance increases after only a few trials.

Heuristic based theories of deduction

The theory of deduction based on probabilistic heuristics (Oaksford & Chater, 2007) assumes that individuals reason by employing simple heuristics based on informativeness and probabilistic entailment. A claim is informative if it rules out possibilities; thus, the universal statement All of the swans are white is more informative than the existential statement Some of the swans are white, because the universal rules out the possibility that some swans are not white, whereas the existential statement has no such constraint. Oaksford and Chater (2007) argue that people use heuristics based on this knowledge of informativeness to select and test conclusions along with heuristics based on probabilistic entailment, i.e., knowledge about whether one premise probabilistically follows from another. For instance, the statement All swans are white probabilistically entails the statement Some swans are white. The authors detail several ways in which individuals might apply the heuristics based on informativity and probabilistic entailment to test and derive conclusions, and show that the predictions made by the heuristics are a good fit for the difficulty of certain syllogisms, i.e., arguments with two or more quantified premises (Oaksford & Chater, 2007, Ch. 7).

Oaksford and Chater argue that for a particular syllogism, individuals construct and test conclusions based on heuristics that require the following pieces of information: 1.) a complete ordering of premises on their informativeness; 2) the quantifier of the least informative premise; 3) a complete account of probabilistic entailments; 4) the most informative premise. Oaksford and Chater suggest that (1) and (3) are immutable whereas (2) and (4) are calculated from the premises of each new problem. According to their theory, a human's departure from a normative answer provided by logic need not be suboptimal, as logic is the wrong normative baseline by which to assess rationality. If humans "err", they do so not because they do not provide the answer sanctioned by classical logic, but because they are equipped with inexpensive heuristics that are fallible. Chater and Oaksford's (1999) analysis of difficult syllogisms suggest that to provide probabilistically valid responses to syllogistic reasoning problems, reasoners are required to apply apply all of the probabilistic heuristics specified in the model. As Copeland and Radvansky (2004) observe, the need to apply more heuristics taxes working memory as it requires individuals to hold in mind both the heuristics themselves as well as the results of each heuristic. Feedback may trigger improved performance by inducing reasoners to apply all heuristics instead of just a subset.

Theories of reasoning based mental models

The mental model theory of deduction (Johnson-Laird, 1983) is based on the notion that individuals reason, both deductively and inductively, by constructing representations of possibilities. The theory proposes that the process of deduction goes through three stages: individuals first use the meaning of sentences and their knowledge to envisage what is possible given the propositions expressed in the premises, and they represent the possibilities as a single mental model. Second, the model is scanned for information not made explicit in the premises, and if any such information is found it is considered a putative conclusion. Third, individuals assess the conclusion by looking for counterexamples, i.e., alternative models of the premises where the conclusion is false. If a counterexample exists, then the conclusion is dismissed and individuals return to the second stage to construct an alternative explanation. Improved reasoning as a result of evaluative feedback can be attributed to the diligence with which individuals form models and search for counterexamples. Reasoners would then use feedback as a cue to fully flesh out multiple mental models and search for counterexamples. These processes require working memory resources to hold the relevant models in mind and operate on them. Thus, increases in executive control as a result of evaluative feedback may improve the ability to consider alternatives and search for counterexamples.

In summary, we showed how three accounts of reasoning can explain why deduction is enhanced by feedback. Mental rules theories predict that feedback must affect individual rules to improve general performance on non-identical problems; however, substantial experience would be required for the reasoner to learn the individual relations between feedback and the use of particular rules across many contexts. The probability heuristics theory holds that individuals apply a series of simple heuristics based on approximations of statistical calculations. Regardless of the normative baseline used, feedback for a particular problem should cue participants to apply all relevant heuristics instead of a subset. Mental models theory posits that individuals flesh out models and search for counterexamples in order to obtain correct answers, and can make mistakes

when they fail to do either. Thus, feedback may prompt the reasoner to search for counterexamples more assiduously, and could lead to general increases in performance across many types of problems.

The results we report demonstrate that feedback and reinforcement can improve the efficacy of conscious reasoning. Theories of reasoning can be extended to handle feedback effects explicitly in the ways we outlined above, and doing so may allow future studies to identify and test the ways in which feedback information implicitly changes reasoners' representations and inferential processes.

Acknowledgments

This research was supported by a National Science Foundation Graduate Research Fellowship to both authors, and by National Science Foundation Grant No. DRMS 0844851. We thank Jeremy Boyd, Andy Conway, Sam Glucksberg, Adele Goldberg, Geoffrey Goodwin, Matt Johnson, Phil Johnson-Laird, Mike Oaksford, and Laura Suttle for their suggestions and critiques.

References

Chapman, L. J., & Chapman, A. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology*, 58, 220–226.

Chater, N. & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, 38, 191-258.

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

Braine, M. D. S., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Lawrence Erlbaum Associates, Mahwah, NJ.

Bucciarelli, M., & Johnson-Laird, P.N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, 23 (3), 247-303.

Copeland, D. E., & Radvansky, G. A. (2004). Working memory and syllogistic reasoning. *Quarterly Journal of Experimental Psychology*, 57, 1437-1457.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N., & Bara, B. (1984). Syllogistic inference. *Cognition*, 16, 1–61.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, 111 (3), 640-661.

Khemlani, S., & Johnson-Laird, P.N. (2009). Disjunctive illusory inferences and how to eliminate them. *Memory & Cognition,* 35 (5), 615-623.

Khemlani, S., & Johnson-Laird, P.N. (in press). Theories of the syllogism: A meta-analysis. *Psychological Bulletin.*

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review*, 94, 211-228.

Leevers, H.J., & Harris, P.L. (1999). Persisting effects of instruction on young children's syllogistic reasoning with incongruent and abstract premises. *Thinking and Reasoning,* 5 (3), 145-173.

Neth, H., Khemlani, S., & Gray, W. (2008). Feedback design for the control of a dynamic multitasking system: Dissociating outcome feedback from control feedback. *Human Factors,* 50, 643-651.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, 5, 349-357.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making,* 5, 411-419.

Polk, T.A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, 102, 533-566.

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

Stenning, K., & Van Lambalgen, M. (2008). *Human reasoning and cognitive science: Logical foundations for the psychology of reasoning.* Cambridge, MA: MIT Press.

Walsh, C. R., & Johnson-Laird, P. N. (2009). Changing your mind. *Memory & Cognition*, 37 (5), 624-631.

Wason, P.C. (1960). On the failure to eliminate hypothesis in a conceptual task. *Quarterly Journal of Experiment Psychology,* 12, 129-140.

Wason, P.C. (1964). The effect of self-contradiction on fallacious reasoning. *Quarterly Journal of Experimental Psychology*, 16, 30-34.