Cognitive Basis of Conditional Reasoning: Insight Through Eye Movements

by

Oufei Dong
Class of 2013

A thesis submitted to the faculty of Wesleyan University in partial fulfillment of the requirements for the Degree of Bachelor of Arts with Departmental Honors in Psychology

Middletown, Connecticut April, 2013
Abstract

The current study used eye movements to investigate the cognitive processes underlying deductive reasoning. Connections were made between eye-tracking research on problem-solving and existing accounts of deductive reasoning. Eighteen college students solved conditional reasoning problems using three valid inference rules from propositional logic: *modus ponens*, *modus tollens*, and hypothetical syllogism. Their eye movements during the task were analyzed for insights toward the nature of the deductive reasoning process: specifically, whether it is driven by mental models or syntax-based, domain-independent inference rules. Findings suggest that the reversal of propositional terms between *modus ponens* and *modus tollens* problems had a significant effect on the accuracy of participants’ responses. Despite this, general eye movement patterns during those two problem types remained similar. Although these findings do not entirely dismiss the role played by syntax-based rules, they offer potential evidence toward the mental model account of reasoning.
Cognitive Basis of Conditional Reasoning: Insight Through Eye Movements

Since the mid- to late 20th century, eye-tracking has become an increasingly popular method for studying cognitive processes, particularly during visual search, scene perception, decision-making, reading, and various types of problem-solving. This is likely because eye movements can inform us on a detailed level about the online processes (i.e., processes occurring during the course of a given task) driving these activities. This type of information typically cannot be gained through traditional measures such as reaction time and error rate. Moreover, eye movements are good measures of overt attention and precise indices of mental processing. Finally, eye-tracking data can be combined with those gathered using traditional measures for a more comprehensive understanding (see Rayner, 2009, for a review of research on eye movements during reading, scene perception, and visual search).

The present study investigates the largely unexplored intersection between eye movements and deductive reasoning, specifically conditional reasoning (i.e., reasoning based on “if…then…” statements). The sections below provide an overview of how eye movements have been used to study various types of reasoning tasks, including diagram-based problem-solving (Grant & Spivey, 2003; Hegarty & Just, 1993; Just & Carpenter, 1985), arithmetic problems and number line estimation tasks (Green, Lemaire, & Dufau, 2007; Hegarty, Mayer, & Monk, 1995; Schneider et al., 2008; Sullivan, Juhasz, Slattery, & Barth, 2011; Verschaffel, De Corte, & Pauwels, 1992;), and deductive reasoning (Aaron & Spivey, 1999; Espino, Santamaría, Meseguer, & Carreiras, 2005; Halberda, 2006; Johnson-Laird & Bara, 1984). Then, the argument forms *modus ponens*, *modus tollens*, and hypothetical
syllogism are briefly described. Finally, existing theories of deductive reasoning and their relationships to biases in human logical inference are discussed (Ball, Phillips, Wade, & Quayle, 2006; Johnson-Laird & Wason, 1970; Oberauer, Hörnig, Weidenfeld, & Wilhelm, 2005; Wason & Shapiro, 1971). Together, these studies demonstrate that eye movements are a valid medium through which insights about reasoning processes may be obtained.

**Eye Movements and Problem Solving**

Existing literature documents eye movements during various problem-solving tasks. Several studies on the subject have focused on diagram-based problems. For example, Just and Carpenter (1985) made use of eye movements to study spatial reasoning ability, particularly the mental rotation of visual stimuli. The study used the Cube Comparison test, in which participants were asked to determine whether two drawings of differently oriented cubes could depict the same object. The researchers also looked at performance in the Shepard-Metzler task, which required participants to compare two images of 3-dimensional, asymmetrical objects. Differences between individuals with high and low spatial reasoning skills were examined. Also investigated were different mental rotation strategies. These entail the encoding of objects with respect to various cognitive coordinate systems, in which mental representations of said objects are rotated. In both tasks, eye movements helped to indicate that subjects with low spatial reasoning ability took longer to mentally rotate the given object, and that they were less efficient than their counterparts with high spatial reasoning ability at keeping track of multiple aspects of the object. In addition,
in the Cube Comparison task, subjects with high spatial reasoning skills were found to be more flexible with their mental rotation strategies.

Hegarty and Just (1993) examined eye movements of people as they constructed mental models of simple machines based on written descriptions and diagrams. A preliminary experiment showed that people tend to understand mechanical systems better through a combination of text and diagrams than through either alone. A second experiment investigated the process of integrating information from these two media. Subjects’ eye movement patterns suggested that they had read the descriptions incrementally, rereading relevant sections prior to the construction of mental models. In addition, they had engaged the diagrams first on a local level, as individual components, and later on a global level, as connected systems of many parts. It was suggested that subjects had used the diagrams as external representations of their mental models. In doing so, they had freed up cognitive resources to mentally integrate the individual parts into a whole.

More recently, eye movements have been used to examine and even actively influence the process of solving insight problems, the answers to which cannot be merely deduced using logic. Grant and Spivey (2003) demonstrated that guiding eye movements can increase the likelihood of success in an insight- and diagram-based problem known as Duncker's radiation problem. Participants were given a diagram representing a tumor surrounded by healthy tissue. The diagram contained four regions of interest: an oval “tumor” at the center, a larger surrounding oval representing skin, a white region of “healthy tissue” between the tumor and the skin, and the region outside of the skin. Participants were then asked to think of a way,
using hypothetical lasers that destroy tissue at a sufficient intensity, to cure the tumor without damaging the healthy tissue. The solution is to fire multiple low-intensity lasers from outside the healthy region so that they converge at the tumor at a sufficient combined intensity. Perhaps surprisingly, it was found that the successful problem solvers spent a significantly higher percentage of time looking at the “skin” than the unsuccessful ones did. Moreover, in a follow-up experiment, the researchers were able to significantly increase the success rate by subtly animating the skin region. It was suggested that, in the successful problem solvers, the numerous fixations on the skin region were a side effect of frequent eye movements from the tumor to the outside region and back, and that these movements simulated the process of directing of multiple lasers at the tumor. In conclusion, the study showed that eye movements may actively direct cognition rather than solely being directed by cognition.

**Eye Movements and Mathematical Reasoning**

While it may be intuitive to use eye movements to study diagram-based problem solving, which clearly involves spatial cognition, eye movements have also been informative of cognitive processes involved in seemingly abstract problem-solving tasks requiring mathematical skill, such as number line estimations and arithmetic problems. An eye-tracking study by Sullivan et al. (2011) examined the numerical-spatial translation skills of adults using number line estimation tasks, as well as the effect of the initial number’s size on their performance. Participants were asked to estimate, using mouse clicks, the positions of pseudo-randomly generated numbers on a number line. It was found that adults’ numerical-spatial translation
skills were rapid and accurate, and that the size of the target number influenced the location of the first fixation. In addition, because the size of the initial number affected both looking behavior and estimation performance, it was suggested that participants had used on-line calibration. That is, they had calibrated their estimates in real-time as they were completing the task as opposed to relying on memorized number-to-location correspondences. Finally, patterns of estimation error suggested bias caused by a proportional-reasoning strategy, wherein participants made biased judgments of numerical magnitude by considering the part (the number whose location is to be estimated) in relation to the whole (the number line). This was reflected by participants’ tendencies to fixate on the midpoint as a central reference point.

Schneider et al. (2008) conducted a similar study, in which the number-line estimation task was presented to children. The study also used both eye-tracking and behavioral methods. Specifically, participants indicated their answers with fixations as well as mouse clicks. It was concluded that elementary school children’s number sense, as indicated by the number-line estimation task, increases with grade level; that eye fixations on the number line are closely related to mouse clicks on the same; that in Grade 2, fixations are related to the ability to solve addition problems; and that children also use the midpoint of the line as a reference.

Green et al. (2007) conducted a study using eye movements as reflections of complex addition strategies in younger and older adults. Two strategies for adding two three-digit numbers were tested: adding the hundred digits first, then the decades, and then the units; and starting with the unit digits, then the decades, and then the
hundreds. Each strategy was associated with a certain eye movement pattern. In one experimental condition, participants were given a choice between the two strategies. In a second condition, participants were told to use one of the two strategies. The study revealed that each strategy was suited to a different type of problem. For example, when no carrying was required during the process of addition, it was shown to be faster to start with the hundreds digits. It was also found that age did not significantly affect addition speed, but that young participants chose between two strategies more adaptively than older ones. Importantly, participants’ reported strategies of choice were reflected in their eye movements, thus validating eye-tracking as a method for studying arithmetic problem-solving.

Verschaffel et al. (1992) conducted a series of eye movement experiments examining the comprehension errors made by students in arithmetic word problems involving relational terms such as “more” or “less.” They concluded that, with a sufficiently high demand on working memory, error rates were higher for word problems that used words associated with addition (such as “more” and “altogether”) but actually requires subtraction, or vice versa, compared to problems that used relational terms consistent with the operation. This was suggested to be a result of the need to mentally rearrange the relational sentences to make them consistent with the associated arithmetic operations. In a related study, Hegarty et al. (1995) used differences in eye movement patterns to compare the strategies of successful and unsuccessful solvers of arithmetic word problems containing relational terms. It was proposed that successful problem solvers tended to create mental representations of
the problems, whereas unsuccessful ones tended to focus on specific relational keywords such as “less” or “more.”

To summarize, eye movements have been widely used to study the cognitive processes and strategies involved in problem-solving, both mathematical and otherwise. Because of the crucial role that deductive reasoning plays in mathematical reasoning, it may be that many of the general findings in the aforementioned studies also apply to the former. In particular, certain eye movement patterns may be reflective of specific deductive reasoning strategies. The current study explores the possibility of this connection.

Eye Movements and Deductive Reasoning

While an abundance of published eye movement studies exists in the general realm of problem solving, very few of them have focused on deductive reasoning. Eye-tracking has been used to study a particular rule of deductive logic, namely disjunctive syllogism (process of elimination), as a word learning strategy in both children and adults (Halberda, 2006). However, the focus of the study was on word learning and not the process of deduction itself. Given an image of a familiar object and a novel object, participants were asked to apply a label to one of the objects by fixating on it. If the label was novel, participants tended to look at the known object before fixating on the novel one. Halberda proposed that this gaze pattern was a result of their rejecting the known object as the referent of the label, and that this process enables word-learners to map novel labels to novel objects.

More relevantly, Espino et al. (2005) studied eye movements during the process of drawing conclusions from categorical syllogism. They examined the
difficulty of syllogistic reasoning problems and the arrangement of the terms in the premises on early and late processing of the syllogisms. A categorical syllogism consists of a major premise, a minor premise, and a conclusion, all of which can be one of four forms:

- All A are B
- Some A are B
- No A are B
- Some A are not B

A and B each represent a term, such that a total of four terms appear in the two premises of each argument. The difficulty of the problem was found to influence the total reading times of both premises. Meanwhile, the arrangement of the four terms was found to have an effect on the total reading time of only the second premise. This effect was thought to be related to the difficulty with which people can integrate information from the second premise with that from the first. Both variables had significant effects on the error rates of conclusions. These findings were interpreted as evidence for the mental model theory of reasoning (see below). The theory postulates that the premises in syllogisms are represented as mental models, and that the arrangement of the terms influences the construction of said models and therefore the integration of the premises. This is known as “figural bias” or the “figural effect” (Johnson-Laird & Bara, 1984).

The current study is largely inspired by Aaron and Spivey’s (1999) abstract on the application of human logical inference to automated theorem proving. Students who had taken courses involving calculational logic each wrote several calculational
proofs while their eye movements were recorded. Each proof problem consisted of a statement to be proved and a list of the axioms of calculational logic, or premises, which remained constant for all problems. One of the problems was presented in two equivalent variations, one of which had two additional propositional letters each connected by a disjunction (logical “or”) operator. This difference was found to impact the fixation patterns on the premise list, an effect similar though not identical to figural bias. Interestingly, this attentional difference did not significantly alter the written proof. In addition, participants devoted a significant number of fixations to unused premises, a phenomenon not fully explained in the abstract.

In summary, previous eye-tracking research on microcognition during deductive reasoning leaves much unexplored. For example, do aspects of strategies used in arithmetic word problems carry over to logic problems? Do biases found in arithmetic problem-solving or syllogistic reasoning also show up in the comprehension of relational arguments? The current study may provide a starting point for answering at least some of these questions.

**Logical Argument Forms: Modus Ponens, Modus Tollens, and Hypothetical Syllogism**

Deductive logical reasoning is the process of arriving at a logically certain conclusion based on one or more general statements. This type of reasoning is central to propositional logic, in which one proposition may be inferred from two or more others. The inferred proposition is referred to as the conclusion, and the propositions that lead to it are known as the premises. Propositional logic deals almost exclusively with syntax. Propositions are represented by propositional letters and connected by
logical operator symbols, and the semantic meaning of each proposition is limited to one of two truth values (“true” and “false”) assigned to it (see Dipert, 1981, for a review of propositional logic; see Johnson-Laird, 1999, for a review of deductive reasoning).

An argument consists of a set of propositions. In a valid argument, the last proposition, the conclusion, is logically entailed by the premises, and the conclusion is true as long as the premises are true. For example, a simple argument may consist of a conditional premise, a secondary premise (hereupon referred to as the hypothesis for the sake of clarity), and a conclusion. The premise is a conditional statement in the form of “if $P$, then $Q$,” where $P$ and $Q$ may represent any two propositions. $P$ is called the antecedent, and $Q$ the consequent. The hypothesis asserts that $P$ is true. Based on the conditional premise and the hypothesis, it can be concluded that $Q$ is necessarily true. This argument form is referred to as *modus ponens* (see Johnson-Laird, 1999).

A related argument form is *modus tollens*, also known as denying the consequent. We are given the same conditional premise form, “if $P$, then $Q$. $Q$ is false, leading to the conclusion that $P$ is necessarily false (see Johnson-Laird, 1999).

The final argument form employed in this study is the hypothetical syllogism. Essentially a two-steps *modus ponens*, classical hypothetical syllogism is an argument form in which the two premises and the conclusion are all conditional statements:

If $P$, then $Q$.

If $Q$, then $R$. 
Therefore, if $P$, then $R$.

Thus, the conclusion itself is nearly identical to a single conditional premise, usable in a *modus ponens* argument wherein if a hypothesis asserts $P$ to be true, $R$ must also be true (see Dipert, 1981). The set of inference rules used in propositional reasoning is by no means limited to the three mentioned here, but they are the only ones employed in the current study. *Modus ponens* and *modus tollens* were chosen because they have been traditionally used in studies of conditional reasoning (see Johnson-Laird, 1999). Hypothetical syllogism, as described above, is comparable to a double *modus ponens*. As such, it is a good way to test the effect of an increased cognitive load on reasoning processes.

Traditionally, studies on deductive reasoning have involved making inferences in the forward direction. That is, participants are usually presented with one or more premises and asked to determine if a given conclusion drawn from them is valid (Johnson-Laird & Wason, 1970). Ball et al. (2006) used this paradigm in their study on the effect of prior belief on reasoning, described in the next section. However, exceptions can be found in the field of automated computer simulation of reasoning, such as Rips’s (1983) ANDS model of mental operations underlying propositional reasoning, which incorporates backward deduction rules. For example, according to its backwards *modus ponens* rule, if the primary premise “if $P$, then $Q$” is known to be true, and the goal is to deduce the conclusion $Q$, then the secondary hypothesis $P$ may be be inferred from other existing axioms. However, this still differs from the paradigm used in our study, described below, in which subjects are required to infer the primary premise given the hypothesis and the conclusion.
Theories of Deductive Reasoning and Limitations of Human Logical Inference

In his review of accounts of deductive reasoning, Johnson-Laird (1999) outlines three major competing theories on the subject. The first, dating back to antiquity and pervasive throughout history, asserts that deductive reasoning relies on syntactic, domain-independent, formal rules of logic. Rips’s (1983) ANDS model, for instance, presupposes this theory. The second school of thought maintains that a basis of factual knowledge underlies deductive reasoning. According to this theory, reasoning is removed from logic and instead depends on memories of previous inferences, which may then lead to new inferences. The problem with this theory is that it does not readily explain how people are able to reason about unknown or abstract things. The final school of thought views deduction as based on semantics. According to this theory, people construct mental models of situations described by statements, and reasoning is based on the manipulation of these models, though the exact nature of such models is debated (see Johnson-Laird, 1999; Johnson-Laird & Bara, 1984; Chater & Oaksford, 2001).

Of course, humans are anything but infallible logicians. *Modus ponens*, *modus tollens*, and hypothetical syllogism are all valid argument forms. However, it has been shown that depending on the semantic content of the propositions, the validity of these types of arguments may or may not be readily apparent to people. In syllogistic reasoning, people are more likely to accept conclusions they find believable than ones they find unbelievable, regardless of the actual validity of the argument (Ball et al., 2006). Moreover, in conditional reasoning, people tend to be prone to a “matching
bias,” in which they ignore the negatives preceding propositional terms. This effect makes *modus tollens* more difficult than *modus ponens* (see Johnson-Laird, 1999).

Relatively, prior knowledge has been found to have an effect on performance in the Wason selection task and its variations. In the original task, subjects are presented with four cards showing (for instance) a vowel, a consonant, an odd number, and an even number. The experimenter provides a statement, such as “every card which has a vowel on one side has an even number on the other.” Subjects are then asked to select all the cards, and only those cards, that would need to be flipped in order to determine the truth of the statement. A response is considered erroneous if it either identifies a card that does not need to be flipped or fails to identify one that does. The majority of responses were erroneous, usually due to affirmation of the consequent. However, when given a version of the problem using concrete terms describing realistic situations, the number of correct responses dramatically increased (see Johnson-Laird, 1999; Wason & Shapiro, 1971).

In a similar vein to the study on figural bias by Espino et al. (2005), it has been suggested that a semantic directionality is inherent to most statements used in deductive reasoning, since they tend to express relationships between objects. For example, the statement “all X are Y” has the opposite directionality as “some X are Y,” since X is the larger set in the first statement and the smaller set in the second. Likewise, the statements “if it is raining, then the street is wet” and “if it the street is wet, then it is raining” have opposite directions of implication. This directionality, along with a recency effect of working memory, is thought to modulate the effect of premise figure on the time required to read the premise. This has been observed in
reading times during both the premise integration stage and the conclusion derivation stage in syllogistic arguments (Oberauer et al., 2005). In their study, participants were presented with arguments containing multiple premises one line at a time. Following each set of premises, they were presented with a conclusion and asked to judge its validity using key presses. The reading times of each premise and the reaction times to the conclusion were measured, though no eye-tracking was involved. It was found that certain forms of premises were much more efficiently processed than others. As an example, premise pairs in the form of “all A are B; all B are C” took less time to read than those in the form of “all B are A; some C are B.” These differences in order also affected their reaction times to conclusions. A follow-up experiment revealed similar effects when participants, instead of judging the validity of a given conclusion, were shown a list of possible conclusions and asked to select the correct one. These findings suggested that subjects preferred certain combinations of semantic directionality over others. Once again, the researchers interpreted their results as evidence for the mental model account of reasoning because it predicts the figural effect, whereas the syntactic account predicts no such effect.

**Current Study**

The purpose of the current study is to use eye movements to investigate microcognition during deductive reasoning. An attempt is made to connect existing eye-tracking studies on problem-solving with literature related to deductive logical reasoning. Although error rates were examined, they were not the main focus of the analyses. As mentioned, the advantage of using eye movements is that they allow us
to glimpse the cognitive processes underlying problem solving in real time, which
traditional measures such as error rate and reaction time cannot address.

The reasoning tasks used in this study applied certain rules of propositional
logic, upon which the system of calculational logic used in Aaron and Spivey’s (1999)
study is based. However, the current study differs drastically from theirs in its
experimental design. Firstly, instead of limiting our participant pool students of
mathematical logic, the present study used typical college students not necessarily
trained in logical reasoning for a more random sample. Secondly, our reasoning
problems consisted of relational statements in the form of sentences (albeit ones
containing made-up words) rather than symbolic propositions, thus requiring no
formal training on the participants’ part. Furthermore, according to the three
aforementioned argument forms, the premises in our statements were simply
conditional statements, and each premise appeared only once throughout the
experiment. Finally, participants reasoned mentally rather than writing out formal
proofs, as the motions of writing introduce head movements that would reduce eye-
tracking accuracy.

As mentioned, rather than using the traditional process of deducing
conclusions from premises and hypotheses, the type of task used in the present study
required the backwards deduction of conditional premises using the hypothesis and
the conclusion. This structure was designed to mimic that of calculational logic
problems such as those used by Aaron and Spivey (1999). By reasoning backwards,
participants essentially constructed a mental proof for every hypothesis-conclusion
A number of complications must be kept in mind when using eye-tracking to study logical reasoning. In human subjects, the process of constructing proofs using a given set of premises almost inevitably incorporates elements of visual search. Therefore, the locations of propositions inevitably exert an influence on people’s gaze. To account for this, we varied the location of each premise type (described in detail below) from trial to trial. Secondly, a general difficulty of studying human reasoning is that certain effects may only manifest under a sufficient cognitive load, and it is difficult to gauge what exactly constitutes “sufficient.” This is also difficult to answer without prior testing and may be addressed in future experiments. Finally, the words themselves may have biased fixations in unpredictable ways. For example, it has been shown that people devote a longer fixation time to high-frequency words than low-frequency ones. Fixation time has also been shown to be longer for words with two equally likely meanings than those with one very likely meaning (Rayner & Duffy, 1986). By incorporating non-words in the arguments and thereby creating semantically incoherent sentences, we hoped to reduce effect of belief on reasoning as well as minimize any other effects that real words may have on fixations.

The first question this study seeks to address is how eye movements may provide evidence (or lack thereof) for existing theories of deductive reasoning. Since our problems contain non-words, the knowledge-based theory is beyond the scope of this study. As described above, the mental model account of reasoning predicts an effect of the arrangement of propositional terms on eye movement patterns, whereas
the syntactic account does not. In *modus ponens* arguments, terms in the conclusion appear in same order as those in the premise (first $P$, then $Q$). In *modus tollens* arguments, the order is reversed. Does this reversal result in different fixation patterns on the premises, the hypothesis, or the conclusion? Since the problems in the first two sections only required one inference step, a high error rate was not expected for them. However, it was thought that participants would spend time looking at irrelevant premises even if they ultimately arrived at the correct answer; and that certain irrelevant premises would be more distracting than others. Particularly, the more similar a premise is to the correct answer, the more distracting it should be. This should be reflected in the gaze durations on those premises as well as the number of double-checks made to them.

The current study also investigates a question raised by Aaron and Spivey (1999). That is, how much attention do people devote to necessary premises--or, in our case, correct ones--compared to the others (i.e., the distractors) during the deductive reasoning process? Fixation patterns on premises likely vary among argument types and differ between successful and unsuccessful reasoners, although the last part may be most relevant to the hypothetical syllogism problems. Since they essentially require a two-step *modus ponens* deduction, they should pose a higher demand on working memory than the other two argument forms. Therefore, it was hypothesized that response accuracy would be lowest for those problems, in accordance with their relatively high difficulty. Accuracy rate was hypothesized to be lower for *modus tollens* problems than *modus ponens* ones because of the reversal of propositional terms; however, since *modus tollens* problems required the participant
to select only one premise, they were hypothesized to be easier to solve than hypothetical syllogisms.

A final aim of this study is to determine the extent to which eye movement patterns related to decision-making and informational search carry over to the realm of logical reasoning tasks. In particular, we examine differences in the tendency to gaze at blank space in those who self-report to be decisive versus their indecisive counterparts. Patalano, Juhasz, & Dicke’s (2010) eye-tracking study on decision-making revealed that compared to decisive individuals, indecisive individuals spent significantly more time looking away from parts of the stimulus containing decision information in a college course selection task. It would be interesting to see how eye movement patterns in decision-making tasks replicate themselves in deductive reasoning tasks, which incorporate elements of decision-making but do not rely on personal preference.

Method

Participants

Eighteen Wesleyan University students participated in the study. Two were excluded because over 20% of their trials were removed from analysis based on criteria discussed below in the “Design and Analysis” section. All had normal or corrected vision. Each participant was tested individually in a session that lasted approximately 30 to 45 minutes. Participants were recruited by e-mail from a list of former participants; they were each paid $10 in compensation.
Apparatus

Participants’ eye movements were recorded using an EyeLink 1000 (SR Research, LTD) eye tracker with binocular viewing. Eye position was sampled from the right eye every 1 ms. The logic problems were displayed on a 20-inch ViewSonic CRT monitor positioned 83 cm from the participants’ eyes.

Stimuli

The task required participants to complete partial deductive arguments. Fifteen deductive reasoning problems were presented. The first five followed the argument form of *modus ponens*, the second five that of *modus tollens*, and the final five that of hypothetical syllogism (see Appendix for a complete list of problems).

For each problem in the first two sections, an incomplete deductive argument consisting of only the hypothesis (either \(P\) for *modus ponens* or \(\neg P\) for *modus tollens*) and the conclusion (either \(Q\) or \(\neg Q\), respectively) was displayed on the left side of the screen. On the right side was a list of six conditional statements, only one of which could serve as the primary premise (if \(P\), then \(Q\)) of the argument. That is, one statement, when considered together with the hypothesis, logically entailed the conclusion displayed. The third section, which involved a two-step reasoning, followed a similar form. The hypothesis \(P\) and the conclusion \(R\) were displayed on the left. Eight conditional statements were shown on the right. Two of those statements could serve as valid premises (if \(P\) then \(Q\), if \(Q\) then \(R\)) of the argument. The remaining premises consisted of different types of distractors. For each of the *modus ponens* and *modus tollens* problems, the six premises had the same forms.
For example, in each modus ponens problem, participants were presented with a pair of hypothesis and conclusion in the form “A. Therefore, B.” The correct premise had the form “if A, then B” (AB). The five distractors had the forms “if A, then C” (AC), “if B, then A” (BA), “if B, then C” (BC), “if D, then B” (DB), and “if C, then D” (CD), though not necessarily in that order. Note that C and D are entirely different terms from A and B, such that there were varying degrees of similarity to the correct answer among the distractor types. In each modus tollens problem, participants were presented with the hypothesis and conclusion in the form “Not B. Therefore, not A.” The premise types had the same forms as those in the modus ponens problems. Finally, in the hypothetical syllogisms, participants were given the hypothesis and conclusion in the form “A. Therefore, C.” The two correct premises had the forms “if A, then B” (AB) and “if B, then C” (BC). The six distractors had the forms “if A, then E” (AE), “if B, then A” (BA), “if C, then A” (CA), “if C, then B” (CB), “if D, then C” (DC), and “if D, then E” (DE). As with the previous two sections, D and E are entirely different terms from A, B, and C.

Rather than using letters and logic operator symbols, arguments were presented in the form of English sentences. Letters were replaced by a combination of words and non-words to avoid potential bias while remaining syntactically unambiguous. Logic operators were replaced by their English equivalents (“if...then...” and “not”).

In addition to the reasoning tasks, Frost and Shows’s (1993) Indecisiveness Scale was administered. The questionnaire contained 15 Likert items rated on a scale
from 1 (strongly disagree) to 7 (strongly agree). It included items such as “I try to put off making decisions” and “I often worry about making the wrong choice.”

Procedure

At the start of each experiment, the experimenter briefly explained the notion of deductive reasoning, providing an example of each of the three argument forms used in this study. The participant was then brought to the testing room. Prior to the start of each section, an example of the specific argument form presented in the section was shown to the participant. These examples consisted entirely of real words, with the correct answer(s) labeled. For example, prior to the *modus ponens* section, participants were presented with example containing the following hypothesis and conclusion, respectively:

Jamie is ill.

Therefore, Jamie stays home.

The statements were accompanied by the following list of possible premises:

1. If Jamie stays home, then Jamie reads a book.
2. If Jamie stays home, then Jamie is ill.
3. If Jamie is happy, then Jamie stays home,
4. If Jamie reads a book, then Jamie is happy.
5. If Jamie is ill, then Jamie reads a book.
6. If Jamie is ill, then Jamie stays home. (correct premise)

A full-screen 9-point calibration of the eye tracker was then performed. This procedure was conducted before every section, before every question in the hypothetical syllogism section, and whenever the calibration was deemed inaccurate.
After the calibration, the five problems in the section were presented to the participant one at a time. For each question in the first two sections, the participant was told to select what he or she thought was the valid premise. Answers were selected using a mouse click anywhere on the premise. In the third section, the participant was asked to find the two valid premises in no particular order. Upon arriving at a solution, he or she signaled the completion of the problem using a mouse click before verbally stating his or her two choices. After completing all 15 problems, the participant completed the Indecisiveness Scale.

**Design and Analysis**

All fixations that lasted under 80 ms were removed. A trial was excluded from the final analysis if it met one of four criteria: the maximum calibration error exceeded 1° of visual angle, the average calibration error exceeded 0.60° of visual angle, 20% or more of that trial’s data was lost (from a combination blinks, track losses, and fixations under 80ms), or the participant did not understand the instructions during that trial (in which case, they were clarified after the trial so that the remaining trials could be properly completed). In addition, if three (20%) or more of a given participant’s trials were removed based on these criteria, then the participant’s data were excluded from the data analysis entirely. Two participants were removed in this manner.

The 16 participants included in the analysis contributed 240 trials, nine (3.75%) of which were removed. In addition, 7.38% of the remaining fixations were removed.
Results

Since a large number of comparisons were used in the present study, a Bonferroni correction was applied to all contrasts. The corrected alpha level is reported prior to each set of contrasts.

Response Accuracy

Participants’ responses were assessed based on accuracy, with each correct answer scoring 1 and each incorrect one scoring 0. For each hypothetical syllogism, each of the two premises contributed its own score. The overall average scores for each section are reported in Table 1. A repeated-measures ANOVA showed a significant main effect of problem type on accuracy rate $F(1.57, 23.50) = 9.62, p = .002$. A paired samples $t$-test was used to compare the accuracy rates. Since three contrasts were used for the present analysis, an alpha value of .017 was used. Accuracy rate was found to be significantly higher in section 1 than section 2, $t(15) = 3.73, p = .002$. It also revealed that accuracy rate was marginally lower in section 2 than section 3, $t(15) = -2.60, p = .020$. No significant difference was found between sections 1 and 3 ($p > .017$).

Eye Movement Patterns

Overall dwell times for each trial were recorded. The problem space was initially divided into two interest areas: the left side, which contained the hypothesis and conclusion; and the right side, which contained the premises. Total dwell time, first run dwell time, and run count were examined for each interest area without accounting for the order in which the problems were presented. For further analysis,

---

1 Analyses taking into the account the orders of problems were conducted, but no significant effects or interactions were found with the problem order variable.
the problem space was re-divided into the hypothesis by itself, the conclusion by itself, and each individual premise. The same variables were then examined for each of those interest areas. For both sets of interest areas, repeated measures ANOVAs were conducted to investigate the effects of interest area and problem type on dwell time, first run dwell time, and run count. As noted above, a Bonferroni correction was applied to all contrasts. Also, because Mauchly’s sphericity test indicated unequal variance between conditions, a Huynh-Feldt correction was applied to the degrees of freedom in each analysis.

**Overall trial dwell time by problem type.** The mean overall trial dwell times for sections 1 (*modus ponens*), 2 (*modus tollens*), and 3 (hypothetical syllogism) are reported in Table 2. A repeated measures ANOVA revealed a significant effect of problem type on overall trial dwell time, $F(1.17, 18.14) = 16.73, p < .001$. Since three contrasts were used for the present analysis, an alpha value of .017 was used. A paired samples $t$-test showed that average dwell time was significantly shorter in the section 1 than section 3, $t(15) = -4.55, p < .001$. A significant difference was also found between sections 2 and 3, $t(15) = -3.91, p = .001$, with average dwell time being shorter in section 2. No significant difference was found between the dwell times in sections 1 and 2 ($p > .017$).

**Analyses of interest areas: hypothesis and conclusion versus premises.** Since three contrasts were used for each of the analyses in this section, a Bonferroni-corrected alpha value of .017 was used. Average dwell times—i.e., the total amount of time spent looking at the interest area—on the left (hypothesis-conclusion) and right (premises) sides for each section are shown in Table 3. A significant main effect
of problem type on total interest area dwell time was found, $F(1.24, 18.60) = 14.91, p < .001$. A significant main effect was also found for interest area, $F(1.00, 15.00) = 18.21, p < .001$. There was also a significant interaction between problem type and interest area, $F(1.03, 15.48) = 30.83, p < .001$. A paired samples $t$-test showed a significant difference between average dwell times on the two interest areas in section 1, $t(15) = -2.95, p = .010$, with dwell time being longer on the right side. A significant difference was also found in section 3, $t(15) = -5.07, p < .001$, also with dwell time being longer on the right side. No significant difference was found in section 2 ($p > .017$).

The “first run dwell time” for an interest area was defined as the amount of time from when the gaze first entered that interest area to when it first left that area. Average first run dwell times on the left and right sides for each section are shown in Table 3. A significant main effect of interest area was found on first run dwell times, $F(1.00, 15.00) = 18.83, p < .001$. No significant effect was found for problem type ($p > .100$). However, a significant interaction was found between problem type and interest area, $F(1.79, 26.88) = 5.34, p = .013$. A paired samples $t$-test showed significant differences between first run dwell times in the two interest areas in section 1, $t(15) = 5.64, p < .001$; as well as in section 2, $t(15) = 5.86, p < .001$. In both cases, first run dwell time was longer on the left side. No significant difference was found in section 3 ($p > .017$).

The “run count” for an interest area was defined as the total number of times the gaze entered that area from another region. Average run counts for the left and right sides are shown in Table 3. There was a significant main effect of problem type
on run count, $F(1.43, 21.54) = 15.33, p < .001$. A significant main effect of interest area was also found, $F(1.00, 5.64) = 8.64, p = .010$. A significant interaction was found between problem type and interest area, $F(1.17, 17.54) = 11.43, p = .002$. A paired samples $t$-test showed that in section 3, the right side had a significantly higher run count than the left, $t(15) = -3.49, p = .003$. No significant differences were found in sections 1 and 2 ($p > .017$).

**Analyses of interest areas: hypothesis and conclusion.** As above, because three contrasts were used for each analysis in this section, an alpha value of .017 was used. Average dwell times on the hypothesis alone and conclusion alone for each section are reported in Table 4. No significant effect of problem type on dwell time was found ($p > .100$). A marginally significant main effect of interest area was found, $F(1.69, 3328) = 3.34, p = .088$. A significant interaction between problem type and interest area was found, $F(1.23, 18.52) = 7.64, p < .009$. A paired-samples $t$-test showed no significant differences between dwell times on the conclusion and the hypothesis in any section ($p > .017$). However, a marginally significant difference was found in section 3 only, $t(15) = -2.77, p = .022$, such that dwell time was higher in the conclusion region.

Average first run dwell times on the hypothesis alone and conclusion alone are shown in Table 4. A significant main effect of problem type on first run dwell time was found, $F(1.49, 22.31) = 8.90, p = .003$. No significant effect of interest area was found ($p > .100$). A significant interaction was found between problem type and interest area, $F(2.00, 30.00) = 3.73, p = .036$. A paired-samples $t$-test showed no significant differences between first run dwell times on the conclusion and the
hypothesis in any section \((p > .017)\). However, there was a marginally significant difference in section 1 only, \(t(15) = 2.39, p = .030\), such that first run dwell time was higher in the hypothesis region.

Average run counts on the hypothesis alone and conclusion alone are shown in Table 4. A significant main effect of problem type on run count was found, \(F(1.59, 23.77) = 7.87, p = .004\). No significant effect of interest area was found \((p > .100)\). A significant interaction was found between problem type and interest area, \(F(1.45, 21.81) = 11.81, p < .001\). A paired-samples \(t\)-test showed that in section 3, the conclusion region had a significantly higher run count than the hypothesis region, \(t(15) = -2.774, p = .014\). No significant difference was found for the other two sections \((p > .017)\).

**Analyses of interest areas: premises.** For each problem type, premises were analyzed according to the order in which they appear. In the first two sections, premises were also analyzed according to their similarity to the correct answers. As per the Method section, the labels AB, AC, BA, BC, DB, and CD will be used to refer to the various premise types in the analyses.

Table 5 shows the average dwell times on premises by order and problem type. A significant effect of premise order on dwell time was found only in section 1, \(F(2.64, 39.59) = 15.85, p < .001\). A marginally significant effect of the same was found in section 2, \(F(3.80, 56.98) = 2.43, p = .061\).

Average total dwell times on the premises in sections 1 and 2 by their similarity to the correct answers are shown in Table 6. A significant main effect of similarity to correctness on dwell time was found, \(F(2.25, 33.76) = 38.78, p < .001\).
No significant effect of problem type was found. There was a significant interaction between problem type and similarity to correctness, $F(2.57, 38.61) = 3.38, p = .034$. A paired-samples $t$-test was conducted to compare dwell time on the correct premise to that on each distractor. Since 10 contrasts were used in total, an alpha value of .005 was used. In section 1, total dwell time was significantly longer on the correct premise than on every distractor type: AC ($t(15) = 6.19, p < .001$), BA ($t(15) = 6.91, p < .001$), BC ($t(15) = 7.66, p < .001$), DB ($t(15) = 8.60, p < .001$), and CD ($t(15) = 8.55, p < .001$). In section 2, total dwell time was significantly longer on the correct premise than on distractor types AC ($t(15) = 3.62, p = .002$), BC ($t(15) = 4.14, p < .001$), DB ($t(15) = 4.17, p < .001$), and CD ($t(15) = 4.80, p < .001$); no significant difference was found between the correct premise and distractor type BA ($p > .005$).

Table 6 shows the average first run dwell times on the premises in sections 1 and 2 by their similarity to the correct answers. A significant main effect of similarity to correctness on first run dwell time was found, $F(3.92, 58.77) = 17.21, p < .001$. No significant effect of problem type was found ($p > .100$), nor did it significantly interact with similarity to correctness ($p > .100$).

Table 6 shows the average run counts for the premises in sections 1 and 2 by their similarity to the correct answers. A significant main effect of similarity to correctness on run count was found, $F(3.01, 45.17) = 31.49, p < .001$. No significant effect of problem type was found. There was a significant interaction between similarity to correctness and problem type, $F(3.84, 57.92) = 2.74, p = .039$. A paired-samples $t$-test was performed to compare differences in run count between the correct premise and the distractors. As above, an alpha value of .005 was used. In section 1,
run count was significantly higher for the correct premise than each distractor type: AC \((t(15) = 5.39, p < .001)\), BA \((t(15) = 5.88, p < .001)\), BC \((t(15) = 8.23, p < .001)\), DB \((t(15) = 8.30, p < .001)\), and CD \((t(15) = 7.33, p < .001)\). In section 2, run count for significantly higher for the correct premise than distractor types BC \((t(15) = 4.01, p < .001)\) and CD \((t(15) = 4.71, p < .001)\); a marginally significant difference was found for AC \((t(15) = 3.24, p = .005)\) and DB \((t(15) = 3.00, p = .009)\); no significant difference was found for BA \((p > .005)\).

**Decision Making Analyses**

Participants were divided into decisive and indecisive groups based on a median score of 3.5 on the Indecisiveness Scale, with indecisive individuals scoring higher. The decisive group had a mean score of 2.73 \((SD = 0.55)\); the indecisive group had a means score of 4.45 \((SD = 0.83)\). An independent samples \(t\)-test was used to compare the mean trial dwell times, mean trial fixations counts, and mean dwell times on blank space between the two groups. However, no significant difference was found for any of the above variables \((p > .05)\).

**Discussion**

**Evidence for Mental Model Account of Deductive Reasoning**

The mental model theory claims that deductive reasoning is semantics-based, and inferences are made through the manipulation of mental models. The syntactic-based account of reasoning essentially equates the rules of human logical inference to the rules of formal logic, where the emphasis is on sentential connectives (see Johnson-Laird, 1999). Support for the former theory has been found in the biases
reiterated in this section, though the latter theory still underlies many studies and
models of deductive reasoning (Ball et al., 2006; Rips, 1983).

The findings related to eye movement patterns and response accuracy appear
to be aligned with a mental model account of reasoning. Though the evidence is by no
means conclusive, the two sections below outline the basis for this proposal. Two
main points are covered. Firstly, the reversal of propositional terms between the
*modus ponens* and *modus tollens* problem types had a significant impact on both
accuracy rate and fixations on premises. This fits into existing research on figural
effects of caused by the arrangement of propositional terms in syllogistic reasoning
tasks (Espino et al., 2005; Johnson-Laird & Bara, 1984; Oberauer et al., 2005).
Secondly, it appears that the reversal of terms had a somewhat greater negative
impact on the reasoning process than an increased working memory load, as reflected
by the marginally higher accuracy rate in section 3 than section 2. This is consistent
with previous findings on the “matching bias,” which describes this very tendency
(see Johnson-Laird, 1999).

**Fixation patterns.** Predictably, participants spent significantly more time on
the hypothetical syllogism problems (section 3) than both the *modus ponens* (section
1) and *modus tollens* (section 2) problems; meanwhile, no significant time difference
was found between the latter two problem types. This was likely because the
problems in section 3, unlike those in the other two sections, required participants to
both read more text and select more premises. Since the problems in sections 1 and 2
were structured similarly, they required similar amounts of time to solve. These
results suggest that the hypothetical syllogisms did indeed present the highest demand on working memory.

Interesting differences emerged when comparing fixation patterns on the left side (hypothesis and conclusion) versus the right side (premises) of the stimulus. Sections 1 and 3 were similar in terms of total dwell times, whereas sections 1 and 2 were similar in terms of first run fixation times and run counts. In sections 1 and 3, a significantly greater amount of time was devoted overall to the right side than the left. This was unsurprising, since the right side contained the most the text. However, in section 2, no significant difference in dwell time was found, perhaps because participants had gotten accustomed to the structure of the problems after completing the first section, which was nearly identical in the presentation of the premise list. Despite this, the accuracy rate was much lower in section 2 (see below). An alternative explanation is that the negation of the terms in the hypothesis and the conclusion might have resulted in longer processing times on the left side, thus evening out the processing times of the two sides. The mean dwell times indeed show such a difference (Table 3).

In both sections 1 and 2, the left side’s first run dwell time was significantly higher than the right side’s. However, there was no significant difference between the first run dwell times on the two sides in section 3. This makes sense, since in sections 1 and 2, all the content words in the left side appeared in the premises on the right side, and so the redundancy likely resulted in a decreased initial processing time of the premises. The same was not true for section 3. This fact, combined with the greater number of premises, means that participants probably took a longer amount of
time to find a relevant premise. Section 3 had the highest run counts on both sides, with more runs on the right side. This shows that participants made many “double-checks” to the hypothesis and conclusion, and even more to the premise list. This difference was not evident in sections 1 and 2. The double-checking behavior was probably directly related to cognitive load, which is again consistent with the assumption that hypothetical syllogisms posed the highest demand on working memory.

Fixation patterns on the hypothesis alone versus the conclusion alone were also examined. In section 1, participants spent a marginally longer first run on the hypothesis than the conclusion, a difference that disappeared in the next sections. Once again, this may simply indicate that participants became accustomed to the task throughout the course of the experiment. In section 3 only, both a longer (albeit marginally so) total dwell time and more run counts were devoted to the conclusion than the hypothesis. This was probably because after initially processing the conclusion, participants had to double-check it during the second step (“if Q, then R”) of the reasoning.

Since both the mental model and syntactic accounts can explain an increased difficulty with a higher number of inference steps, the implications of the high run counts in section 3 are ambiguous. However, they do perhaps indicate that people used the text on the screen as external representations of their mental models so that they might be able to devote more cognitive resources to making inferences. A similar suggestion, as mentioned earlier, was made by Hegarty and Just (1993) about the role of diagrams in helping people understanding simple machines.
Finally, fixations on correct and irrelevant premises were examined. In section 1, participants mostly paid attention to the correct answer, for which the run count was also the highest. In section 2, on average, around the same amount of time was devoted to the correct answer and premise type BA (in which the order of the content words was reversed). Once again, the same pattern held for run counts. These results suggest that participants found premise type BA most distracting, an effect reflected in the dramatic decrease in response accuracy. No other premises, except perhaps premise type AC in section 2, seemed to detract significantly from participants’ ability to solve the problems. Other than premise type BA in section 2, participants generally paid less attention to a premise the further it was from being correct, as predicted. Thus, the reversal of propositional terms in section 2 seemed to have posed the greatest challenge to participants, many of whom committed the fallacy of denying the antecedent. This suggests that unsuccessful reasoners might have exhibited a matching bias, paying less attention to the logical connectives relating the propositional terms than the terms themselves. This difficulty with processing relationships between concepts was reflected in the accuracy rate, as discussed below.

**Response accuracy.** It was hypothesized that response accuracy would be highest in section 1, lower in section 2, and lowest in section 3. Although participants spent around the same amount of time on section 1 as they did on section 2, the response accuracy was, as hypothesized, significantly lower on the latter. However, contrary to what was hypothesized, the response accuracy was marginally lower in section 2 than section 3. This might be partly because the two responses in each hypothetical syllogism problem were scored separately: if a participant selected one
correct premise and one erroneous one, they received an average score of .5 for that problem. Nonetheless, it is interesting and perhaps surprising that a reversal of terms seemed to have presented more difficulty than an extra step of deduction.

In summary, although general fixation patterns (with the exception of fixations on distractor type BA) were similar in sections 1 and 2, response accuracy was dramatically lower in the latter. Moreover, on average, participants paid nearly as much attention to distractor type BA as they did to the correct premise in section 2. Although it cannot be said for certain, these findings combined gesture toward a possible explanation: Errors in section 2 might have been caused by the construction of mental models of propositional terms without taking into account the differences in the direction of the “if…then…” relation. These results are also consistent with the study by Espino et al. (2005) on categorical syllogisms, in which the arrangement of terms in the premises significantly affected the error rates of conclusions drawn by participants. As with their study, this finding may be interpreted as evidence for the use of mental models rather than a syntax-driven approach, at least in participants who made errors. Interestingly, this particular result is reminiscent of the findings by Verschaffel et al. (1992) on to the possible difficulty of mentally rearranging relational sentences in arithmetic word problems. This suggests that mental models may be used in various types of reasoning tasks. Similarly, Hegarty et al. (1995) found that successful solvers of arithmetic problems tend to create mental representations of problems rather than focusing on keywords. However, the present study shows a potential shortcoming of mental models: Namely, they may not be very useful for representing the relationships between existing mental models.
The syntactic theory predicts that difficulty would increase with the number of inferences (see Johnson-Laird, 1999). However, this was not reliably found in the present study: Although participants made the highest number of double-checks to all interest areas in the hypothetical syllogism problems, accuracy rate on these problems did not significantly differ from accuracy rates in *modus ponens* and *modus tollens*. Recall, however, that for the hypothetical syllogisms, no analyses were performed on the premises based on their similarity to the correct answers. Thus, it may be that these difficulties were exhibited more evidently on a microcognitive level.

These findings combined suggest that conditional reasoning is semantics-based. Interestingly, this appears to be true even when non-words are involved. The aforementioned studies on biases during syllogistic reasoning have used real words (Espino et al., 2005; Johnson-Laird & Bara, 1984). However, if the mental model theory applied to conditional reasoning, the current study’s findings suggest that mental models are created even for nonsensical words, perhaps as placeholders, for the sake of making inferences. This then raises a question about the very nature of mental models: If they are capable of representing meaningless terms, then why do they seem to be less effective at representing the relationship between these terms? One might observe that abstractions are involved in both cases, and yet people seem to be much better at creating mental representations of the former. Moreover, the notion that the mind is less adept at representing relationships between objects is be evidence against the syntactic model, which emphasizes the sentential connectives that describe these relationships.
Limitations and Questions

Due to limitations in the software used to present the problems, participants indicated their responses via mouse clicks in sections 1 and 2 but gave their answers verbally in section 3. This inconsistency in the means of response might have affected eye movements, especially toward the end of each trial. For similar reasons, problems were presented in three groups according to their argument forms in order of *modus ponens*, *modus tollens*, and hypothetical syllogism. In future studies of this kind, it may be better to present the problems in mixed order to minimize the predictability of solutions. As previously mentioned, the order of the premises influenced eye movements. Despite our efforts, this effect was not entirely eliminated, as shown in the analyses of fixation patterns on premises by order of presentation. Another obvious shortcoming of the current study was that, due to the high number of analyses already present, fixation patterns on premises in section 3 were not analyzed based on similarity to the correct answers. However, it would not be difficult to perform such a follow-up analysis in the future. Finally, the current study failed to generate conclusive evidence regarding the relationship between indecisiveness and gazing at blank space. Since the tasks in the current study did not involve an element of personal preference as they did in the original study by Patalano et al. (2010). Follow-up investigations may test whether this relationship only holds in preference-based decision-making tasks.

Although potential evidence was found for the mental model account of deductive reasoning, our findings do not entirely rule out the syntax-based account. A natural next step would be to compare fixations patterns on content words versus
words corresponding logical operators. For example, a higher dwell time on content words may suggest the use of mental models; on the other hand, a higher dwell time on logical operator words may provide evidence of a syntax-driven, domain-independent reasoning process. Another possible follow-up analysis is the comparison of eye movement patterns between successful and unsuccessful problem solvers. Perhaps, as is the case with arithmetic word problems (Hegarty et al., 1995) and diagram-based insight problems (Grant & Spivey, 2003), successful and unsuccessful deductive reasoning strategies may each be associated with different eye movement patterns. As it is, the present study gives little indication of whether the unsuccessful reasoners were the only ones who had used mental models, or if they had merely used mental models less effectively than their successful counterparts. Finally, there are a few speculative directions for future research. Along the lines of Grant and Spivey’s study on Duncker’s radiation problem, it may be possible to increase accuracy rate in conditional reasoning by guiding eye movements in certain fashions. It may also be interesting to study the role of spatial reasoning ability in the construction of mental models during conditional reasoning and related tasks, which do not explicitly involve diagrams.

**Conclusions**

The current study has reinforced the notion of eye movements as a valid means of examining the cognitive basis of conditional reasoning. This was particularly evidenced by the correspondence between error rates and fixation patterns on premises in *modus tollens* problems. The findings discussed above replicate the effect of figural bias in conditional reasoning, observable in both
accuracy rates and eye movements, thus offering potential support for the mental
model account of deductive reasoning. Undoubtedly, many questions remain unasked
about the intersection between eye movement research and deductive reasoning
processes. We hope that the current study will help to spur numerous in-depth
investigations toward a cohesive account of human reasoning.
References


Table 1

Mean Response Accuracy Scores by Problem Type

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modus Ponens</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Modus Tollens</td>
<td>0.56</td>
<td>0.47</td>
</tr>
<tr>
<td>Hypothetical Syllogism</td>
<td>0.86</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Table 2

*Mean Overall Trial Dwell Times (ms) by Problem Type*

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modus Ponens</td>
<td>9,855</td>
<td>4,223</td>
</tr>
<tr>
<td>Modus Tollens</td>
<td>10,799</td>
<td>6,136</td>
</tr>
<tr>
<td>Hypothetical Syllogism</td>
<td>27,690</td>
<td>17,965</td>
</tr>
</tbody>
</table>
Table 3

*Mean Dwell Times (ms), First Run Dwell Times (ms), and Run Counts on Left (Hypothesis & Conclusion) and Right (Premise List) Sides*

<table>
<thead>
<tr>
<th>Interest Area</th>
<th>Modus Ponens</th>
<th>Modus Tollens</th>
<th>Hypothetical Syllogism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>M</em></td>
<td><em>SD</em></td>
<td><em>M</em></td>
</tr>
<tr>
<td>Total Dwell Time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HC</td>
<td>21,100</td>
<td>7,447</td>
<td>25,953</td>
</tr>
<tr>
<td>Premises</td>
<td>27,528</td>
<td>14,567</td>
<td>25,199</td>
</tr>
<tr>
<td>First Run Dwell Time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HC</td>
<td>2,508</td>
<td>775</td>
<td>2,922</td>
</tr>
<tr>
<td>Premises</td>
<td>1,297</td>
<td>685</td>
<td>1,411</td>
</tr>
<tr>
<td>Run Count</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HC</td>
<td>3.43</td>
<td>0.84</td>
<td>3.86</td>
</tr>
<tr>
<td>Premises</td>
<td>3.59</td>
<td>0.10</td>
<td>3.79</td>
</tr>
</tbody>
</table>

*Note.* HC = interest area consisting of the hypothesis and the conclusion.
Table 4

*Mean Dwell Times (ms), First Run Dwell Times (ms), and Run Counts on Hypothesis and Conclusion*

<table>
<thead>
<tr>
<th>Interest Area</th>
<th>Problem Type</th>
<th>Modus Ponens</th>
<th>Modus Tollens</th>
<th>Hypothetical Syllogism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>Total Dwell Time</td>
<td>2,026</td>
<td>631</td>
<td>2,558</td>
</tr>
<tr>
<td>Conclusion</td>
<td>First Run Dwell Time</td>
<td>2,047</td>
<td>839</td>
<td>2,758</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>Run Count</td>
<td>981</td>
<td>365</td>
<td>980</td>
</tr>
<tr>
<td>Conclusion</td>
<td>Run Count</td>
<td>780</td>
<td>265</td>
<td>1,149</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>Run Count</td>
<td>3.40</td>
<td>0.90</td>
<td>4.13</td>
</tr>
<tr>
<td>Conclusion</td>
<td>Run Count</td>
<td>3.71</td>
<td>1.40</td>
<td>3.80</td>
</tr>
</tbody>
</table>
Table 5

Mean Dwell Time (ms) on Each Premise by Order

<table>
<thead>
<tr>
<th>Premise Order</th>
<th>Modus Ponens</th>
<th>Modus Tollens</th>
<th>Hypothetical Syllogism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>1</td>
<td>712</td>
<td>114</td>
<td>935</td>
</tr>
<tr>
<td>2</td>
<td>1,122</td>
<td>127</td>
<td>1,279</td>
</tr>
<tr>
<td>3</td>
<td>1,311</td>
<td>208</td>
<td>789</td>
</tr>
<tr>
<td>4</td>
<td>1,256</td>
<td>153</td>
<td>746</td>
</tr>
<tr>
<td>5</td>
<td>528</td>
<td>117</td>
<td>597</td>
</tr>
<tr>
<td>6</td>
<td>558</td>
<td>114</td>
<td>849</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 6

*Mean Dwell Times (ms), First Run Dwell Times (ms), and Run Counts on Each Premise by Similarity to Correct Answer*

<table>
<thead>
<tr>
<th>Premise Type</th>
<th>Problem Type</th>
<th>Modus Ponens</th>
<th>Modus Tollens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td></td>
<td>Total Dwell Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB (Correct)</td>
<td></td>
<td>2,295</td>
<td>963</td>
</tr>
<tr>
<td>AC</td>
<td></td>
<td>928</td>
<td>437</td>
</tr>
<tr>
<td>BA</td>
<td></td>
<td>863</td>
<td>627</td>
</tr>
<tr>
<td>BC</td>
<td></td>
<td>460</td>
<td>403</td>
</tr>
<tr>
<td>DB</td>
<td></td>
<td>482</td>
<td>496</td>
</tr>
<tr>
<td>CD</td>
<td></td>
<td>459</td>
<td>468</td>
</tr>
<tr>
<td></td>
<td>First Run Dwell Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB (Correct)</td>
<td></td>
<td>902</td>
<td>444</td>
</tr>
<tr>
<td>AC</td>
<td></td>
<td>640</td>
<td>271</td>
</tr>
<tr>
<td>BA</td>
<td></td>
<td>534</td>
<td>194</td>
</tr>
<tr>
<td>BC</td>
<td></td>
<td>432</td>
<td>142</td>
</tr>
<tr>
<td>DB</td>
<td></td>
<td>462</td>
<td>234</td>
</tr>
<tr>
<td>CD</td>
<td></td>
<td>387</td>
<td>289</td>
</tr>
<tr>
<td></td>
<td>Run Count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB (Correct)</td>
<td></td>
<td>2.82</td>
<td>1.06</td>
</tr>
<tr>
<td>AC</td>
<td></td>
<td>1.42</td>
<td>0.64</td>
</tr>
<tr>
<td>BA</td>
<td></td>
<td>1.61</td>
<td>0.70</td>
</tr>
<tr>
<td>BC</td>
<td></td>
<td>1.04</td>
<td>0.75</td>
</tr>
<tr>
<td>DB</td>
<td></td>
<td>0.94</td>
<td>0.52</td>
</tr>
<tr>
<td>CD</td>
<td></td>
<td>1.07</td>
<td>0.52</td>
</tr>
</tbody>
</table>

*Note.* Premise types are labeled as described in the Method section.
Appendix

List of Stimuli

*Modus ponens 1:*

1. If you can pount, then you can pleit.
2. If you can pleit, then an ethis vecides.
3. If you can pleit, then you can pount.
4. If an ethis vecides, then it warates.
5. If you can pount, then an ethis vecides.
6. If an ethis warates, then you can pount.

You can pleit.
*Therefore, you can pount.*

*Modus ponens 2:*

1. If you have a spourmo, then the mixle will teter.
2. If the mixle contains pincture, then you can retoil.
3. If the mixle teters, then the mixle contains pincture.
4. If the mixle contains pincture, then the mixle will teter.
5. If the mixle teters, then you can retoil.
6. If you can retoil, then you have a spourmo.

The mixle contains pincture.
*Therefore, the mixle will teter.*
Modus ponens 3:

There is a skize.
Therefore, there is a punit.

1. If there is a punit, then there is a skize.
2. If there is a skize, then there is a punit.
3. If you were born in Kaloon, then you know how to sertave.
4. If you know how to sertave, then there is a skize.
5. If there is a skize, then you were born in Kaloon.
6. If there is a punit, then you were born in Kaloon.

Modus ponens 4:

The sturch is on.
Therefore, the plind is off.

1. If the bair is blue, then the kape is green.
2. If the plind is off, then the sturch is on.
3. If the sturch is on, then the bair is blue.
4. If the sturch is on, then the plind is off.
5. If the plind is off, then the bair is blue.
6. If the kape is green, then the plind is off.
### Modus ponens 5:

<table>
<thead>
<tr>
<th>The wamp is small.</th>
<th>Therefore, the tume is near.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. If the bact approaches, then the tume is near.</td>
<td></td>
</tr>
<tr>
<td>2. If the tume is near, then the rebic waves.</td>
<td></td>
</tr>
<tr>
<td>3. If the tume is near, then the wamp is small.</td>
<td></td>
</tr>
<tr>
<td>4. If the wamp is small, then the rebic waves.</td>
<td></td>
</tr>
<tr>
<td>5. If the rebic waves, then the bact approaches.</td>
<td></td>
</tr>
<tr>
<td>6. If the wamp is small, then the tume is near.</td>
<td></td>
</tr>
</tbody>
</table>

### Modus tollens 1:

<table>
<thead>
<tr>
<th>You do not see treits.</th>
<th>Therefore, you are not a spair.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. If you are a spair, then you see treits.</td>
<td></td>
</tr>
<tr>
<td>2. If you see treits, then you are a spair.</td>
<td></td>
</tr>
<tr>
<td>3. If you hear a slein, then you see treits.</td>
<td></td>
</tr>
<tr>
<td>4. If you are a spair, then you see a feln.</td>
<td></td>
</tr>
<tr>
<td>5. If you see treits, then you see a feln.</td>
<td></td>
</tr>
<tr>
<td>6. If you see a feln, then you hear a slein.</td>
<td></td>
</tr>
</tbody>
</table>
Modus tollens 2:

1. If the yart is open, then you nobilate.
2. If the meim is open, then the yart is open.
3. If the meim is open, then you nobilate.
4. If you nobilate, then you dilify.
5. If the yart is open, then the meim is open.
6. If you dilify, then the meim is open.

The meim is not open.
Therefore, the yart is not open.

Modus tollens 3:

1. If you vaceded, then you saw the jarp.
2. If you saw the jarp, then you went to the kein.
3. If you went to the kein, then you passed the rennel.
4. If you saw the jarp, then you passed the rennel.
5. If you passed the rennel, then you vaceded.
6. If you went to the kein, then you saw the jarp.

You did not see the jarp.
Therefore, you did not go to the kein.
Modus tollens 4:

**The glake can not be seen.**
**Therefore, you are not facing the vallume.**

1. If you are facing the vallume, then the glake can be seen.
2. If the gryn is red, then the glake can be seen.
3. If the glake can be seen, then you are facing the vallume.
4. If you are facing the vallume, then the sase is new.
5. If the glake can be seen, then the sase is new.
6. If the sase is new, then the gryn is red.

Modus tollens 5:

**The iver does not increase.**
**Therefore, the ander does not ring.**

1. If the spull speaks, then the tert hears.
2. If the iver increases, then the ander rings.
3. If the tert hears, then the iver increases.
4. If the ander rings, then the iver increases.
5. If the ander rings, then the spull speaks.
6. If the iver increases, then the spull speaks.
Hypothetical syllogism 1:

<table>
<thead>
<tr>
<th>The sorson is here.</th>
<th>1. If the bumole is free, then the sorson is here.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therefore, calove is likely.</td>
<td>2. If the sorson is here, then the bumole is free.</td>
</tr>
<tr>
<td></td>
<td>3. If the ronad moves, then the berda falls.</td>
</tr>
<tr>
<td></td>
<td>4. If the bumole is free, then calove is likely.</td>
</tr>
<tr>
<td></td>
<td>5. If the sorson is here, then the berda falls.</td>
</tr>
<tr>
<td></td>
<td>6. If calove is likely, then the bumole is free.</td>
</tr>
<tr>
<td></td>
<td>7. If the ronad moves, then calove is likely.</td>
</tr>
<tr>
<td></td>
<td>8. If calove is likely, then the sorson is here.</td>
</tr>
</tbody>
</table>

Hypothetical syllogism 2:

<table>
<thead>
<tr>
<th>The modgone goes.</th>
<th>1. If the luvet stays, then the modgone goes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therefore, the luvet stays.</td>
<td>2. If the jockle is pink, then the sorpo is orange.</td>
</tr>
<tr>
<td></td>
<td>3. If the modgone goes, then the safome sits.</td>
</tr>
<tr>
<td></td>
<td>4. If the safome sits, then the modgone goes.</td>
</tr>
<tr>
<td></td>
<td>5. If the jockle is pink, then the luvet stays.</td>
</tr>
<tr>
<td></td>
<td>6. If the modgone goes, then the sorpo is orange.</td>
</tr>
<tr>
<td></td>
<td>7. If the safome sits, then the luvet stays.</td>
</tr>
<tr>
<td></td>
<td>8. If the luvet stays, then the safome sits.</td>
</tr>
</tbody>
</table>
Hypothetical syllogism 3:

1. If the sustan is blank, then there are five scites.
2. If there are five scites, then the votis is new.
3. If there are five scites, then the sustan is blank.
4. If they have a lell, then you have a creard.
5. If the sustan is blank, then the votis is new.
6. If they have a lell, then there are five scites.
7. If the votis is new, then you have a creard.
8. If the votis is new, then the sustan is blank.

Hypothetical syllogism 4:

The vaxle is checked.

1. If the glone is round, then the prelda is red.
2. If the klab moves, then the glone is round.
3. If the prelda is red, then the vaxle is checked.
4. If the glone is round, then the vaxle is checked.
5. If the vaxle is checked, then the prelda is red.
6. If the klab moves, then the abers gather.
7. If the prelda is red, then the glone is round.
8. If the vaxle is checked, then the abers gather.
Hypothetical syllogism 5:

1. If the glire is found, then the plam approaches.
2. If the plam approaches, then the glire is found.
3. If the seng is black, then the plam approaches.
4. If the glire is found, then the biom waits.
5. If the biom waits, then the hinger increases.
6. If the biom waits, then the glire is found.
7. If the seng is black, then the hinger increases.
8. If the plam approaches, then the biom waits.

The biom waits.
Therefore, the plam approaches.