Summer 2012

Individual differences and basic logic ability

Christopher R. Runyon
James Madison University

Individual Differences and Basic Logic Ability

Christopher R. Runyon

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

August 2012

Acknowledgments

I would like to thank Dr. Richard West for his help and willingness to allow me to choose an independent topic for my thesis, in spite of the extra work that it brought upon himself and the other members of my thesis committee. I also would like to thank Dr. Laine Bradshaw for her insightful and detailed comments on the methodological aspects of this paper. I am also grateful to Dr. Michael Hall, as he spent several hours assisting me with the construction of this manuscript over the summer months, far beyond the necessary obligations of a committee member. I also would like to thank Dr. Bill Knorpp, Dr. Thomas Adajian, and Dr. Tracy Lupher for their collective help on the philosophical aspects of this manuscript, as well as their added comments in other sections.

I also would like to thank the undergraduate cognitive science research groups (PSYC 402) at James Madison University from the fall of 2010 until the spring of 2012. I could not have asked for a more enjoyable, intelligent group of students to aid in data collection efforts and for their willingness to allow me to present parts of the present manuscript to them for helpful feedback.

I am also grateful for my future wife, Becca, who bonded with me over the perils of graduate school and always provided me with unconditional support. Lastly, I would like to thank my family for their love and support in my decision to return to graduate school, as it allowed me to continue to pursue my love of academia and higher education.

Table of Contents

List of Tables

List of Figures

Abstract

The study of reasoning and information processing in cognitive science has often used problems derived from classical propositional logic inference rules in order to see how people make decisions, often comparing the qualities of those that can and cannot successfully complete these tasks. However, the majority of research that has been done has only focused on one inference rule: the material conditional. This narrow focus does not allow for inferences to be made about the role of logical ability *simpliciter* in cognitive science research. In order to better understand the relationship between cognitive ability and successfully completing tasks based on four binary logical connectives (conjunction, disjunction, material implication, and biconditional), 338 participants were given the Propositional Logic Test (PLT), a N-Back task, a Belief Bias Syllogisms Task, and the Cognitive Reflection Test, that latter two of which have been used in support of a dual-process theory of reasoning. Because no previous research exists examining the dimensionality of the PLT, multiple confirmatory factory analyses (CFA) were performed on the PLT to determine its factor structure. The best fitting model was a 2-factor model with a disjunction factor and conditionals factor, indicating that the PLT is multi-dimensional and there are limitations on its use as a summed score. Multiple regression analyses were then performed on the PLT and the two factors present to reveal what differences between participants may be masked by using the PLT as a summed score. The results indicate that ability to properly make the deductive inferences on the PLT is strongly associated with measures of Type 2 thinking and moderately associated with general intelligence. Furthermore, the disjunction factor was moderately related to both traditional measures of cognitive ability and Type 2 processing, and the

conditionals factor was strongly related to the ability to engage in Type 2 processing and only weakly related to traditional measures of cognitive ability. Thus, the ability to engage in specific types of deductive inferences requires different cognitive abilities, and the ability to engage in basic logical reasoning is significantly predicted by measures of general intelligence, but this alone is not sufficient.

CHAPTER 1

Introduction

The study of the psychology of reasoning has a long history in the tradition of psychology proper, but the topic also has roots in philosophy and religion that precedes this formal study by thousands of years. How people reason, and more narrowly, how people make specific types of inferences is central to the study of reasoning and theories of information processing. Not surprisingly, this study in psychology has often turned to evaluating how people make decisions or choices in tasks that are analogous to traditional (deductive) logical[1] inferences. The purpose of the present study is to identify what different measures of cognitive ability can significantly predict participants' performance on a series of tasks that include English translations of propositional logic connectives. However, it is first necessary to give an overview of propositional logic before introducing previous research that has been done on examining the role of logic in reasoning and decision-making.

Propositional Logic

The study of logic dramatically changed starting in the nineteenth century, as the classic syllogistic form of inference in the West (historically associated with Aristotle) was augmented with a richer conception of logic that introduced quantifiers and the propositional logic structure, providing a mathematical, symbolic treatment of logic. That is, Aristotelian formal logic only dealt with inferences that depended upon the quantificational form of arguments; modern logic kept this, but added to it, recognizing

[1] "Deductive logic" will seem redundant to some; here it is included to indicate that when I speak of logic in this manuscript it is intended to only include deductive logic.

other ways that arguments could be deductively valid (B. Knorpp, personal communication, June 19, 2012). More recently, logic has been applied in computer science and in the development of artificial intelligence, and it “has found radically new and important roles in computation and information processing” (Priest, 2010, p.vii).

In general, logic “consists of a formal or informal language together with a deductive system” (Shapiro, 2009, p.1)[2] that concerns the study of valid and invalid reasoning. More plainly, logic is a formal system[3] that allows for the analysis of expressions to see what other information can be validly derived from these expressions by using deductive inference rules. Mathematical logic treats logic as an abstract mathematical system that consists of a vocabulary of primitive symbols, rules that specify which strings of those symbols are well-formed, and a set of rules for transforming one well-formed string of rules into another (T. Adajian, personal communication, June 19, 2012).

The most well-known formal system is a type of mathematical logic called propositional logic, which is also known as the propositional calculus, sentential logic, or the sentential calculus. This formal system is often taught in the college philosophy departments, but may also be found in computer science or mathematics departments. It includes simple symbols that stand for atomic propositions (declarative sentences that cannot be further broken down into smaller sentences) and symbols that operate on the atomic propositions to form compound propositions, the latter of which are often referred

[2] This definition is contested, but will serve for present purposes.

[3] For more information on formal systems, see Kleene (1967) or Smullyan (1961).

to as logical connectives. The main binary operators that are traditionally taught are conjunction (represented by $\wedge$, and commonly translated into English as meaning "and"), disjunction ($\vee$, commonly translated into English as the inclusive "or"[4]), the material conditional[5] ($\rightarrow$, "If, then"), and the biconditional ($\leftrightarrow$, "if and only if" or "just in case")[6].

For example, the letter "P" may designate the sentence "Tom dances," and the letter "Q" may designate the sentence "Mary sings." These two atomic propositions can be combined by means of logical connectives to form a compound expression. Using the four binary connectives, one can now form several meaningful compound expressions: $P \wedge Q$ can symbolize "Tom dances and Mary sings;" $P \vee Q$ can symbolize "Tom dances or Mary sings;" $P \rightarrow Q$ can symbolize "If Tom dances, then Mary sings;" and $P \leftrightarrow Q$ can symbolize "If, and only if, Tom dances, then Mary sings" or "Tom dances just in case Mary sings."

A formal language is employed in formal systems so that the meaning of the words (sentence letters) in the language is fixed so that inferences can be studied independently of the non-logical elements. In an informal language, such as English, it

---

[4] It has been argued that the English "or" is linguistically indeterminate between the inclusive or exclusive "or." The difference between the two is when both of the given propositions are true. When both P and Q are true, then the inclusive "or" is true, whereas when both P and Q are true obtain, the exclusive "or" is false.

[5] The nature of the conditional is debated even within various systems of formal logic. See chapter 7 of Humbleton (2011).

[6] The symbols used to designate the different binary connectives may vary across different logic texts, although the concepts and English translations of these symbols are the same. The present symbols and translations can be found in Allen and Hand (2001).

may be unclear what is meant due to ambiguity in natural languages, and this can make it difficult to understand the inferences that are being made. Thus, propositional logic uses formal semantics to focus on what conclusions can be validly derived from a set of premises. That is, given that the premises are true, by using the rules of inference in the system one can then infer what other information *does* and *does not* necessarily follow (Beall, 2010).

The truth-value of an expression is determined by the inference rules of the formal system and the truth-values of the atomic propositions (sentence letters) involved in the expression. In classical propositional logic, propositions can only take one of two values: true or false. Thus, by having knowledge of the truth-value of the atomic propositions (if it is true that “Tom dances” and if it is true that “Mary sings”) and the how a specific logical connective functions in the system, it can be determined (via deduction) whether or not an expression formed from those atomic propositions is true. For example, if it is true that Tom dances, but not true that Mary sings, then “$P \wedge Q$” is false, “$P \vee Q$” is true, “$P \rightarrow Q$” is false, and “$P \leftrightarrow Q$” is false. Thus, there are four possible conditions that may obtain between any two propositions: both are true, one is true and the other is not (and its inverse), and both are false. See Table 1 for the truth tables for these four (conjunction, disjunction, material conditional, and biconditional) binary logical connectives.

CHAPTER 2

Review of the Literature

Previous Research on Logic in Psychology

In psychology, the study of logic in thinking and information processing has focused on only a few narrow aspects of the broad field of logic. This previous work has been centered around two lines of study: (1) seeing if one can complete a task that mirrors classical logical inferences (most often involving the material conditional) and comparing the qualities of those than can and those that cannot complete the task correctly; and (2) studying how material conditionals ("If, then" statements) are understood in tasks that are structurally similar but differ in their content.

The best known experiment on the role of logic in thinking is the Wason 4-card selection task, developed by Peter Cathcart Wason (1966). Participants were provided with a rule had the form of the material conditional from classical sentential logic, stating that (parenthetically) "If the card has a vowel on one side, then it has an odd number on the reverse side." Participants were then shown a series of cards: one with a vowel face-up, one with a consonant face-up, one with an odd number face-up, and one with an even number face-up. By giving the participant these cards, the participant is given (respectively) a true antecedent, a false antecedent, a true consequent, and a false consequent. Participants were then asked to turn over the cards that would invalidate the rule that they were given.

Most all participants turned over the vowel face-up card in order to check if the back has an odd number on it, checking to see if the true antecedent has a true consequent. However, most participants faltered in making an additional selection (if they did make an additional selection) – many turned over the odd-numbered card (a false consequent), checking to see if the reverse side has a vowel or consonant on it (a true consequent).

Unfortunately, turning over the odd number does nothing to test the validity of the rule. The rule doesn't stipulate that only vowels will have an odd number on their reverse side (which would be the case if the rule were a biconditional) – only that if it has a vowel on one side, it will have an odd number on the other. The second correct choice in this task is the even-numbered face-up card – the false consequent. Turning over this card to reveal a vowel could invalidate this rule, producing a situation with a true antecedent and false consequent. Usually fewer than 10% of participants choose this option (Stanovich & West, 1998; Wason & Johnson-Laird, 1972).

The Wason selection task has been replicated hundreds of times with each replication slightly varying different aspects of the task, such as altering the content of the problem to see if this influences participant responses (Evans, Newstead, & Byrne, 1993; Newstead & Evans, 1995). An interesting variant of the Wason task is the drinking-age problem (Griggs & Cox, 1982). In this problem, participants are presented with the rule of "if a person is drinking beer, then the person must be over 19 years of age," and presented with the cards of "drinking a beer," "drinking a coke," "16 years of age," and "22 years of age." When presented with this information, participants were much more successful in correctly completing the task - 73% of participants correctly select both of the cards necessary to test the validity of the rule (Griggs & Cox, 1982).

This marked difference in the ability to correctly complete the selection task has produced more research looking at differences between the two problems, as the Wason selection task utilizes abstract rules, whereas the drinking-age problem uses deontic rules (Stanovich & West, 1998). Stanovich and West (1998) gave participants a variety of selection tasks, with both deontic rules and nondeontic (abstract) rules. Two clear patterns of correct responses emerged: those participants that were able to get the nondeontic problems were "people of higher cognitive ability" (Stanovich & West, 1998, p.217) - those that had statistically significantly higher combined SAT scores (SAT Verbal and SAT Mathematics together). However, in deontic selection tasks, this difference in cognitive ability decreases significantly or disappears altogether.

In the Stanovich and West (1998) experiments, the SAT total score was used to compare the groups, thus leaving it unclear if there were more specific differences in the cognitive profiles of the participants that contributed to them correctly answering the different types of tasks. Additionally, only the SAT subscales were used to compare the groups, so it is unclear if additional characteristics of cognitive ability, such as those measured by non-traditional cognitive ability measures, might be more revealing as to predicting success on these tasks.

Psychological Theories of Reasoning with Conditionals

The interpretation of Wason's task, and, more broadly, the way humans use the conditional when reasoning, has given rise to several different theories of interpreting the way that people interpret "If, then" statements in English. The Mental Models Theory was made most famous by Philip Johnson-Laird, one of Wason's students (see Johnson-

Laird 2006; Johnson-Laird, Byrne, & Schaeken, 1992; Wason & Johnson-Laird, 1972). In this theory, it is hypothesized that people create mental models of the world when reasoning with conditionals, and that the outcomes of decisions are modeled when doing so. This theory assumes that humans are equipped with inference rules that preserve the truth-functional relationship with the antecedent and conditional found in truth tables of interpreting the material conditional in classical propositional logic. Thus, in employing this theory, reasoning with false antecedents is relevant in entertaining other possible ways that the consequent could obtain.

The suppositional theory (most associated with Jonathan St. B. T. Evans - another one of Wason's students – and David Over, e.g., in Evans & Over, 2004) posits that people reason given that the antecedent is true. That is, they consider only the cases in which the antecedent is true. Thus, when reasoning through conditionals they normally (heuristically) reason through those scenarios in which the antecedent is true, and may, analytically, reason through various counter-factual scenarios where the consequent may obtain without the conditional being true. On the suppositional theory, false antecedents are irrelevant in reasoning with "If, then" statements, as this false antecedent has no bearing on the consequent obtaining. The "If, then" statement is not truth-preserving in the same respect as the material conditional in classical logic; it not truth-functional in the manner that the mental models theory suggests.

The probabilistic reasoning theory (also known as Bayesian rationality – presently most associated with Mike Oaksford and Nick Chater, e.g., in Chater & Oaksford, 2004; Oaksford & Chater, 2007) suggests that people reason probabilistically. They determine the probability of the antecedent obtaining and then infer the probability of the consequent obtaining given the antecedent's probability. Thus, in this view, a false antecedent is given a low (or infinitesimal) probability, so reasoning with a false antecedent to a given outcome results in an expected low (but not necessarily zero) probability of the consequent obtaining given the antecedent.

A variant of the Mental Models theory has also recently been argued for by (e.g., Schaeken, Vandierendonck, Schroyens, & d'Ydewalle, 2007). According to this variant, people heuristically reason with conditionals probabilistically, but analytically reason with mental models. That is, when making faster, more intuitive judgments, people often employ the probabilistic approach to interpreting the conditional, and when people slowly deliberate about the conditional in question, they then take on the framework of the mental models theory.

While these theories of psychological interpretations of the conditional may seem different, if one looks at these theories within a dual-processing theory of reasoning framework a common theme emerges. Dual-process theories of cognition posit "two distinct processing mechanisms, which employ different procedures and may yield different, and sometimes conflicting, results" (St. B. T. Evans & Frankish, 2009, p.1). One of the best-known characterizations of the dual-process theory of cognition stems from Stanovich and West's extensive research on rationality that focuses on the differences between Type 1 and Type 2 processing[7] (Stanovich & West, 2000).

---

[7] This distinction has been characterized in several different ways with slight nuances between each of the descriptions. For a list of alternative terms used by theorists for Type 1 and Type 2 processing, see Stanovich, 2011, p. 18.

This distinction provides a framework in which many aspects of cognition can be detailed (such as thinking heuristics and biases), as these two types of processing describe two very different aspects of cognition. The most salient feature of Type 1 processing is its autonomy; the execution of Type 1 processing is rapid, requires little conscious awareness, and this processing is “mandatory when the triggering stimuli are encountered” (Stanovich, 2010, p.128). These processes also include emotional regulation, which “can produce responses that are irrational in a particular context if not overridden” (Stanovich, 2010, p.129). Type 1 processing is independent of high level control systems (possibly as an artifact of its evolutionary development), and can operate in parallel with other Type 1 or Type 2 processes without interference. This is a default mode of processing that employs several heuristics and biases, which can be both beneficial and detrimental in making decisions. This reliance on heuristics allows for quick processing and problem-solving, but non-optimal choices can often result due to this quick processing (Stanovich, 2010).

Type 2 processing is in stark contrast to Type 1 processing. Type 2 processing is a very controlled, serial, conscious type of processing that is very slow and computationally taxing, and a “critical function of Type 2 processing is its ability to override Type 1 processing” (Stanovich, 2010, p.129). This large amount of computational expenditure results in a narrow focus of awareness (a limitation of conscious attention) when engaging in Type 2 processing, and it is constituted by its more deliberate procedure to decision making.

Thus, when people think quickly and heuristically (engage in Type 1 processing) through a problem analogous to the material conditional, they do not entertain the possibility in which the consequent could obtain with a false antecedent, or even of scenarios with a false antecedent altogether. However, when people engage in Type 2 processing, they may entertain scenarios in which the antecedent fails to obtain, but yet the consequent obtains. All of the models suggest that participants that more deliberatively think through tasks employing the material conditional are more likely to be able to correctly complete the task.

Each of the previous theories described are supported by empirical studies and often the purpose of these studies is to provide evidence that supports one theory over another. For example, a study by Sevenants, Dieussaert, and Schaeken (2011) was comprised of three experiments to compare the mental models theory and the suppositional theory by testing the idea that a false antecedent was "irrelevant.[8]" Participants were given a series of selection tasks, and then were asked to indicate if specific information about the task was relevant to reasoning through the problem. The respondents were divided into two groups, those who indicated that a false antecedent was relevant and those who indicated a false antecedent was irrelevant. The study then compared the groups' abilities on a working-memory task developed by De Neys, d'Ydewalle, Schaeken, and Vos (2002) that had been shown to correlate highly with measures of intelligence (Engel, 1999), as well as with performance on the Cognitive Reflection Test (Frederick, 2005) as measures of cognitive ability. The study found that

[8] This idea is not new in thinking about conditionals; Pierce (1898) points out a similar debate between the Hellenistic philosophers Philo and Diodorus (p.125).

those participants who chose the false antecedent as being irrelevant to the problem were associated with higher cognitive ability, and, conversely, the group of participants who considered the false antecedent as relevant had lower cognitive ability.

The study of how people reason with conditionals and the differences between those who can and cannot complete tasks analogous to traditional logical paradigms has contributed to the literature on reasoning and information processing, but the scope of these studies in understanding the role of logic *simpliciter* in these fields is insufficient. While this study of logic in psychology has produced numerous books and research programs (for instance, besides those previously mentioned: Rips, 1994; Garnham & Oakhill, 1994; Leighton & Sternberg, 2004; Johnson-Laird, 2006; Manktelow & Chung, 2004; Stenning, 2002), it has been limited in its study of the place of basic logical ability in cognitive architecture and information processing. Given that logic entails understanding what does and does not *necessarily* follow from given information, better understanding the role of logic in information processing is an essential component to understanding how people reason. Thus, the development of measures of logical thinking that are psychometrically sound and that can be used in cognitive science research is prudent.

Propositional Logic Test

A search was made to identify any previously existing measures that would fulfill the role of measuring basic logic ability in psychological science research. The Propositional Logic Test (PLT; Piburn, 1989) was identified as a potential measure of interest, as it was the only measure found that attempts to test all four of the basic logical connectives used in propositional logic and thus would be able to give a more complete understanding of the ability of the participants to make specific types of deductive inferences. The test consists of 16 problems, four problems of each of the four traditional binary logical connectives: conjunction, disjunction, material conditional, and biconditional.

An advantage of the PLT is that it provides a simple way of testing the ability to understand and apply translated logical connectives in a simple (but slightly abstract) context, and it also tests all four of the basic logical connectives. The participants are given a sentence that includes a translation of a logical connective (e.g., “If it is round, then it is striped”) and four figures are displayed below the rule (e.g., a white circle, a white square, a striped square, and a round square). Students are then instructed to indicate which figures are allowed by the sentence and not allowed by the sentence. The four figures presented after each stem correspond with the four possible relationships between the two propositions (true ‘P’ and true ‘Q’, true ‘P’ and false ‘Q’, false ‘P’ and true ‘Q’, and false ‘P’ and false ‘Q’). Thus all four possible response options are represented, and whether or not an object is determined to be allowed or not allowed by the rule depends on the truth-conditions of propositions involved in the expression. On the PLT the biconditional was not translated using one of the traditional English translations (“P if and only if Q”); instead, it was translated as a conjunction of two material conditionals in the form of “If P, then Q, and if Q, then P.” This translation emphasizes the close relationship between the material conditional and the biconditional.

It was necessary that a participant mark all four of the figures correctly in order to get the item correct. This showed that the participant had a complete understanding of the logical connective tested in that problem. Additionally, by scoring the four options on a problem as a single item and not each option as a response, the influence of participant guessing was minimized. The number of correct responses on the PLT was summed into a total score. Additionally, subscale scores were created by summing together the number of correct responses on the items that included a translation of a specific logical connective. Thus, four subscale scores could be created that were purported to measure how well participants could correctly interpret conjunction items, disjunction items, material conditional items, and biconditional items.

The PLT has been shown to predict success on a test of scientific reasoning (Piburn, 1990), where both the total PLT score and scores on the four PLT subscales were used in the regression models. Almstrum (1999) used the test in order to identify shortcomings in the abilities of computer science students when applying logical connectives to simple problems, again using both the PLT total score and subscale scores. Almstrum also notes several studies (Kim, 1995; Owens & Seiler, 1996; Stager-Snow, 1985) in which the PLT was used, but results of these studies were non-significant, both statistically and practically. For example, a study by Stager-Snow (1985) found that the PLT had little predictive power for females in predicting final course grades in an introductory computer-science course for non-computer science majors, and no predictive power for males. However, in these previous studies mentioned by Almstrum (1999), it is not clear which scoring method is used for the PLT. Almstrum notes the use of the material conditional subscale in studies by Stager-Snow (1985) and Kim (1995), but is not explicit in mentioning which scoring method Owens and Seiler (1996) used for the PLT. It is unclear if PLT total scores, subscale scores, or both were used in these studies.

The original version of the PLT was a paper-and-pencil test where students were directed to circle those figures that were allowed by the rules and to cross through those figures that were not allowed by the rule. On the reverse side of the 16 problems was a directions page that gave an example of how to complete each of the four types of problems on the test. (See *Appendix A* for the original PLT.) This method of administration presented two possible problems: (1) the participant could use the examples as an aid on the test if they were unsure of how to answer a question (which Piburn said rarely happened; personal communication from Michael Piburn to Thomas Adajian, July 7, 2010), and (2) if a figure was not marked it could not automatically be assumed to be either “allowed by the sentence” or “not allowed by the sentence” If a figure was left unmarked, it could not be validly inferred that the participant intend for the unmarked figure to designate one choice or the other, as the participant may not know whether or not the figure is allowed by the sentence.

Given these potential difficulties with test administration, it was converted into a computerized format where participants clicked on the “allowed by the sentence” figures to turn them green and the “not allowed by the sentence” figures to turn them red. The items were presented one at a time on the screen, and participants were also required to mark an answer for each figure before going on to the next item. (See *Appendix B* for the computerized version of the PLT.)

The computerized version of the test was piloted against the original pencil-and-paper measure, and the computerized version of the test had slightly better coefficient alpha[9] levels than in Piburn's original report. More importantly, no data were lost. On the pencil-and-paper version of the measure administered at approximately the same time, almost half of the participants (17 of 32) were thrown out due to not following the directions. During the pilot testing the proctor repeated the specific method of marking the figures several times when giving the participants the directions for taking the test and also wrote these specific instructions for marking the figure on the classroom whiteboards for emphasis. Thus, it is reasonable to assume that data would be lost due to incomplete response patterns of the participants, should the paper-and-pencil version of the test continue to be used.

While the PLT was identified as a test that could be used as a measure of basic logic ability, research on this test found that the measure's development did not fulfill all of the aspects of Benson's (1998) strong program of construct validation. Benson's program of construct validation details the steps in which a construct undergoes validation, as this process is "the most critical step in test development and use because it is the process by which test scores take on meaning" (1998, p.10).

Benson's program consists of three stages: a substantive stage, a structural stage, and an external stage. In the substantive stage, the construct of interest is researched and

[9] Coefficient alpha is a measure of the reliability (internal consistency) of a scale or test.

theoretical and empirical definitions are made. The theoretical definitions are developed by the theory surrounding the construct of interest, and the empirical definition consists of identifying (or writing) items that would adequately represent the construct (which operationalizes the construct). The structural stage consists of only the empirical items themselves, as it examines how these items covary with one another, specifically seeing if these items covary in the manner specified in the theoretical domain. Often intercorrelations between items and subscales or factor analysis (exploratory or confirmatory) are employed at this stage in order to test the internal characteristics of the measure to ensure that it functions as intended. The final, external, stage comes about when the result of the structural stage analysis reflects that the scale or test is functioning internally as intended. In this stage the measure is then correlated with other established, related measures to see that it is representing the hypothetical construct correctly by having the expected relationships present.

In evaluating the PLT, Piburn (1989) appears to have addressed the substantive stage insofar as he notes that logic instructors reviewed the items and deemed them appropriate for measuring the basic logic connectives. Similarly, he addresses the external stage by correlating the PLT with several alternative measures of logical reasoning. Unfortunately, the internal structure of the test was not examined. This is an important stage, as the results of different tests of the internal structure help to understand if a test is being appropriately scored. A test should only be scored as a summed score and used as such in psychological research when it has been established that the test is unidimensional (Gerbing & Anderson, 1988). While the PLT will be used in the current investigation to address a research question (see below) related to the external stage of validation, a structural analysis of the PLT will additionally be provided as a necessary precursor to focusing on the results at the external stage.

Research Questions

Despite the wealth of the literature on role of logical thinking in the psychology of reasoning, relatively little research has been done to better ascertain the general logical abilities of participants by looking at a wider breadth of deductive reasoning tasks. Additionally, no studies have examined the characteristics of participants to see what types of cognitive ability measures may be significant predictors of this type of logical reasoning ability. While individual differences have been examined for the selection task (e.g., Stanovich & West, 1999) and syllogistic reasoning (e.g., Sá, West, & Stanovich, 1999), they have not been examined for tasks that attempt to measure the ability of participants to apply a wider breadth of propositional logic connectives. The current investigation will be able to provide a more robust understanding the relationship between basic logical reasoning and cognitive ability. In doing so, it will provide insight into the relationship between the ability to make specific types of deductive inferences and the cognitive abilities most strongly associated with making these deductive inferences.

The present study uses the PLT as a measure of basic logic ability because the PLT tests the four binary connectives used in propositional logic and can provide a broader conception of basic logical reasoning than merely testing one of the connectives alone. A multiple regression analysis will then be performed to see what cognitive ability measures best predict scores on the PLT. Of greatest interest is seeing if non-traditional measures of cognitive ability (i.e., the Belief Bias Syllogisms and the Cognitive Reflection Test) will be able to account for a significant amount of unique variance once all of the shared variance has been controlled for in multiple regression models.

As a necessary precursor to using the PLT as a measure of basic logic ability, the factor structure of the measure must be known. This is important for properly understanding the results of the multiple regression analyses. If the PLT is multi-dimensional, then a single regression analysis on the total score of the PLT can mask important differences between the cognitive abilities that are most strongly associated with each dimension of the PLT. Thus, a confirmatory factor analysis will be performed on the PLT to better understand its factor structure, a secondary benefit of the present study.

The mental processes that underlie the ability to use the four binary logical connectives on the PLT are significantly different from one another, and the result of the confirmatory factor analysis is expected to support that the PLT is a multi-dimensional measure. This expectation is based upon the notion that the success of properly employing one of the logical connectives is not necessarily dependent on successfully understanding how to employ the others. One exception to this expectation is between the material conditional and the biconditional; given their conceptual similarity, the ability to solve these types of problems will be represented by a single construct.

I hypothesize that the results of the multiple regression analysis will show that the ability to make the deductive inferences characteristic of logical reasoning on the PLT will be most strongly related to measures of cognitive ability that involve Type 2 thinking. This is because previous research has shown that Type 2 thinking significantly predicts other measures of logical reasoning (Sá, West, & Stanovich, 1999; Stanovich & West, 1999). This should be particularly true for the material conditional and the biconditional connectives, given that they are both cognitively sophisticated and are conceptually similar. Furthermore, given that the material conditional and biconditional connectives are conceptually distinct from the conjunction and disjunction connectives, it also was expected that the cognitive ability measures that significantly predict the PLT will vary based upon the type of deductive inference being made.

CHAPTER 3

Methods

Participants

Participants in the study were 338 students (249 women, 89 men) from a medium-size, mid-Atlantic public state university who were recruited through an introductory psychology participant pool in exchange for partial course credit. The mean age of the participants was 18.79 years old (*SD* = 0.92), and 87% of the participants were Caucasian. Participants were tested in small (2-4) groups in a private computer lab on campus, and participants were supervised by undergraduate research assistants.

Procedure

Upon arriving at the computer lab, participants signed the consent form, giving permission for verified SAT Mathematics and Verbal scores to be obtained from the university. Participants then answered demographic questions and proceeded to complete the study. The measures in the current analysis were presented along with other measures of cognitive ability, as well as various thinking disposition scales. In addition to the Propositional Logic Test (PLT), other measures of cognitive ability included the Cognitive Reflection Test, the Belief Bias Syllogisms task, and the N-back task. In the study the PLT was given after the Cognitive Reflection Test and the Belief Bias Syllogisms task, but before the N-back task. Those measures and scales not pertinent to the present analysis are excluded from the manuscript.

Cognitive Ability Measures

Cognitive Reflection Test. The Cognitive Reflection Test (CRT) was developed by Frederick (2005) as a “simple measure of one type of cognitive ability” (p.26), which is "predictive of the types of choices that feature prominently in tests of decision-making theories" (p.26). The test consists of 3 items, and the task has been described to "measure the tendency to override a prepotent response alternative that is incorrect and to engage in further reflection that leads to the correct response" (Toplak, West, & Stanovich, 2011, p.1275). That is, the items are such that an initial (false) answer quickly arises in participants, and in order to correctly respond to the item the participant must suppress this initial response and think through the problem more carefully[10]. The three items on the test are:

(a) A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost? ____ cents

(b) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ minutes

(c) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____ days

For item (a), the most common answer given, 10 cents, is false. If this were the case, then the bat and the ball would cost $1.20. The correct answer, 5 cents, is given by those that can suppress the heuristic, quick response of 10 cents and think through the implications of this initial response.

---

[10] Frederick (2005) provides substantial discussion for why one may reasonably infer that this is the mechanism underlying the task, including that the false responses are commonly the same response. See Frederick (2005) pp. 26-28 for this discussion.

The CRT has high correlations with measures of general cognitive ability, but has been shown to account for unique variability in a series of heuristics-and-biases tasks after traditional measures of cognitive ability have been controlled for (Toplak, West, & Stanovich, 2011). Thus, it measures a component of good thinking that is not captured on traditional measures of cognitive ability. In the dual-processing theory of cognition paradigm, this task measures how well one can override their Type 1 response and engage in Type 2 processing to correctly answer the problem (Toplak, West, & Stanovich, 2011, p.1275). The problems are in an open-answer format, and participants were only given credit for a correct response when the correct answer (10 [cents], 5 [minutes], 47 [days]) is provided.

Belief Bias Syllogisms. The Belief Bias Syllogism task was developed by Evans, Barson, and Pollard (1983) to test the idea that when people are given an argument to evaluate, they will more often default to using a priori knowledge instead of carefully examining the logical structure of the information that is given to them. The task consists of a series of 16 syllogisms that are presented to the participants, and they are asked to evaluate the syllogisms and indicate if the conclusion logically follows from the premises.

The current version of the task has been adapted from the initial Evans, Barson, and Pollard (1983) task and incorporates problems from the work on George (1995). The task consists of four valid syllogisms with believable conclusions, four valid syllogisms with unbelievable conclusions, four invalid syllogisms with believable conclusions, and four invalid syllogisms with unbelievable conclusions. Thus, a total of 8 of the problems are logically valid and 8 are not logically valid.

Four basic statements are used within the syllogisms: a universal affirmative statement (All X are Y), a universal negative statement (No X are Y), a particular affirmative statement (Some X are Y), and a particular negative statement (Some X are not Y). The tasks given to the participants consisted of presenting them with two premises and a conclusion, and participants were then asked to evaluate if the conclusion logically follows from the premises.

For example, a participant may be given the premises of “all living things need water” and “roses need water,” and then asked if the conclusion that “roses are living things” follows from these premises. This conclusion is believable, as it is true that roses are living things, but the conclusion does not follow from the premises. The first premise does not indicate that being a living thing is the *only* sufficient condition for needing water, and as such other things could need water but not be a living thing. The fact that a rose is, in the actual world, a thing that needs water provides the participant with conflicting results based whether the conclusion follows from the form or the content of the problem. (See *Appendix C* for the Belief Bias Syllogism items.)

Thus, in order to answer each item correctly the participant must evaluate the logical structure of the syllogism free from its content and base their answer on the structure alone. The “belief bias” is said to be present when participants are unable to successfully decouple the form from the content of the problem and responds only to the content of the problem. That is, participants will incorrectly answer a problem because they will be basing their answers on the believability of the conclusion, and not on the structure of the argument.

This task has been shown to provide evidence for the dual-process theory of reasoning, as those participants who were forced to quickly endorse an item after being exposed to it were more likely to demonstrate the belief bias, as they had less time to slowly deliberate through the problem (Evans & Curtis-Holmes, 2005). Thus, it is thought that those participants who get more of the items correct are more likely engaging in the slow, deliberate processing characteristic of Type II processing.

N-back Task. The N-back task (Kirchner, 1958) has traditionally been used as a measure of working memory, and working memory is characterized as referring to "processes used for temporarily storing and manipulating information in the face of ongoing processing and distraction" (Jaeggi, Buschkeuhl, Perrig, & Meier, 2010, p. 394). That is, it is a component of fluid intelligence that allows one to process new information while still keeping other recent information present. Working memory has also been shown to correlate highly with general intelligence (Conway, Kane, & Engle, 2003), but it has also been shown that working memory and general intelligence are not equivalent (Ackerman, Beier, & Bole, 2005).

The N-back task presents participants with a series of letters (in both lower-case and upper-case formats), and asks participants if a specific letter was *n* places back. For example, a participant may be presented with a series of letters: S F s D f d F D s F d. As each letter is presented, they would be asked if that same letter was presented *n* letters back. If the task was 2-back, then the participant would need to enter the following sequence to be correct (corresponding with the above example): N N Y N N Y Y Y N N N.

The current study used a series of 2-back and 3-back tasks, each with 16 possible correct positive responses out of 32 responses. A participant's score is the result of subtracting the percentage of correct positives from the percentage of false positives for the both the 2-back and 3-back task. These two percentages were then averaged for a single N-back score.

Recent research suggests that the traditionally low reliability of these N-back tasks indicates that inferences made from this measure may be limited (Jaeggi, Buschkeuhl, Perrig, & Meier, 2010). However, this study was specifically aimed at testing the psychometric properties of the different variants of the N-back task, where visuospatial, auditory, and dual tasks were tested. It was found that the N-back task may diverge from other tasks of working memory, suggesting that different processes are being measured by the different tasks, and that the N-back task is not a "pure" measure of working memory. However, for present purposes only a proxy of such functioning was desired in order to test the general relationship between working memory, other measures of cognitive ability, and the PLT.

Data Screening

Prior to the analysis, the data were screened to check for invalid response patterns, incomplete data, outliers, and both univariate and multivariate normality. The original sample contained 429 participants; 91 students were removed using listwise deletion for missing data for missing a score on at least one of the measures used in the present study. The largest numbers of participants (70) were excluded from the analysis because their verified SAT Mathematics score or SAT Verbal score was not on record.

Because the PLT is scored such that participants get an item correct or incorrect, a tetrachoric correlation matrix was estimated using PRELIS (Jöreskog & Sörbom, 2007) in order to understand the relationships between the items. Tetrachoric correlations were used because Pearson and Spearman correlations are attenuated in situations where the data are dichotomous (Finney and DiStefano, in press). Table 2 presents the tetrachoric correlation matrix for the items on the PLT, including how many participants got each item correct. Examination of the tetrachoric correlation matrix showed relationships between the items ranging from moderate, negative correlations (-.389) to very strong, positive correlations (.991). The tetrachoric correlation matrix also shows that the disjunction, material conditional, and biconditional items are correlating highly within each subscale, with very high correlations between the material conditional and biconditional items and moderate correlations between the disjunction and material conditional items and the disjunction and biconditional items. The conjunction items have low to moderate correlations within the subscale, and a wide range of correlations with the other items on the PLT, ranging from -.389 to .884.

Table 3 presents the distribution of scores on each of the subscales of the PLT, indicating that a large majority of the participants got all of the conjunction problems correct. Very few participants (5) got 0 or 1 of the conjunction items correct; thus, there is very little variance in the conjunction subscale scores. The scores on the disjunction subscale indicated that, in general, participants did well in completing these items, and there was moderate variability in the scores. Lastly, a large majority of the participants did not get any of the material conditional or the biconditional items correct, but there were some students that got some (and all) of these items correct; the distribution in these categories ranged from 9 – 22 participants. This is consistent with the results of Piburn (1990).

Table 4 presents the polychoric correlations between the subscales on the PLT. Most interestingly, the conjunction subscale scores have almost no relationship to the other subscale scores; at best, there is a very weak positive correlation with the disjunction subscale scores (.015), and there are weak *negative* correlations with the material conditional and biconditional subscales (-.011 and -.054, respectively). This is an artifact of the participants doing well on this subscale, as there is not much variance in the summed scores of the conjunction items. The polychoric correlation matrix also shows that the material conditional and biconditional subscales are highly correlated with each other (.903), and both of these subscales had moderate correlations with the disjunction subscale.

The zero-order Pearson correlations between the cognitive ability measures and the PLT (including the PLT subscales) are presented in Table 5. Examination of this correlation matrix shows that the conjunction subscale does not have a strong relationship with many of the other measures of cognitive ability. Again, because participant are doing so well on this subscale, there is little variance to explain and there are no factors that would distinguish those participants that are doing well (or poorly) on this measure.

The correlations between the material conditional subscale and the biconditional subscale with the other measures of cognitive ability are again similar.

Confirmatory Factor Analysis

A series of confirmatory factor analyses (CFA) were performed on the PLT in order to determine the dimensionality and factor structure of the test, as CFA allows researchers to test models of hypothesized interrelationships among items in a measure. This is desirable in order to understand how one should score and use the test in studies, should a theoretically plausible factor structure be identified and supported by the respective CFA. Previously the PLT was scored as a summed score, which is only appropriate when there is a unidimensional factor structure, as the score then represents the measurement of a single construct. Each of the 4 subscales has also been used in analyses to see how the ability to do well on these subscales relates to the other constructs of interest. This scoring method is only appropriate if there are 4 unique dimensions to the PLT, represented by each of the subscales.

Confirmatory factor analysis is most informative when several models can be specified a priori in order to make comparative and absolute judgments about the hypothesized relationships between the items. The PLT has been used in past studies (e.g., Piburn, 1990; Almstrum, 1999), and thus confirmatory factor analysis was appropriate.

Hypothesized Models

A total of nine models were tested in order to understand the factor structure of the PLT. Three of these models are plausible explanations of the data from all of the items on the PLT, four are intended to test the data from the individual subscales of the PLT without the influence of the other subscales, and two models are to test the hypothesis that a single construct is responsible for representing the data from the material conditional and biconditional items (again without the influence of the other subscales). The purpose of these tests is to identify the best-fitting model that is a plausible explanation of the data from the PLT, as well as to better understand the structure of the subscales and the possible relationships between the subscales. This information is essential to understanding how the PLT should be scored, thus allowing more precise inferences to be made about what measures of cognitive ability predict how well one can make deductive inferences based on the logical connectives used in propositional logic. See Figures 1-9, respectively, for diagrams of the nine hypothesized models.

Model 1 is a unidimensional model with all 16 items loading onto a single factor representing the construct of basic logic ability. This model was of interest because the PLT has previously been used as a summed score (which is only appropriate with a unidimensional factor structure) and the logical connectives tested are the traditional binary connectives that are taught at the same time in an introductory propositional logic class. Thus, it could be thought that the ability to correctly answer problems these binary logical connectives is closely related.

The second model (model 2) is a 4-factor model with each of the individual subscales representing a unique factor, where the subscales are moderately, but not highly, correlated with one another. The hypothesized model has the 4 conjunction items being explained by a construct representing the ability to understand how to properly interpret the conjunction connective (a “conjunction” factor), with the other 3 factors representing the other 3 logical connectives mutatis mutandis. This model represents the hypothesis that the abilities to properly interpret the four binary logical connectives are related to each other, although not enough to be considered to be the product of a single construct.

A similar model is hypothesized in model 3, but this model is a 3-factor model that collapses the distinction between two separate factors explaining the ability to correctly apply the material conditional connective and the ability to correctly apply the biconditional connective, resulting in a single “conditionals” factor. This hypothesis is supported by the similar concepts that the material conditional and the biconditional represent. Other, less common translations of these logical connectives show the similarities between the two: while “If P, then Q” is the most common translation for the material conditional, the connective is also translated as “P is sufficient for Q” or “Q is necessary for P;” the biconditional is also translated as “P is necessary and sufficient for Q” (Allen & Hand, 2001). Thus, the biconditional is the conjunction of two material conditional statements, and the wording of the biconditional items on the PLT makes this link even more explicit.

Models 4, 5, 6, and 7 are one-factor models representing the conjunction, disjunction, material conditional, and biconditional subscales, respectively. Each of these models treats one of the subfactors of Model 2 as a single factor. As in Model 2, each subfactor is measured by 4 items. These models were tested to see if each of the subscales were plausible explanations of the data from only that subscale. Additionally, this also shows the strength of the individual pattern coefficients from the subfactor to each of the items independent of the other subfactors on the PLT. However, if the subscales are significantly correlated, it would not be appropriate to use these scales independently, nor appropriate to use the scores from these subscales individually in research or practice.

The hypothesis that a single construct is responsible for driving the correct responses to the material conditional items and the biconditional items led to the two final models to be tested. Model 8 is a 2-factor model where the material conditional items and the biconditional items are explained by their respective factors, whereas model 9 posits that a single construct, the ability to correctly interpret conditionals, is responsible for explaining the scores on both of these sets of items. These two models provide information about the relationships among these two subfactors independent of the other subfactors.

Method of Estimation and Measures of Model Fit

Structural equation modeling was used to perform the confirmatory factor analyses on the proposed models, and LISREL 8.80 (Joreskog and Sorbom, 2007) was used to estimate the models. While maximum-likelihood (ML) estimation produces less biased estimates of fit and parameter estimates than other types of estimation under conditions of model misspecification (Olsson, Foss, Troye, & Howell, 2000; Olsson, Troye, & Howell, 1999), this is only appropriate when the data have a multivariate normal distribution, including that the data are continuous in nature. Since the indicators in the present models are dichotomous, categorical variable methodology (CVM) is more appropriate to estimate the models (Finney and DiStefano, in press). In such a situation, robust diagonally weighted least squares (DWLS) estimation is more appropriate for modeling the dichotomous data. This estimation method uses only the diagonal elements of the weight matrix used in weighted least squares (WLS) estimation; WLS estimation requires extremely large sample sizes in order to converge to an admissible solution, however DWLS can more easily converge on an admissible solution without such large sample size requirements. It has been shown that DWLS can adjust for departures from normality similarly to the Satorra-Bentler adjustment with ML estimation, and DWLS performs better (is less biased and gives more correct estimates) than WLS (Satorra & Bentler, 1994; Finney & DiStefano, in press).

To assess the fit of the proposed models, several absolute and incremental fit indices are reported. Absolute measures of model fit provide an absolute measure of how well the proposed model reproduces the observed covariance matrix, whereas incremental measures of model fit compare the reproduced model to a baseline model in which no relationships between the items are specified. The χ^2 statistic, the most common absolute fit measure, is reported, as it provides an exact test of the discrepancy between the observed covariance matrix and the model-implied covariance matrix. Hence, a non-significant χ^2 test implies that there is not significant amount of discrepancy between the two matrices, and that the suggested model is a plausible explanation of the hypothesized relationship between the variables. Because the χ^2 test is an exact test that is influenced by sample size, it is supplemented with several other fit indices[11]. Since the data is non-normal, robust χ^2 values from DWLS will be reported, as the robust χ^2 is adjusted for non-normality in the data. This robust χ^2 test has been shown to be fairly accurate in CVM, although there has been limited research on the misspecification of models in CVM (Finney & DiStefano, in press).

The fit indices reported are the standardized root mean square residual (SRMR, an absolute fit index; Joreskog & Sorbom, 1981, Bentler, 1995), the root mean square error approximation (RMSEA, an absolute fit index; Steiger & Lind, 1980; Steiger, 1990), and the comparative fit index (CFI, an incremental fit index; Bentler, 1989, 1990). Measures of approximate fit have been shown to vary under different conditions of model misspecification when using robust DWLS in CVM, but Yu (2002) found that the RMSEA and CFI performed well under these conditions. Additionally, these approximate fit indices are considered to be "robust" versions of the indices, as these are estimated with the adjustments made to the model by employing DWLS estimation.

The SRMR is an absolute index of model fit using a standardized (correlation) metric that is an indication of the average amount of discrepancy between actual and reproduced correlation matrices. This fit index is the very sensitive to misspecified factor covariances and moderately sensitive to factor loading misspecification (Hu and Bentler, 1998). A cutoff value of .08 or less is recommended (Hu and Bentler, 1999), but the results of Nye and Drasgow (2011) have found that more stringent cutoff standards (.04 or .03) are necessary with CVM, especially with large samples.

[11] As sample size increases, the χ^2 test becomes more powerful, rejecting plausible hypothesized relationships due to trivial differences between the covariance matrices.

RMSEA is an absolute fit index based on the non-centrality parameter that "assesses lack of fit due solely to model misspecification and provides a measure of discrepancy per degree of freedom" (Finney, Pieper, & Barron, 2004, p. 375). This index is sensitive to misspecified factor loadings (Hu & Bentler, 1998), and a values of .05-.08 indicate close fit and values of .10 or greater indicate poor fit (Browne & Cudeck, 1993). RMSEA was found to function well in CVM, although more stringent cutoff values are suggested (Nye & Drasgow, 2011).

The CFI is an incremental fit index that is based on the noncentrality parameter, and it compares the fit of the proposed model to an independence model, one where no relationships between the observed variables are said to exist. CFI is moderately sensitive to misspecified factor covariances, very sensitive to factor loading misspecification, and, importantly, it is not sensitive to sample size, meaning that it gives a reliable estimate at smaller sample sizes (Hu & Bentler, 1999). A cutoff of .95 or above is recommended for the CFI under conditions of multivariate normal data (Hu & Bentler, 1999), but Nye and Drasgow (2011) found that the "CFI appears to be less affected [than the Tucker-Lewis Index, another incremental fit index,] by nonnormality, although Type I errors [were] still higher in both moderate and severe skew conditions" (p.560). Thus, the CFI is still reported and the results were interpreted with this caution in mind, and more stringent cutoff values will be used.

In addition to these global fit indices, the standardized covariance residuals were also analyzed. These residuals represent the differences between the actual and reproduced covariance matrix, allowing one to identify areas of local misfit. Because the residual is standardized, these residuals can be interpreted as z-scores, and Byrne (1998) says that absolute values larger than 3 should be considered large. However, these covariance residuals are influenced by sample size, and larger samples will produce larger standardized covariance residuals than smaller samples with identical covariance matrices. Since these standardized covariance residuals are influenced by sample size, these residuals may not properly reflect areas of local misfit in the present sample, and the resulting analyses would be under-powered to detect areas of local misfit[12]. In these cases the residual covariance matrix should be converted into correlation matrix so the magnitude of the residual values can be more appropriately assessed. A cutoff of |.10| for individual correlation residuals has been recommended (Kline, 2011), although this recommendation is based on continuous data and more research is needed to better understand appropriate suggested cutoffs when estimating models with dichotomous data.

[12] This would also be appropriate for SEM procedures that are highly over-powered due to a very large sample size, as this would inflate the standardized covariance residuals.

CHAPTER 4

Results

Confirmatory Factor Analysis Results

The unidimensional, 1-factor model (model 1) for all 16 items converged, but inspection of the output revealed a negative error variance (-0.070) associated with item I4 (a material conditional item), which is an inadmissible solution. Models 2 and 3 also failed to converge to an admissible solution. In model 2, the phi matrix, representing the correlations between the exogenous variables (in this case, between the PLT subscales) was non-positive definite, and problems in estimating the phi matrix can result when there are high correlations between the some of the items or some of constructs in the model. The reproduced correlation between the conjunction factor and the material conditional factor was 2.36, and the reproduced correlation between the conjunction factor and the biconditional factor was 2.38; both out-of-bounds solutions. Additionally, the theta-delta matrix was also non-positive definite, which results when a negative error variance is estimated in one of the items; item I4 (a material conditional item) had a negative error variance (-0.039).

Model 3 failed to converge to an admissible solution, and the problems that were present in estimating model 2 were also present in model 3. The reproduced correlation between the conjunction factor and the conditionals factor was 2.50, and item I4 had a negative error variance (-0.084). Thus, models 1, 2, and 3 were not plausible explanations of the data from the 16 PLT items.

The model representing on the conjunction factor and the four conjunction items (Model 4) is a plausible explanation of the data from these items. (Table 6 presents the fit indices for the models that converged to admissible solutions and allows conclusions to be drawn about the research questions of interest.) The χ^2 test was non-significant, the CFI was above suggested cutoff values, the RMSEA was below suggested cutoff values, and the SRMR was within suggested cutoff limits. Additionally, only 1 standardized (correlation) residual was above |.10|. However, analysis of the pattern coefficients (relationships of the individual indicators and the factor) showed large differences between each indicators and the conjunction factor. The values of the pattern coefficients between the four conjunction items (I1, I2, I3, and I4) and the conjunction factor were (respectively) .33, .47, .79, and .68. (See Table 7 for the standardized pattern coefficients for the models that converged, indicating the correlations between each item and the factor explaining the item.) These values indicate that there is a large amount of variance in the conjunction items that was not explained by the construct that represents the ability to correctly complete the conjunction items. Cronbach's alpha for the factor was .388[13]. These results are consistent with the distribution of scores on the conjunction subscale, as there is not much variance to be explained by the model.

The model representing the disjunction factor and the four disjunction items (model 5) fit the data moderately well. While the χ^2 test was significant, this test is extremely sensitive to exact model fit, so the other fit indices and individual correlation

[13] Cronbach's alpha is a lower-bound (and thus conservative) estimate of the factor reliability. McDonald's (1999) omega coefficient is more commonly reported, as it produces more precise estimates of reliability (DeShon, 1998), however this is only in cases of linear structural equation modeling (SEM). In the cases of non-linear SEM, such as the present models, a different method of calculating the reliability of the construct is advised (Green & Yang, 2009).

residuals were examined to further understand the fit of the model. The CFI was above suggested cutoff values, and but the RMSEA was larger than the suggested cutoff values. However, because the RMSEA assesses misfit per degrees of freedom and thus “penalizes” models with low numbers of degrees of freedom (model 4 only has 2 degrees of freedom), the SRMR and standardized residuals were examined in order to draw conclusions about model fit. The SRMR was within suggested cutoff limits, and only 1 standardized residual was above |.10|, so the model was deemed to have adequate fit. The values of the pattern coefficients between the four disjunction items (D1, D2, D3, and D4) and the disjunction factor were (respectively) .79, .86, .75, and .86. These values indicate that the construct of the ability to solve disjunction problems was able to account for a modest amount of variance in each of the disjunction items. Cronbach’s alpha for the factor was .752.

Model 6, representing the material conditional factor and the four material conditional items, did not converge to an admissible solution. Item I4 had an error variance of -0.16, a pattern coefficient of 1.01 (representing a correlation higher than 1 with the factor), and the amount of variance explained in the item by the factor was 102%, all out-of-bounds values. Thus, the model is not a plausible explanation of the data from the items on the material conditional subscale.

The model with the biconditional items being explained by a biconditional factor (model 7), is a plausible explanation of the data from the biconditional items. The χ^2 test was non-significant, but the CFI was above suggested cutoff values and the RMSEA was below suggested cutoff values. The SRMR was extremely low (.022), and no standardized residuals were above |.10|. The values of the pattern coefficients between the four biconditional items (B1, B2, B3, and B4) and the biconditional factor were (respectively) .96, .95, .98, and .98. These values indicate that the construct of the ability to solve biconditional problems was able to account for a large amount of variance in each of the disjunction items. Cronbach's alpha for the factor was .899.

The models testing the hypothesis that the material conditional items and the biconditional items are explained just as well by a single factor (model 9) as by two separate factors (model 8) also did not converge to admissible solutions. Again, the error variance associated with item I4 was negative in both models.

Model Modifications

None of the models hypothesized to be plausible explanations of all 16 PLT items (models 1, 2, and 3) converged on an admissible solution. Given that a negative error variance was associated with item I4 in these models, as well as in models 6, 8, and 9, the tetrachoric correlation matrix was analyzed to diagnose the problem of model convergence. The tetrachoric correlations between the four material conditional items were very high within the subscale, and highest between I3 and I4 (.991). I4 and I3 also had similar correlations with I1 and I2, suggesting that these items are redundant. A cross tabulation of how participants responded to items I3 and I4 found that only 9 participants were on the off-diagonal, indicating that 329 of the participants either got both items correct or incorrect. This was the smallest off-diagonal number found when all four of the material conditional items were compared to one another.

The intent of the CFAs was to identify the dimensionality of the PLT so that the proper scoring method can be employed in order to better understand which measures of cognitive ability best predict performance on the factor (or factors) in the selected model.

In order to continue to try and identify the dimensionality of the PLT, item I4 was dropped and the tests of model 1 (the 1-factor model), model 2 (the 4-factor model), model 3 (the 3-factor model), model 8 (the 2-factor model explaining the material conditional and biconditional items), and model 9 (the 1-factor model explaining the material conditional and biconditional items) were tested again with the remaining 15 PLT items. These modified models are designated with the suffix “b” to reflect this change; see Figures 10-14 for diagrams of the updated models. Model 6, the model representing that the material conditional factor explaining the material conditional items, could not be tested as it was now saturated and just-identified, as it only has 3 indicators.

By examining the fit indices for model 1b, the unidimensional model with all 15 items representing a single “logical ability” factor, this model is not a plausible explanation of the data. The χ^2 test was significant, the CFI value was well below suggested cutoff values, the RMSEA was well above suggested cutoff values, and the SRMR was large, indicating several areas of local misfit. The pattern coefficients for the items ranged from -.11 to .98, indicating items ranged from being weakly and negatively associated with the factor to strongly and positively associated with the factor; see Table 7 for the standardized factor pattern coefficients for models 1b, 4b, and 5b.

Models 2b and 3b both did not converge to an admissible solution again. In both of these solutions, the phi matrix was non-positive definite. In Model 2b, the reproduced correlation between the conjunction subscale and the material conditional subscale was 2.31, and the reproduced correlation between the conjunction subscale and the biconditional subscale was 2.32; both out-of-bounds solutions. Similarly, in Model 3b the reproduced correlation between the conjunction subscale and the conditionals subscale was 2.54. Given that the conjunction subscale had little variance and did not correlate well with the other subscales, the inclusion of this subscale in models 2b and 3b was the reason that models were not able to converge to an admissible solution.

Model 8b, the 2-factor model with the material conditional factor explaining the 3 material conditional items and the biconditional factor with 4 items, is a plausible explanation of the data. The χ^2 test was significant, so the other fit indices were examined to further understand the fit of the model. The CFI was above suggested cutoff values, the RMSEA was below the suggested cutoff values, and the SRMR was very low. Importantly, no standardized (correlation) residuals were above |.10|. The values of the pattern coefficients between the three material conditional items (I1, I2, and I3) and the material conditional factor were (respectively) .93, .97, and .96. The values of the pattern coefficients between the four biconditional items (B1, B2, B3, and B4) and the biconditional factor were (respectively) .96, .96, .98, and .97. The correlation between the material conditional factor and the biconditional factor was .94. These values indicate that both of these constructs are able to account for a large amount of variance in these items.

Model 9b, a 1-factor model representing the conditionals explaining the 3 material conditional items and 4 biconditional items, is also a plausible explanation of the data, as the fit statistics for this model were similar to those found in model 10b. Again, no standardized residuals were above |.10|. The values of the pattern coefficients between the seven conditionals items (I1, I2, I3, B1, B2, B3, and B4) and the conditionals factor were (respectively) .91, .96, .94, .95, .95, .97, and .97. These values indicate that a single factor representing the conditionals items accounted for a large amount of variance in each of the conditional items. Cronbach's alpha for the factor was .924.

Model Modifications 2

Admissible solutions were now obtained for models 8b and 9b, so conclusions can be drawn about the hypothesis that a single conditionals factor explains the data from the material conditional and biconditional items. The fit statistics for model 8b and model 9 are similar, and the correlation between the two factors in model 8b was very high (.94). Thus, the hypothesis that a single factor explains the data of these items is supported, and the more parsimonious model is said to best represent the data from these two subscales.

However, no models still were able to provide insight into the dimensionality of the remaining 15 PLT items, as model 1b did not fit the data well and models 2b and 3b did not converge to an admissible solution. The conjunction subscale had little variance, which can lead to problems in obtaining an admissible solution when this subscale was included in a model, so this subscale and the accompanying four conjunction item were dropped. A final model was tested (3c), a 2-factor model with a disjunction factor and a conditionals factor, in order to see if this model could be a plausible explanation of the remaining 11 items on the PLT. See Figure 15 for a diagram of the hypothesized model.

This final model is a plausible explanation of the data from the disjunction, material conditional, and biconditional items. The χ^2 test was significant, but the CFI was above suggested cutoff values and the RMSEA was below the suggested cutoff values, even when using the more stringent values than suggested for models with non-normal data, as per Nye and Drasgow (2011). The SRMR was lower than suggested cutoff values and seven standardized residuals were above |.10|: two had values of -.17, two had values of |.15|, one had a value of -.14, one had a value of -.11, and one had a value of .10. The pattern coefficients between the disjunction subscale and D1-D4, were, respectively, .79, .85, .72, and .90, indicating that this factor is explaining a moderate amount of variance in these items. The pattern coefficients between the conditionals subscale and items I1-I3 and B1-B4, were, respectively, .91, .96, .94, .95, .94, .97, and .97, indicating that this factor is explaining a large amount of variance in their respective items. The correlation between the conditionals factor and the disjunction factor was .40. Thus, the 2-factor model (3c) was said to have adequate fit and to be a plausible explanation of the data from the 11 PLT items.

Multiple Regression Results

With the factor structure of the PLT properly understood, multiple regression analysis could be used to answer the main research questions of interest. I hypothesized that the cognitive ability measures that are associated with Type 2 thinking, the Belief Bias Syllogisms (Evans & Curtis-Holmes, 2005) and the Cognitive Reflection Test (Toplak, West, & Stanovich, 2011), would be significant predictors of performance on the PLT. I also hypothesized that the cognitive ability measures that significantly predict the ability to make the deductive inferences necessary on the two subscales present in the PLT will be different. Furthermore, because the items on the conditionals subscale are more cognitively sophisticated than those on the disjunction subscale, the BBS and CRT were expected to be more strongly associated with scores on the conditionals subscale than the scores on the disjunction subscale.

In light of the multi-dimensional factor structure of the PLT, it is possible that some of the significant predictors of performance on the PLT could be masked if relying solely upon a total score. To avoid this potential limitation, three multiple-regression analyses were planned. These included comparisons of which measures of cognitive ability are significant predictors of: (1) the participant's performance of the 11-item PLT; (2) the participant's performance on the disjunction subscale alone; and (3) the participant's performance on the conditionals subscale alone.

A final multiple regression model was planned to see which predictors might be able to best explain the variance on the conjunction subscale model. However, the factor analysis revealed that this subscale did not consistently explain a significant proportion of the variance in the conjunction items. Despite this lack of variance that limits the inferences than can be made from the conjunction subscale, the results of this analysis are reported for the sake of completeness.

In order to best understand the relationship and unique contribution of each predictor in the models, the variables were entered in four steps. The SAT Mathematics and SAT Verbal scores were entered first, as these scores have consistently been found to correlate highly with general intelligence in psychological science research. The N-back task was entered next; the inclusion of this task is to serve as another measure of general intelligence that would account for additional unique variance over and above the SAT scores, allowing for more informative interpretations of the other cognitive ability measures included. The Belief Bias Syllogisms task scores were entered next and the Cognitive Reflection Test scores were entered last. These last two measures are thought to be able to account for unique variance associated with Type II processing and higher levels of cognitive sophistication, after controlling for the shared variance between the measures. While the variables were entered one at a time in the model in order to better understand the effect of each measure in the models given the previous variables already in the model, the final models (including all of the cognitive ability measures) are of most interest. As such, only the final models for each dependent variable will be interpreted.

Importantly, even though the range of scores on the disjunction and conjunction subscales was limited (0 – 4), multiple regression analyses were used because this analysis has been demonstrated to be robust to violations of the assumption that the underlying variable of interest is normally distributed (i.e. large kurtosis and skewness absolute values) when sample sizes are large (greater than 200; Waternaux, 1976). The data from the conjunction subscale suggests that that this variable is not normally distributed in the sample, so even more caution is warranted (in addition to the previous cautions) when making inferences based on the results of the model predicting the conjunction subscale.

11-item PLT. The final model including all of the cognitive ability measures accounted for a significant percentage of the variance in the 11-item PLT ($R^2 = .343$, F (5,332) = 34.610, $p < .001$, 95% CI: .252 to .417), and more variance than the model with only SAT scores, the N-back task, and BBS ($R^2\Delta = .030$, $F\Delta$ (1,332) = 22.349, $p < .001$). The CRT ($\beta = .256$, $p < .001$, $sr = .210$), the Belief Bias Syllogisms ($\beta = .230$, $p < .001$, $sr = .192$), the N-Back Task ($\beta = .134$, $p = .008$, $sr = .120$), and SAT Verbal scores ($\beta = .119$, $p = .027$, $sr = .099$) all contributed to the model, while SAT Mathematics scores did not ($\beta = .070$, $p = .259$, $sr = .050$). Summaries of the predictors in each of the models can be found in Table 9. Table 9 shows that as the N-Back, BBS, and CRT are added into the models, the semi-partial correlations between SATM and SATV and the 11-item PLT decrease substantially, and that the measures indicative of Type II processing account for a significant amount of variance in the 11-item PLT scores. The multiple regression model summaries can be found in Table 10, showing that the addition of the N-Back, BBS, and CRT all produced statistically significant increases in the amount of variance explained in the 11-item PLT.

Disjunction Subscale. The final model including all of the measures accounted for a significant percentage of disjunction subscale score variance (R^2 = .218, F (5,332) = 30.830, p < .001, 95% CI: .134 to .289), and more variance than the model with only SAT scores, the N-back task, and BBS ($R^2\Delta$ = .010, $F\Delta$ (1,332) = 4.068, p = .045). The N-back task (β = .125, p = .021, sr = .112), Belief Bias Syllogisms (β = .164, p = .005, sr = .137), and the Cognitive Reflection Task (β = .119, p = .045, sr = .098) all contributed to the model, with SAT Math (β =.130, p = .053, sr = .094) and SAT Verbal (β = .106, p = .071, sr = .088) no longer contributing to the model. Summaries of the predictors in each of the models can be found in Table 11. Table 11 also shows that as the N-Back, BBS, and CRT are added into the models, the semi-partial correlations between SATM and SATV and the disjunction subscale scores decrease substantially, and that the measures indicative of Type II processing account for a significant amount of variance in the disjunction subscale. The multiple regression model summaries can be found in Table 12, showing that the addition of the N-Back, BBS, and CRT all produced significant increases in the amount of variance explained in disjunction subscale scores.

Conditionals Subscale. The model including all of the measures accounted for a significant percentage of the Conditional subscale variance (R^2 = .213, F (5,332) =

17.954, $p < .001$, 95% CI: .129 to .284), and more variance than the model with only SAT scores, the N-back task, and BBS ($R^2\Delta = .047$, $F\Delta$ (1, 332) = 19.670, $p < .001$). Interestingly, only the BBS scores (β = .192, $p = .001$, sr =.161) and CRT scores (β =.263, p = .<.001, sr = .216) contributed to the model, with SAT Mathematics (β = -.003, p = .960, sr = -.002), SAT Verbal (β =.083, p = .158, sr = .069), and the N-back task (β =.089, p = .104, sr = .079) not contributing to the model. Summaries of the predictors in each of the models can be found in Table 13, again showing that as the N-Back, BBS, and CRT are added into the models, the semi-partial correlations between SATM and SATV and the conditionals subscale decrease substantially and that the measures indicative of Type II processing account for a significant amount of variance in the 11-item PLT scores. The multiple regression model summaries can be found in Table 14, showing that the addition of the N-Back, BBS, and CRT all produced statistically significant increases in the amount of variance explained in the conditionals subscale.

When comparing the results of the multiple regression analysis on the 11-item PLT to the results on of the multiple regression analyses of the disjunction and conjunction factors, specific differences should be noted. Because there are seven items that are explained by the conditionals factor on the 11-item PLT, the significant predictors in the conditionals multiple regression model are more strongly related to the significant predictors in the 11-item PLT multiple regression model. The BBS and CRT in the 11-item PLT (semi-partial correlations of .192 and .210, respectively) are more similar to the relationships between these measures and the conditionals factor (.161 and .216) than in the disjunction factor (.137 and .098). The SAT Verbal score was more strongly related to the 11-item PLT score (.099) than to both the disjunction items alone

(.088) or the conditionals items alone (.069). The N-Back task also was more strongly related to the 11-item PLT (.120) than to the disjunction items alone (.112) and the conditionals items alone (.079).

It is important to note that inclusion of the N-Back task, Belief Bias Syllogisms, and the Cognitive Reflection Test impact the semi-partial correlations differently in the conditionals subscale multiple regression model than in the disjunction subscale multiple regression model. With inclusion of these measures in the models predicting the disjunction subscale scores and conditionals subscale scores, SATM has a semi-partial correlation of .094 (p = .053) with the disjunction subscale scores, whereas this correlation is -.002 (p = .960) in the conditionals model. SATV scores were less so affected, with a semi-partial correlation of .088 (p = .071) in the last disjunction multiple regression model and a semi-partial correlation of .069 (p = .158) in the last conditionals multiple regression model. The rank-order of the semi-partial correlations also differed in the models. The disjunction subscale was best predicted by (in order from largest to smallest semi-partial correlation) the BBS, the N-Back task, the CRT, SATM, and SATV. The conditionals subscale was best predicted by the CRT, BBS, the N-Back task, SATV, and SATM.

Conjunction Subscale. None of the models tested accounted for a significant percentage of conjunction subscales scores. However, while the final model did not account for a significant amount of variance in conjunction subscale scores, the N-Back task was a significant predictor in the model, (β = .124, p = .041, sr = .111). See Table 15 for summaries of the predictors in each of these models, showing that in the last model was the N-Back identified as a significant predictor. Table 16 presents the multiple regression model summaries, showing that no set of predictors were significant of the multiple regression models tested on the conjunction subscale. An ancillary regression analysis was performed using only the N-Back task as a predictor of the conjunction subscale scores, and it was found that this model was significant (p = .016). Because it was the only predictor in the model, the zero-order correlation between the N-Back task and the conjunction subscale (.131) is the same as the standardized coefficient (β).

CHAPTER 5

Discussion

As expected, the confirmatory factory analysis revealed that the PLT is not a unidimensional measure of logical ability. Rather, a 2-factor model (3c), with separate factors representing the items from the disjunction subscale and the items from the conditionals subscale, was found to best fit the data. While the PLT has been used as a summed score (Almstrum, 1999; Piburn, 1990), this finding suggests that using the PLT as such is not appropriate if the PLT is intended to be used as a unidimensional measure of basic logic reasoning. Additionally, it is not well-advised to use the four subscales present on the PLT as individual subscores (Almstrum, 1999; Kim, 1995; Piburn, 1990; Stager-Snow, 1985) for two reasons: (1) the conjunction subscale doesn't function properly due to the ceiling effect of the scores (See Table 3), causing a lack of variation in the items responses; and (2) the subscales created by the material conditional items and the biconditional items are explained by a single construct.

The PLT can still be important in cognitive science research. If a general understanding of the participants' ability to make the types of deductive inferences present in the disjunction, material conditional, and biconditional items is desired, the test may still be used as a summed score, although it must be understood that it is multidimensional and thus measuring several aspects of deductive reasoning. Additionally, given the high correlations within the items on the material conditional and biconditional subscales, fewer items may be used in order to assess these abilities, enabling a shorter measure to be created.

The two dimensions on the PLT can also be used as individual subscales. This is supported by the moderate fit of the disjunction subscale (model 5) and good fit of the conditionals subscale (model 9b) when these subscales were tested without the inclusion of the items from the other subscales. In the final model (model 3c), these factors were only moderately correlated (.40), further indicating that these abilities are distinct from one another.

More importantly, several conclusions also can be drawn about the relationship between basic logical reasoning and the other measures of cognitive ability. For example, as hypothesized, the result of the multiple regression analysis on the 11-item PLT was most strongly predicted by the measures indicative of Type 2 processing. This result means that the ability to make the deductive inferences on the 11-item PLT are more strongly associated with Type 2 processing than with general cognitive ability. Thus, when making certain deductive inferences that are shown to be valid in propositional logic, general intelligence alone is not sufficient to be able to understand what does and does not follow by necessity. The ability to engage in Type 2 processing, when slow, computationally taxing thinking occurs (Stanovich, 2010) is necessary in order to be able to make the correct choices on these tasks. This is consistent with suggestions from previous research on other types of logical thinking (e.g., Evans & Curtis-Holmes, 2005; Toplak, West, & Stanovich, 2011).

Because the PLT is multi-dimensional, it was also hypothesized that the relationship between the different factors on the PLT and the measures of cognitive ability would be different. The disjunction subscale was most strongly predicted by the Belief Bias Syllogisms, then the N-Back task, and then the Cognitive Reflection Test, the three statistically significant predictors in the model. However, while SATM and SATV were not statistically significant by traditional cutoff standards (p = .053 and p = .071), the total range of semi-partial correlations between the items was .088 to .137. These results have two implications: (1) that traditional general cognitive ability is influential in correctly completing the items testing the disjunction connective; and (2) that Type II processing is also influential in completing the disjunction items. That is, the type of thinking necessary to make the deductive inferences necessary to complete the disjunction problems is driven both by general intelligence and Type II processing, but neither of these dominates this type of thinking.

A contrasting outcome was obtained for the conditionals subscale. Specifically, the cognitive ability measures that have been associated with engaging in Type II processing were much stronger predictors of performance on the conditionals subscale. For example, the Cognitive Reflection Test had a .216 semi-partial correlation with the scores on the conditional subscale, and the Belief Bias Syllogisms had a .161 semi-partial correlation with the scores on the conditional subscale. Both of these semi-partial correlations are stronger than any of the semi-partial correlations in the disjunction multiple regression model. In contrast, SATV and N-Back, both of which are more indicative of general intelligence, were not statistically significant predictors (p = .158 and p = .104, respectively) and had much weaker semi-partial correlations with conditional subscale scores (.069 and .079, respectively). Likewise, SATM had no unique relationship to the conditionals subscale score (sr = -.002, p = .960). Taken collectively, these findings suggest that the type of thinking associated with Type 2 processing more strongly drives this type of deductive inference. One must be able to engage in slow, deliberate processing in order to be able to understand what does and does not follow by necessity when reasoning with the material conditional and biconditional.

Two similar mistake patterns in the responses to the material conditional items and the biconditional items indicate that the thinker that is not engaging in the type of thinking characteristic of Type II processing, and this was most responsible for the participant not correctly completing these items. Participants often did not select figures in which the antecedent of the material conditional is false as being allowed by the sentence, and participants often did not select figures in which the both atomic propositions were false in the biconditional as being allowed by the sentence. Being able to solve both of these types of problems requires the participant to engage in deliberative thinking, as they must consider the implications of the sentence for the figures which are *not* directly spoken about in the sentences provided to the participants.

For example, the first material conditional item has the sentence of "If it is large then it is round," and is followed by a large circle, a small square, a large square, and a small circle (these are all striped and tailed, but this information is irrelevant to the stem). In order to correctly mark all of the figures correctly, the participant must consider what implications the rule has when the figure is *not large* (small). A participant that is engaging in quick, heuristic thinking will not engage in the deliberative type of thinking necessary to think through the implications of the rule for the small items. That is, since the rule starts with "If it is large," a participant engaging in heuristic processing will not consider what this rule implies for figures that are *not* large. It is necessary to understand that "If it is large then it is round" does not prohibit small figures from being allowed, and thus they should be marked as "allowed by the sentence."

Correctly marking all of the figures for the biconditional items follows a similar, deliberative method of reasoning in which figures that are not directly spoken about are not prohibited by the sentence. For example, the first biconditional item has the sentence of "If it is round it is striped and if it is striped it is round," and is followed by a white square, a striped circle, a striped square, and a white circle. In order to correctly mark all of the figures correctly, the participant must think through what the implications are of the sentence for the figure that is *not round* and is *not striped* (a white square). Again, there is nothing present in the sentence that prohibits a white square from being allowed by the sentence, and correctly identifying this figure as being allowed by the sentence requires the participant to follow the implications of the sentence all the way through for all of the figures presented.

A large number of participants often make these mistakes in the material conditional and biconditional items. They fail to mark the figures with false antecedents as "allowed by the sentence" in sentences that are translations of the material conditional, and a similar number of participants fail to mark those figures in which neither proposition obtains in the biconditional as being "allowed by the sentence." See Table 17 for an indication of the abundance of incorrect responses to these specific figures.

In summary, the results of the present study show that the ability to make the types of basic deductive inferences found in propositional logic is related more to the ability to engage in sophisticated thinking than to general intelligence. Furthermore, the degree of cognitive sophistication required to make these inferences differs as a function of the form of the deductive inference being made. The ability to logically reason through deductive tasks is multidimensional and different cognitive abilities are necessary to reason through problems on these different dimensions. Thus, while the ability to engage in logical reasoning is clearly associated with general intelligence, intelligence cannot solely predict the ability to make valid deductive inferences.

Future Research with the Propositional Logic Test

Future research should continue to develop tests that appropriately measure different aspects of deductive reasoning in propositional logic. Classical propositional logic concerns the study of valid inferences and no current measure (or set of measures) appropriately captures the breadth of this study. The PLT only captures a small part of this breadth, and revisions to the PLT and the creation of other measures would be beneficial to better understand the relationship between deductive inference, decision-making, and cognitive ability.

More broadly, it is important to note that the classical view of symbolic logic has been rivaled by different philosophies of logic and philosophical logics, such as temporal, modal, relevantist, intuitionist, many-valued, and conditional logics (e.g., Beall, 2010; Burgess, 2009; Priest, 2008). These views of logic question the axioms, inference rules, and truth-functionality of classical logic, as well as expand upon the very foundations of logic itself. That is, current measures in psychological science are insufficient to assess classical propositional logic inferences, and classical propositional logic is only a small part of the current study of logic *simpliciter*. There has been very little research to date in psychological science non-classical logic. However, there has been some work (e.g., Ripley, 2009) on these topics in the field of experimental philosophy, a new branch of philosophy the employs experimental methodology often found in psychological science studies to have empirical data to inform philosophical questions. Future research in psychology should incorporate these advances in the study of logic in order to better understand the role of deductive inference in reasoning.

In addition to the insufficient study on classical propositional logic and its variants, there is a rich history of the rigorous study of logic in non-Western cultures that has not yet been examined in modern cognitive science. For example, there is a long history of Tibetan monastic debate (Perdue, 1992; Perdue, in press) and extensive scholarship of signs and semiotics in the Tibetan monastic tradition (Rogers, 2009), but this study has not yet been formalized or incorporated into the study of reasoning in the West.

Table 1
Truth Tables for Conjunction, Disjunction, Conditional, and Biconditional Logical Connectives

P	Q	$\wedge$	$\vee$	$\rightarrow$	$\leftrightarrow$
T	T	T	T	T	T
T	F	F	T	F	F
F	T	F	T	T	F
F	F	F	F	T	T

Note: $\wedge$ = conjunction; $\vee$ = disjunction; $\rightarrow$ = material conditional; $\leftrightarrow$ = biconditional.

Table 2

Tetrachoric Correlations of the Items on the Propositional Logic Test

	C1	C2	C3	C4	D1	D2	D3	D4	I1	I2	I3	I4	B1	B2	B3	B4
C1	--															
C2	.297	--														
C3	.198	.355	--													
C4	.188	.275	.554	--												
D1	.284	-.101	-.018	-.289	--											
D2	.186	-.282	.155	-.009	.751	--										
D3	.092	.062	.223	-.123	.486	.609	--									
D4	.195	-.083	.094	-.274	.657	.695	.725	--								
I1	.260	.863	-.184	-.019	.238	.162	.233	.363	--							
I2	.849	.289	-.289	-.054	.236	.282	.281	.376	.910	--						
I3	.226	.281	.074	.031	.325	.488	.305	.489	.882	.940	--					
I4	.273	.328	-.352	-.002	.260	.340	.264	.432	.885	.978	.991	--				
B1	.856	.328	-.233	-.082	.260	.284	.193	.432	.867	.899	.848	.889	--			
B2	.845	.110	-.253	.128	.358	.315	.101	.343	.858	.905	.848	.868	.940	--		
B3	.168	.373	-.389	-.216	.301	.215	.217	.323	.836	.875	.886	.892	.922	.889	--	
B4	-.041	.884	-.285	-.155	.381	.353	.217	.450	.856	.786	.886	.857	.908	.889	.967	--
Incorrect	24	30	31	32	124	97	188	120	298	302	303	296	296	305	288	288
Correct	314	308	307	306	214	241	150	218	40	36	35	42	42	33	50	50
% Incorrect	7.1	8.9	9.2	9.5	36.7	28.7	55.6	35.5	88.2	89.3	89.6	87.6	87.6	90.2	85.2	85.2
% Correct	92.9	91.1	90.8	90.5	63.3	71.3	44.4	64.5	11.8	10.7	10.4	12.4	12.4	9.8	14.8	14.8

Note. C1-C4 = the four conjunction item on the PLT (PLT Items 1, 6, 9, and 14, respectively); D1-D4 = the four disjunction items on the PLT (PLT Items 2, 5, 10, and 13, respectively); I1-I4 = the four material conditional items on the PLT (PLT Items 3, 8, 11, and 16, respectively); B1-B4 = the four biconditional problems on the PLT (PLT Items 4, 7, 12, and 15, respectively).
N = 338.

Table 3

Distribution of Scores on the Propositional Logic Test Subscales

	Score							
Subscale	0	1	2	3	4	5	6	7
Conjunction	1	4	19	63	251	--	--	--
Disjunction	58	33	56	86	105	--	--	--
Material conditional	283	12	9	13	21	--	--	--
Biconditional	274	11	17	14	22	--	--	--
Conditionals	266	12	10	9	12	7	10	12

Note: Range of scores on the conjunction, disjunction, material conditional, and biconditional subscales was 0-4. The conditionals subscale range of scores was 0-7; one material conditional item was deleted because it was redundant.

Table 4

Polychoric Correlations between the Propositional Logic Test Subscales

	Conjunction	Disjunction	Material Conditional	Biconditional
Conjunction	1.00			
Disjunction	.015	1.00		
Material Conditional	-.011	.284	1.00	
Biconditional	-.054	.308	.903	1.00

	Conjunction	Disjunction	Conditionals
Conjunction	1.00		
Disjunction	.015	1.00	
Conditionals	-.022	.302	1.00

Note: Range of scores for the subscales is 0-4. The conditionals subscale score is the combination of scores on the material conditional and biconditional subscales.
N = 338.

Table 5

Zero-order Correlations between the Cognitive Ability Measures and the Propositional Logic Test

	SATV	SATM	N-Back	BBS	CRT	CONJ	DISJ	MAT CON	BICON	CONDS	PLT11	PLT16
SATV	1.0											
SATM	.508	1.0										
N-Back	.397	.307	1.0									
BBS	.471	.439	.339	1.0								
CRT	.555	.331	.303	.350	1.0							
CONJ	.075	.105	.131	.038	-.009	1.0						
DISJ	.377	.322	.301	.356	.322	.006	1.0					
MAT CON	.311	.271	.233	.375	.371	.023	.202	1.0				
BICOND	.289	.264	.259	.307	.350	-.017	.202	.773	1.0			
CONDS	.310	.280	.258	.349	.383	.005	.213	.915	.960	1.0		
PLT11	.433	.381	.353	.450	.455	.007	.708	.774	.806	.841	1.0	
PLT16	.439	.393	.372	.452	.439	.233	.653	.799	.794	.843	.971	1.0
SD	69.166	70.549	0.229	2.598	0.914	0.668	1.448	1.131	1.176	1.892	2.616	2.939
Mean	561.805	557.515	0.536	10.73	0.71	3.64	2.435	0.453	0.512	0.846	3.28	7.06
Skew	0.081	0.014	-0.661	0.332	1.099	-2.167	-0.512	2.391	2.13	2.199	1.01	1.161
Kurtosis	0.252	0.53	0.26	-0.848	0.191	5.06	-1.092	4.219	3.074	3.55	0.857	1.348

Note. SATV = SAT Verbal; SATM = SAT Mathematics; N-Back = N-Back Task; BBS = Belief Bias Syllogisms; CRT= Cognitive Reflection Test; CONJ = PLT Conjunction Subscale; DISJ = PLT Disjunction Subscale; MAT CON = PLT Material Conditional Subscale; BICON = PLT Biconditional Subscale; CONDS = PLT Conditionals Subscale, a combination of the revised Material Conditional Subscale and the Biconditional Subscale; PLT11 = PLT with only the Disjunction and Conditionals Subscales; PLT16 = 16-Item Propositional Logic Test.

N = 338.

Table 6
Fit Statistics for the Proposed Propositional Logic Test Models

Model	χ^2_{NTWLS}	$\chi^2_{S\text{-}B}$	df	$CFI_{S\text{-}B}$	$RMSEA_{S\text{-}B}$	SRMR
Model 1b	36987.78	630.99	90	0.74	0.13	0.280
Model 3c	2313.34	66.45	43	0.99	0.04	0.060
Model 4	15.63	1.32	2	1.00	0.00	0.053
Model 5	75.52	11.57	2	0.99	0.12	0.052
Model 7	248.16	5.42	2	1.00	0.07	0.022
Model 8b	1171.21	29.35	13	1.00	0.06	0.027
Model 9b	1489.33	38.64	14	1.00	0.07	0.038

Note: $CFI_{S\text{-}B}$ = robust comparative fit index; $RMSEA_{S\text{-}B}$ = robust root mean square error of approximation; SRMR = standardized root mean square residual.
N = 338.

Table 7

Standardized Factor Pattern Coefficients from the Confirmatory Factor Analysis of the Propositional Logic Test

	Model 1b	Model 3c		Model 4	Model 5	Model 7	Model 8b		Model 9b
Item	Basic Logic Ability Factor	DISJ	COND	CONJ	DISJ	BICON	MAT CON	BICON	COND
C1	.88	--	--	.33	--	--	--	--	--
C2	.89	--	--	.47	--	--	--	--	--
C3	-.11	--	--	.79	--	--	--	--	--
C4	-.08	--	--	.68	--	--	--	--	--
D1	.56	.79	.00	--	.79	--	--	--	--
D2	.59	.85	.00	--	.86	--	--	--	--
D3	.46	.72	.00	--	.75	--	--	--	--
D4	.64	.90	.00	--	.86	--	--	--	--
I1	.94	.00	.91	--	--	--	.93	.00	.91
I2	.95	.00	.96	--	--	--	.97	.00	.96
I3	.93	.00	.94	--	--	--	.96	.00	.94
B1	.96	.00	.95	--	--	.96	.00	.96	.95
B2	.94	.00	.94	--	--	.95	.00	.96	.95
B3	.96	.00	.97	--	--	.98	.00	.98	.97
B4	.98	.00	.97	--	--	.98	.00	.97	.97

Note. CONJ = conjunction factor; DISJ = disjunction factor; MAT CON = material conditional factor; BICON = biconditional factor; C1-C4 = the four conjunction items on the PLT; D1-D4 = the four disjunction items on the PLT; I1-I3 = the three material conditional items on the PLT; B1-B4 = the four biconditional items on the PLT. One material conditional item (I4) was removed, as models with this item failed to converge.
N = 338.

Table 8

Summary of Multiple Regression Predictors in Predicting the 11-Item Propositional Logic Test

Model & Predictors	B	SE B	β	t	p	sr
Model 1						
SATM	.012	.002	.323	5.776	<.001	.278
SATV	.008	.002	.217	3.883	<.001	.187
Model 2						
SATM	.010	.002	.261	4.543	<.001	.215
SATV	.007	.002	.190	3.427	.001	.162
N-Back Task	2.191	.595	.191	3.683	<.001	.174
Model 3						
SATM	.007	.002	.189	3.257	.001	.150
SATV	.005	.002	.127	2.287	.023	.105
N-Back Task	1.758	.585	.154	3.006	.003	.138
Belief Bias Syllogisms	.255	.055	.254	4.640	<.001	.213
Model 4						
SATM	.003	.002	.070	1.131	.259	.050
SATV	.004	.002	.119	2.214	.027	.099
N-Back Task	1.529	.569	.134	2.687	.008	.120
Belief Bias Syllogisms	.232	.054	.230	4.326	<.001	.192
CRT	.732	.155	.256	4.727	<.001	.210

Note: *B* = unstandardized regression coefficient; *SE B* = standard error of the unstandardized regression coefficient; *β* = standardized regression coefficient; *sr* = semi-partial correlation of the predictor and the conjunction subscale score.
N = 338.

Table 9
Summary of Models Predicting Performance on the 11-Item Propositional Logic Test

Model	R	R^2	df	F	p	$R^2\Delta$	df	$F\Delta$	p
1	.472	.223	2, 335	47.987	<.001	--	--	--	--
2	.503	.253	3, 334	37.712	<.001	.030	1, 334	13.564	<.001
3	.546	.298	4, 333	35.405	<.001	.045	1, 333	21.529	<.001
4	.585	.343	5, 332	34.610	<.001	.044	1, 332	22.349	<.001

Note: $R^2\Delta = R^2$ change; $F\Delta = F$ change.
$N = 338$.

Table 10

Summary of Multiple Regression Predictors in Predicting the Disjunction Subscale Score

Model & Predictors	B	SE B	β	t	p	sr
Model 1						
SATM	0.006	.001	.288	4.963	<.001	.248
SATV	0.004	.001	.176	3.045	.0003	.152
Model 2						
SATM	0.005	.001	.235	3.929	<.001	.194
SATV	.003	.001	.154	2.656	.008	.131
N-Back Task	1.019	.344	.161	2.966	.003	.146
Model 3						
SATM	.004	.001	.185	3.016	.003	.147
SATV	.002	.001	.110	1.867	.063	.091
N-Back Task	.854	.344	.135	2.483	.014	.121
Belief Bias Syllogisms	.098	.032	.175	2.966	.003	.147
Model 4						
SATM	.003	.001	.130	1.939	.053	.094
SATV	.002	.001	.106	1.813	.071	.088
N-Back Task	.795	.344	.125	2.313	.021	.112
Belief Bias Syllogisms	.092	.032	.164	2.831	.005	.137
CRT	.189	.094	.119	2.017	.045	.098

Note: *B* = unstandardized regression coefficient; *SE B* = standard error of the unstandardized regression coefficient; *β* = standardized regression coefficient; *sr* = semi-partial correlation of the predictor and the disjunction subscale score.
N = 338.

Table 11
Summary of Models Predicting Performance on the Disjunction Subscale

Model	R	R^2	df	F	p	$R^2\Delta$	df	$F\Delta$	p
1	.407	.165	2, 335	33.186	<.001	--	--	--	--
2	.432	.187	3, 334	25.572	<.001	.021	1, 334	8.798	.003
3	.457	.208	4, 333	21.921	<.001	.022	1, 333	9.107	.003
4	.467	.218	5, 332	18.512	<.001	.010	1, 332	4.068	.045

Note: $R^2\Delta$ = R^2 change; $F\Delta$ = F change.
N = 338.

Table 12
Summary of Multiple Regression Predictors in Predicting the Conditionals Subscale Score

Model & Predictors	B	SE B	β	t	p	sr
Model 1						
SATM	0.006	0.002	0.226	3.799	<.001	.195
SATV	0.004	0.002	0.165	2.771	.006	.142
Model 2						
SATM	0.005	0.002	0.181	2.918	.004	.149
SATV	0.004	0.002	0.145	2.43	.016	.124
N-Back Task	1.172	0.463	0.142	2.528	.012	.129
Model 3						
SATM	0.003	0.002	0.119	1.882	.061	.094
SATV	0.002	0.002	0.091	1.508	.132	.075
N-Back Task	0.904	0.461	0.109	1.961	.051	.098
Belief Bias Syllogisms	0.158	0.043	0.216	3.635	<.001	.182
Model 4						
SATM	<.001	0.002	-0.003	-0.05	.960	-.002
SATV	0.002	0.002	0.083	1.414	.158	.069
N-Back Task	0.734	0.45	0.089	1.63	.104	.079
Belief Bias Syllogisms	0.14	0.042	0.192	3.306	.001	.161
CRT	0.543	0.123	0.263	4.435	<.001	.216

Note: *B* = unstandardized regression coefficient; *SE B* = standard error of the unstandardized regression coefficient; *β* = standardized regression coefficient; *sr* = semi-partial correlation of the predictor and the conditionals subscale score.
N = 338.

Table 13
Summary of Models Predicting Performance on the Conditionals Subscale

Model	*R*	R^2	*df*	*F*	*p*	$R^2\Delta$	*df*	$F\Delta$	*p*
1	.341	.117	2, 335	22.097	<.001	--	--	--	--
2	.365	.133	3, 334	17.099	<.001	.016	1, 334	6.39	.012
3	.408	.166	4, 333	50.112	<.001	.033	1, 333	13.21	<.001
4	.461	.213	5, 332	51.338	<.001	.047	1, 332	19.67	<.001

Note: $R^2\Delta = R^2$ change; $F\Delta = F$ change.
$N = 338$.

Table 14
Summary of Multiple Regression Predictors in Predicting the Conjunction Subscale Score

Model & Predictors	B	SE B	β	t	p	sr
Model 1						
SATM	.000	.001	.029	.467	.641	.025
SATV	.001	.001	.090	1.429	.154	.078
Model 2						
SATM	<.001	.001	-.006	-.099	.922	-.005
SATV	.001	.001	.074	1.173	.242	.063
N-Back Task	.324	.174	.111	1.861	.064	.101
Model 3						
SATM	<.001	.001	.078	.078	.938	.004
SATV	.001	.001	1.294	1.294	.197	.070
N-Back Task	.342	.176	1.938	1.938	.053	.105
Belief Bias Syllogisms	-.011	.017	-.643	-.643	.521	-.035
Model 4						
SATM	<.001	.001	.635	.635	.526	.034
SATV	.001	.001	1.337	1.337	.182	.072
N-Back Task	.362	.177	2.051	2.051	.041	.111
Belief Bias Syllogisms	-.009	.017	-.513	-.513	.608	-.028
CRT	-.066	.048	-1.376	-1.376	.170	-.074

Note: *B* = unstandardized regression coefficient; *SE B* = standard error of the unstandardized regression coefficient; *β* = standardized regression coefficient; *sr* = semi-partial correlation of the predictor and the conjunction subscale score.
N = 338.

Table 15

Summary of Models Predicting Performance on the Conjunction Subscale

Model	R	R^2	df	F	p	$R^2\Delta$	df	$F\Delta$	p
1	.108	.012	2, 335	1.980	.140	--	--	--	--
2	.148	.022	3, 334	2.484	.061	.010	1, 334	3.463	.064
3	.152	.023	4, 333	1.963	.100	.001	1, 333	0.413	.521
4	.169	.029	5, 332	1.953	.085	.006	1, 332	1.893	.170

Note: $R^2\Delta$ = R^2 change; $F\Delta$ = F change.
N = 338.

Table 16

Number of Participants Correctly Identifying the False Antecedent of the Material Conditional and the Two False Propositions of the Biconditional as "Allowed By The Sentence"

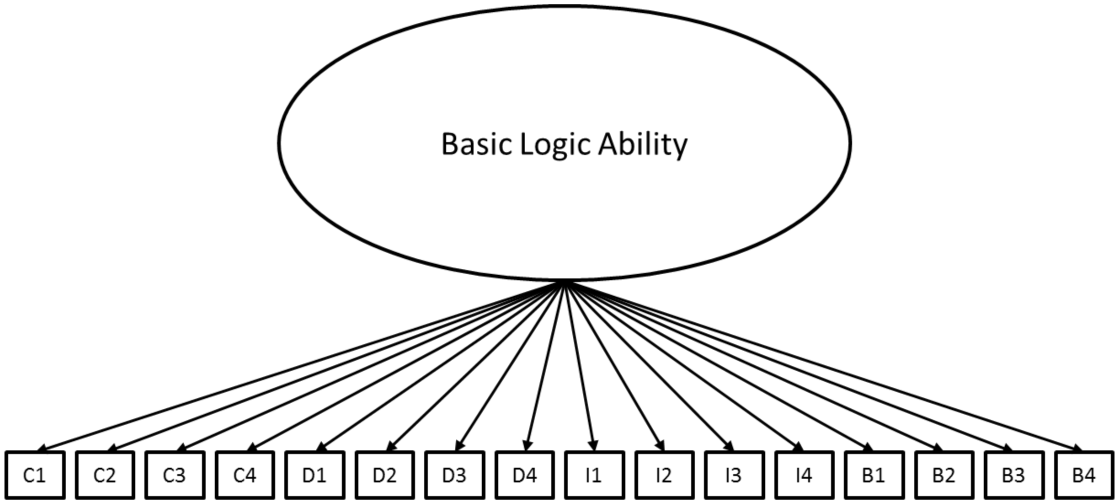
	Material Conditional False Antecedent and False Consequent			Material Conditional False Antecedent and True Consequent		
	I1	I2	I3	I1	I2	I3
Correct	53	57	66	46	43	41
Incorrect	285	281	272	292	295	297

	Biconditional Two False Propositions			
	B1	B2	B3	B4
Correct	49	37	61	61
Incorrect	289	301	277	277

Note. According to the propositional logic truth tables, these items should be "allowed by the sentence."
N = 338.

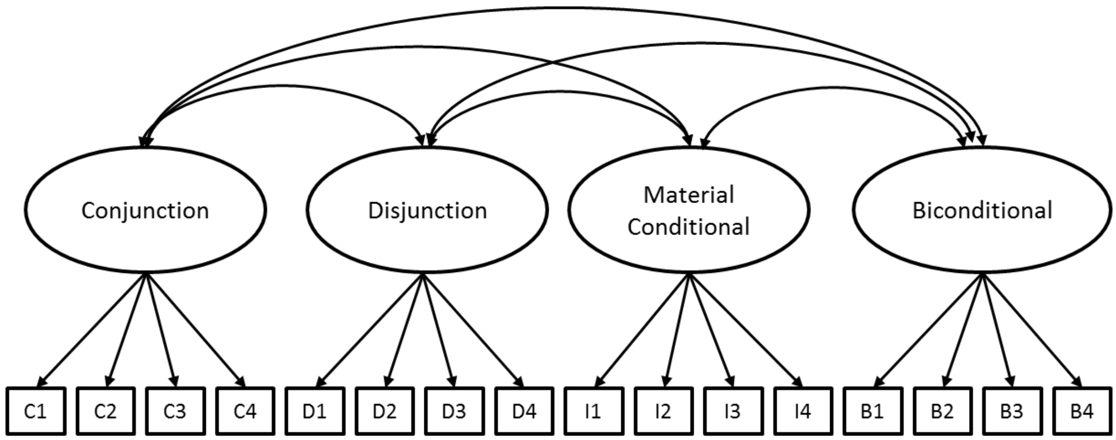
Figures

Figure 1.



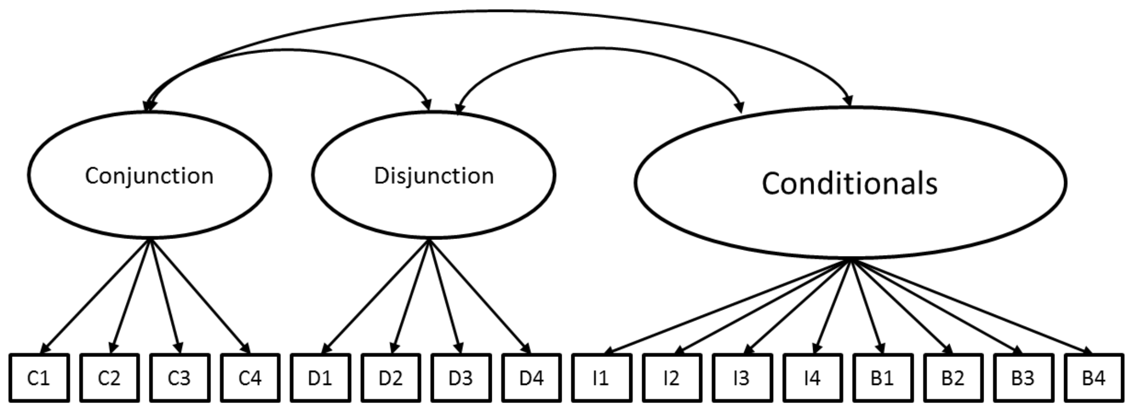
Model 1. A one-factor model based on all 16 Propositional Logic Test items representing the construct of basic logic ability.

Figure 2.



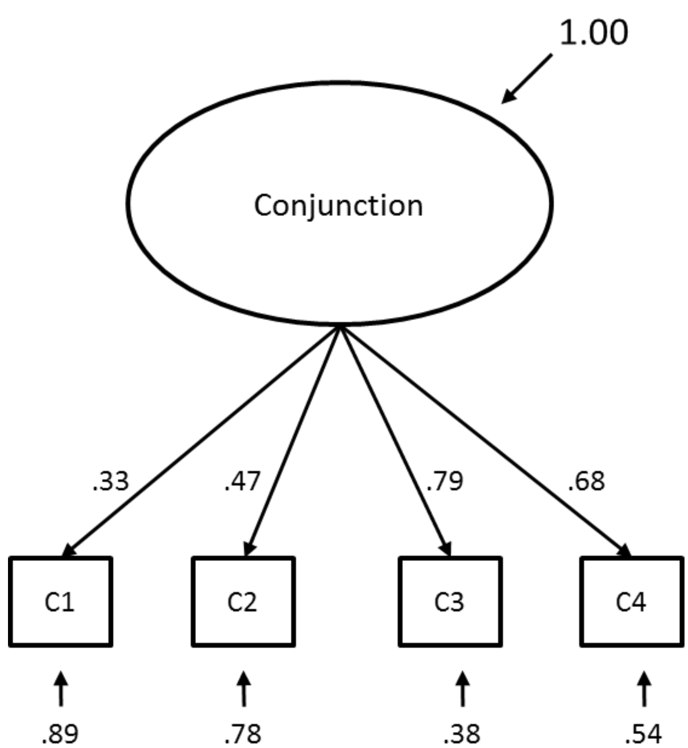
Model 2. A 4-factor model based on all four of the binary logical connectives on the individual subscales of the Propositional Logic Test correlating with one another.

Figure 3.



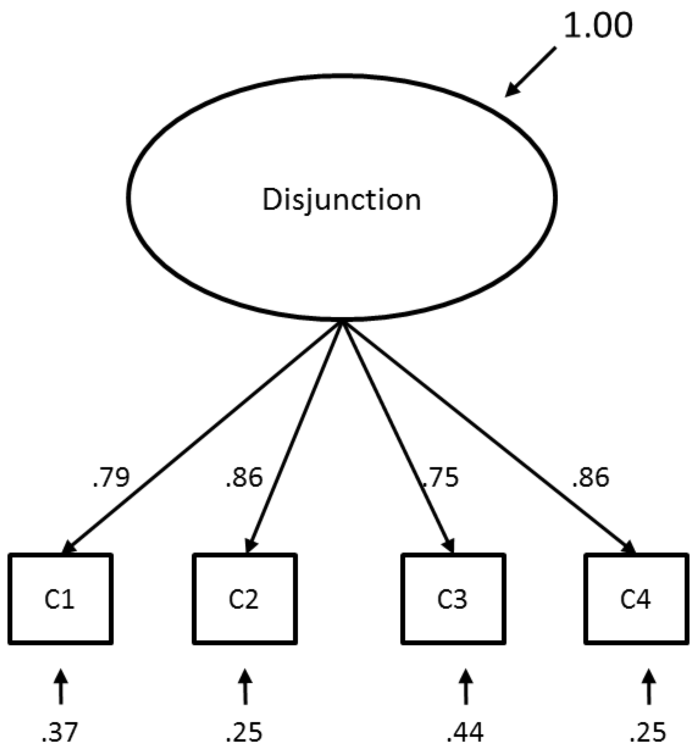
Model 3. A 3-factor model combining the material conditional and biconditional subscales of the Propositional Logic Test into a conditionals subscale, with the three Propositional Logic Test subscales correlating with one another.

Figure 4.



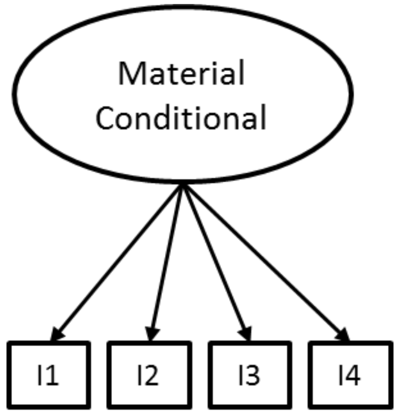
Model 4. A model representing the conjunction subscale of the Propositional Logic Test. Standardized results are reported.

Figure 5.



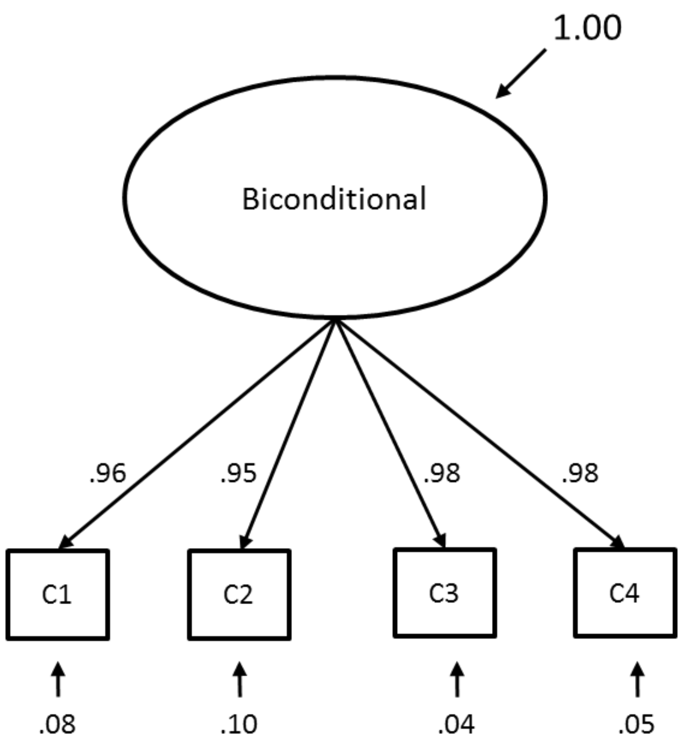
Model 5. A one-factor model representing the disjunction subscale of the Propositional Logic Test. Standardized results are reported.

Figure 6.



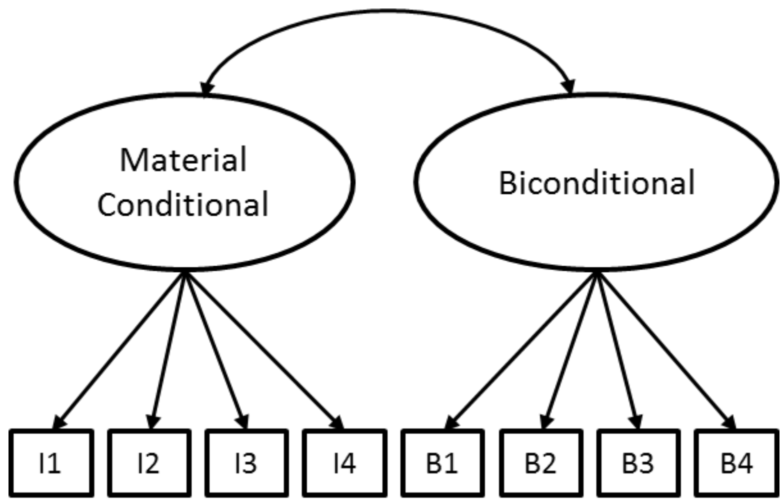
Model 6. A one-factor model representing the material conditional subscale of the Propositional Logic Test.

Figure 7.



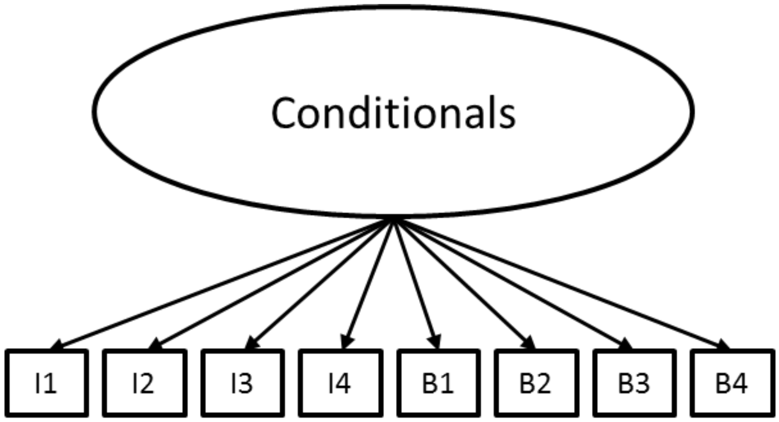
Model 7. A one-factor model representing the biconditional subscale of the Propositional Logic Test. Standardized results are reported.

Figure 8.



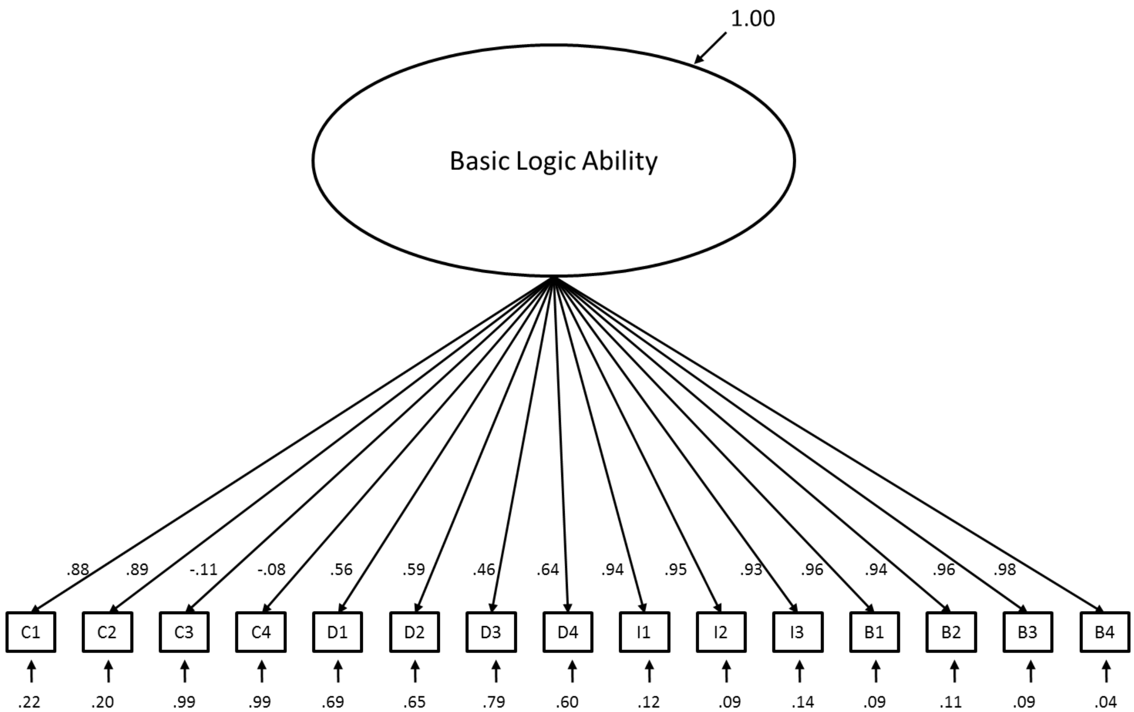
Model 8. A two-factor model representing the material conditional and biconditional subscales of the Propositional Logic Test correlating with one another.

Figure 9.



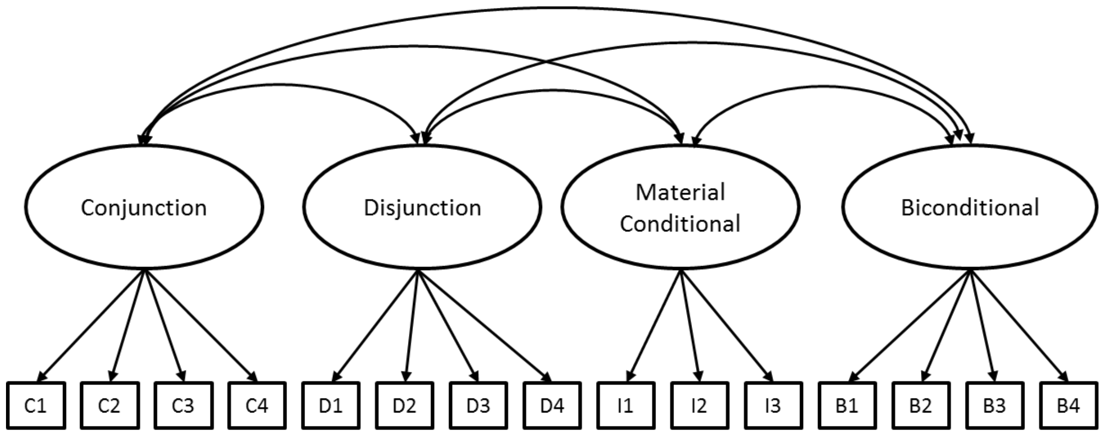
Model 9. A one-factor model representing the combination of the material conditional and biconditional items representing a conditionals factor.

Figure 10.



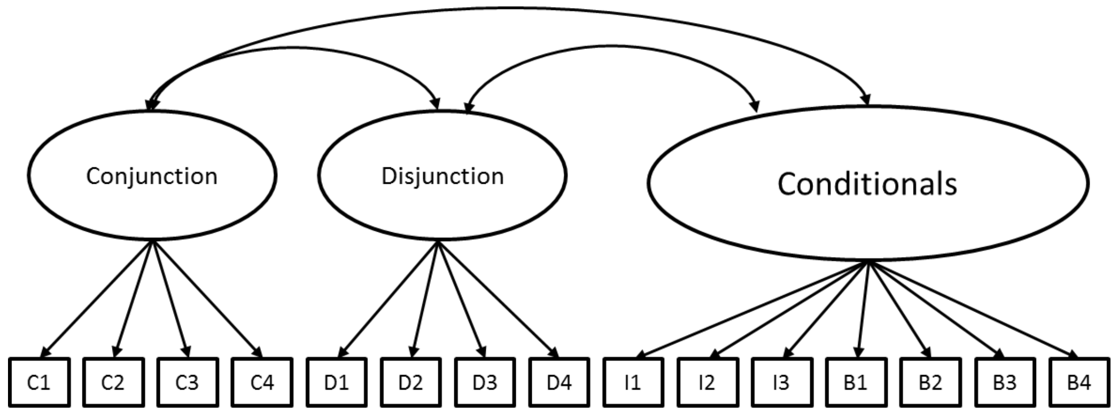
Model 1b. A one-factor model based on 15 Propositional Logic Test items representing the construct of basic logic ability; one of the material conditional items (I4) has been removed from Model 1.

Figure 11.



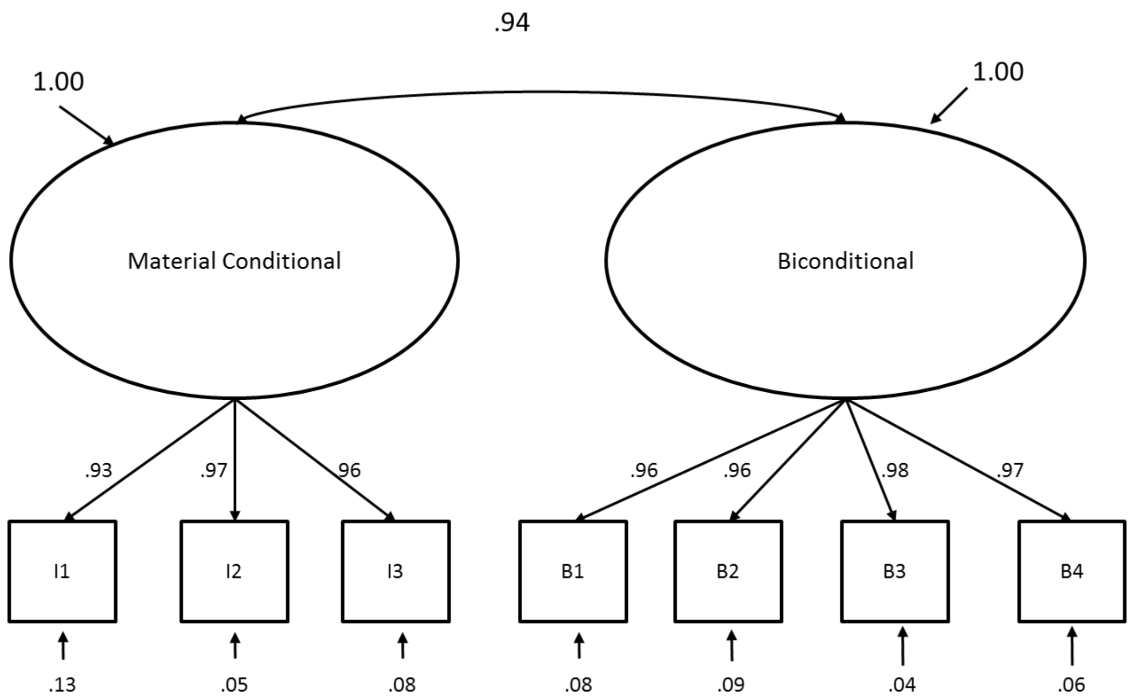
Model 2b. A 4-factor model based on all four of the binary logical connectives on the individual subscales of the Propositional Logic Test correlating with one another; one of the material conditional items (I4) has been removed from Model 2.

Figure 12.



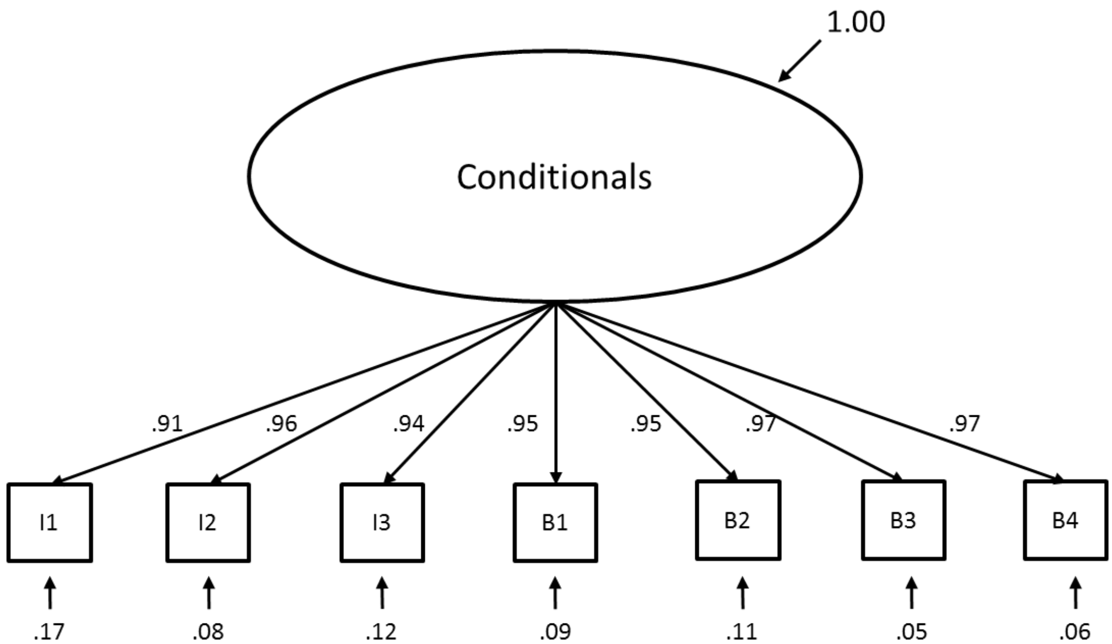
Model 3b. A 3-factor model combining the material conditional and biconditional subscales of the Propositional Logic Test into a conditionals subscale, with the three Propositional Logic Test subscales correlating with one another; one of the material conditional items (I4) has been removed from Model 3.

Figure 13.



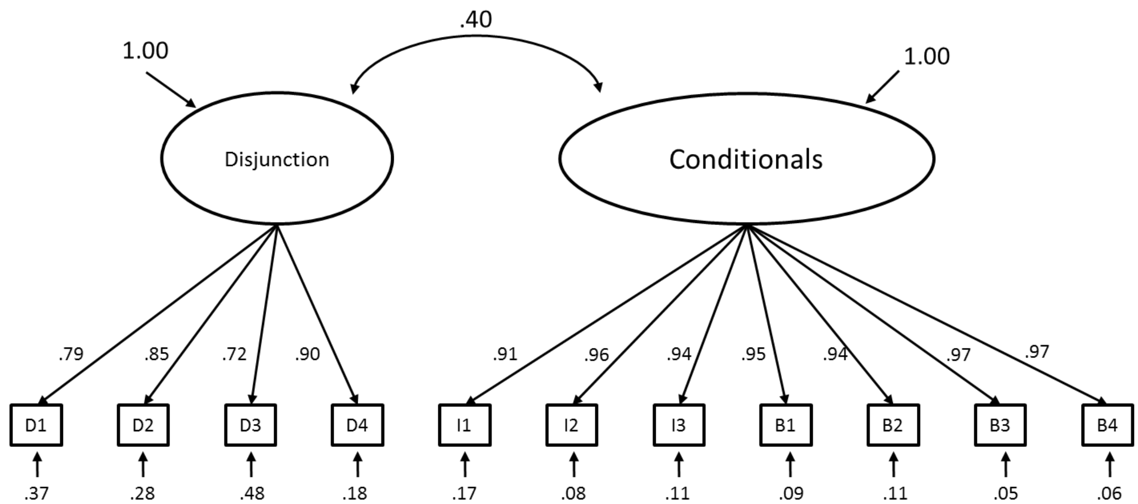
Model 8b. A two-factor model representing the material conditional and biconditional subscales of the Propositional Logic Test correlating with one another; one of the material conditional items (I4) has been removed from Model 8. Standardized results are reported.

Figure 14.



Model 9b. A one-factor model representing the combination of the material conditional and biconditional items representing a conditionals factor; one of the material conditional items (I4) has been removed from Model 9. Standardized results are reported.

Figure 15.



Model 3c. A two-factor model based on the combination of the disjunction subscale and the conditionals subscale of the Propositional Logic Test; one of the material conditional items (I4) has been removed from Model 5. Standardized results are reported.

Appendices

Appendix A

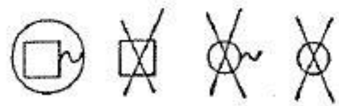
Original Paper and Pencil Version of the Propositional Logic Test

Propositional Logic Test Instructions Page

PLT   Name ________________

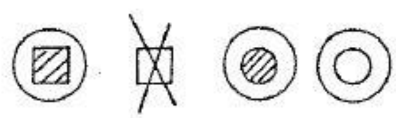
In each of the problems on the following page you will find a sentence followed by four figures. Each figure is either square or round, either large or small, either white or striped, and either tailed (has a tail) or untailed. Your task is to circle those fiqures that are allowed by the sentence and to cross out the ones that are not allowed. Here are some examples with the correct answers to show you what this means. Study them carefully since the problems that follow are very similar.

1. It is square and it is tailed.



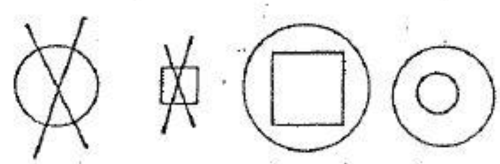
Here it says it must be square <u>and</u> tailed so only the one that is both square and tailed fits. The others are not tailed or are not square or are not both so they should be crossed out.

2. If it is white then it is round.



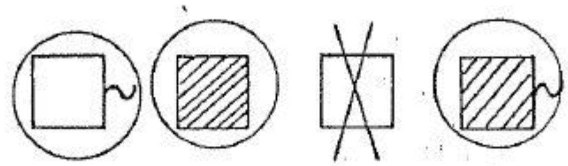
If it is white then it must be round, but if it is striped then it doesn't matter if its round or not. So the white circle fits but the white square does not. The striped figures all fit because the statement only tells us about white figures.

3. If it is round it is small and if it is small it is round.



The round ones that are small fit and so do the small ones that are round. Since the large square isn't round it doesn't have to be small, so it fits. The large circle doesn't fit the first part of the rule and the small square doesn't fit the second part.

4. It is striped or it is tailed or both.

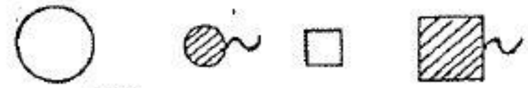


You can circle the first figure because it is tailed. The second figure also fits because it is striped, and the last one fits because it is both striped and tailed. The third figure doesn't fit since it is neither striped nor tailed.

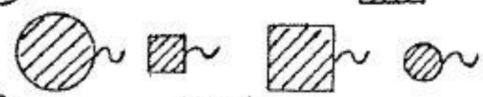
You have 15 minutes to complete this test.

Propositional Logic Test Items

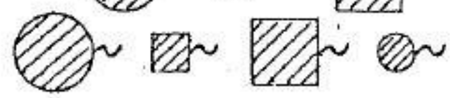
1. It is round and it is striped.



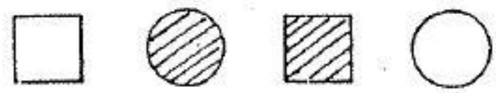
2. It is small or it is round or both.



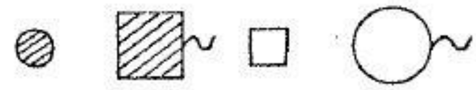
3. If it is large then it is round.



4. If it is round it is striped and if it is striped it is round.

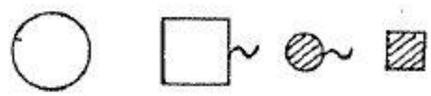


5. It is tailed or it is square or both.

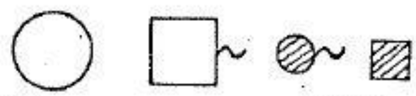


6. It is striped and it is large.

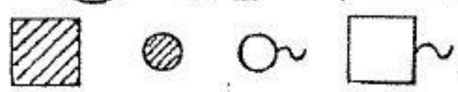
7. If it is large it is square and if it is square it is large.



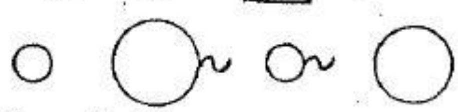
8. If it is small then it is square.



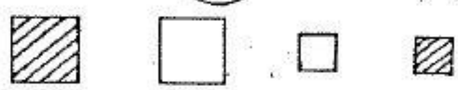
9. It is large and it is tailed.



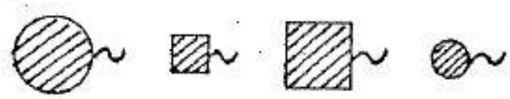
10. It is large or it is untailed or both.



11. If it is white then it is large.

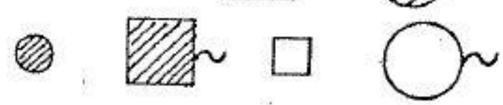


12. If it is small it is square and if it is square it is small.

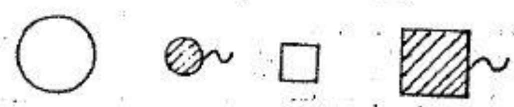


13. It is striped or it is small or both.

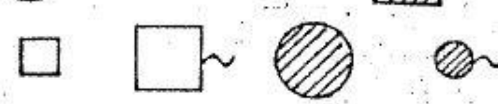
14. It is tailed and it is round.



15. If it is square it is white and if it is white it is square.



16. If it is striped then it is large.



Appendix B

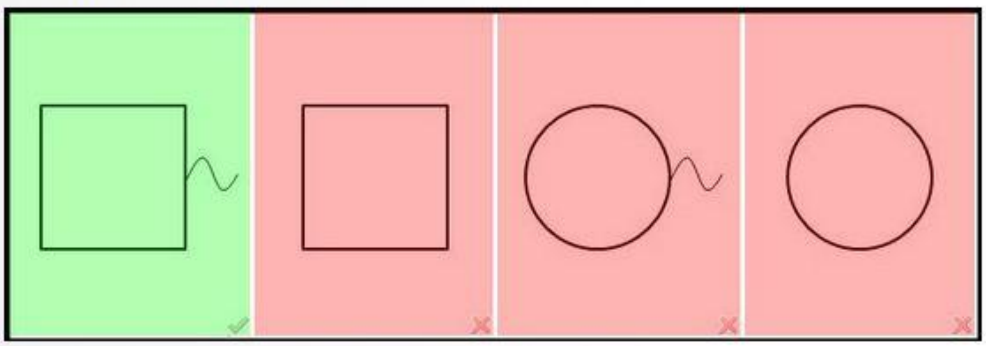
Screenshots of the Computerized Version of the Propositional Logic Test

Computerized Propositional Logic Test Instructions Page

In each of the problems on the following pages you will find a sentence followed by four figures. Each figure is either square or round, either large or small, either white or striped, and either tailed (has a tail) or untailed. Your task is to click on those figures that ***<u>are</u>*** allowed by the sentence ***<u>one time</u>*** (which will turn that area green), and to click on those figures that ***<u>are not</u>*** allowed by the sentence ***<u>two times</u>*** (which will turn that area red).

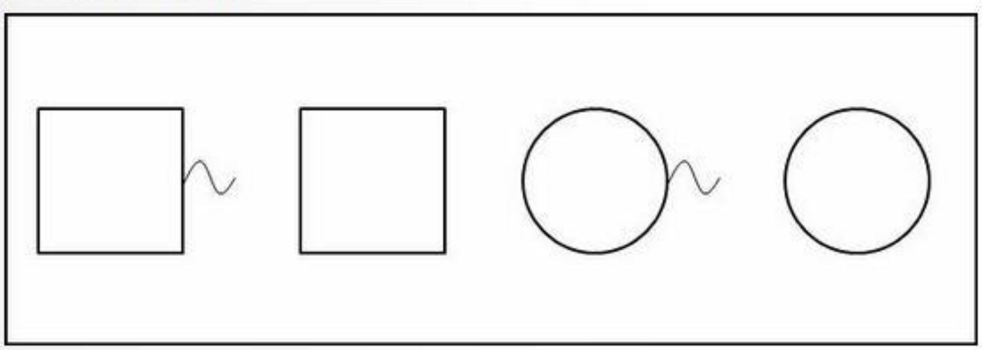
You may change your answer by clicking on the figure again - the area will cycle between green (allowed by the sentence), red (not allowed by the sentence), and unchecked. You must have selected either green (allowed by the sentence) or red (not allowed by the sentence) before you can continue; no figure may be left unmarked.

Example 1. It is square and it is tailed.



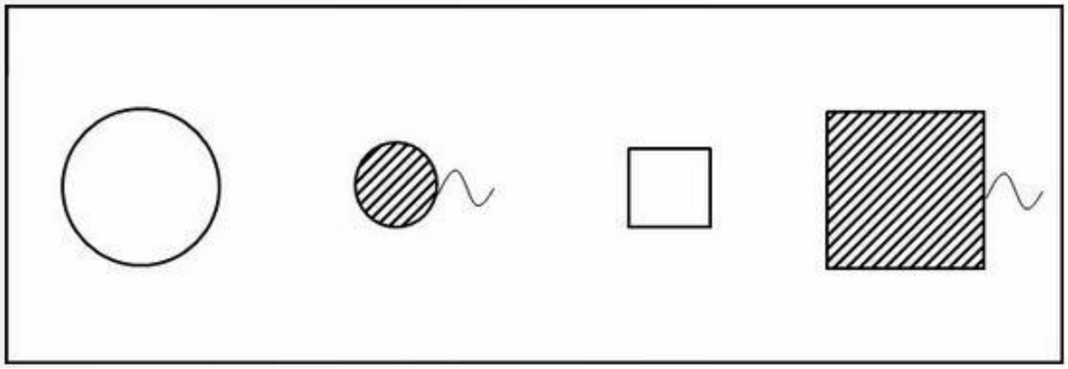
Here it says it must be square <u>and</u> tailed so only the one that is both square and tailed fits. The others are not tailed or are not square or are not both so they should be marked as red - not allowed by the sentence.

Please mark the squares below as you see them above. When finished, proceed to the next page for another example.

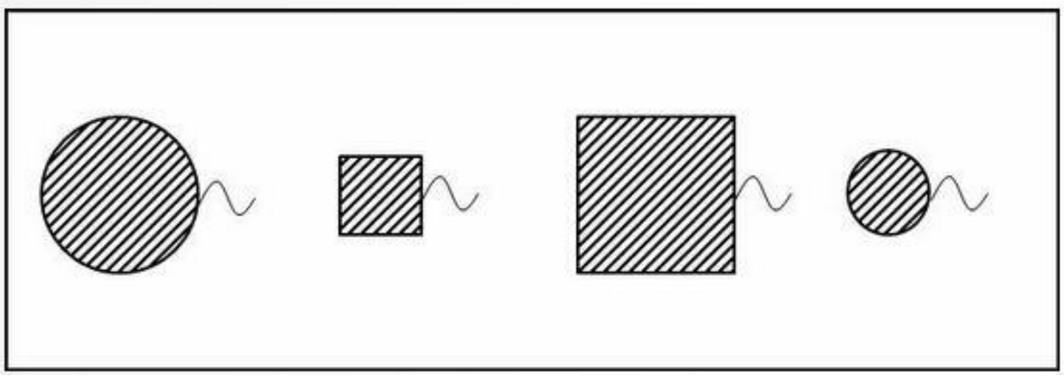


Computerized Propositional Logic Test Items

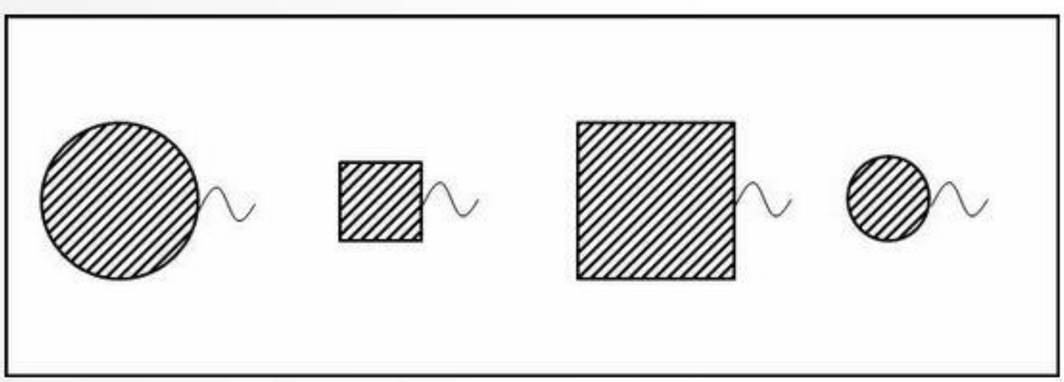
1. It is round and it is striped.



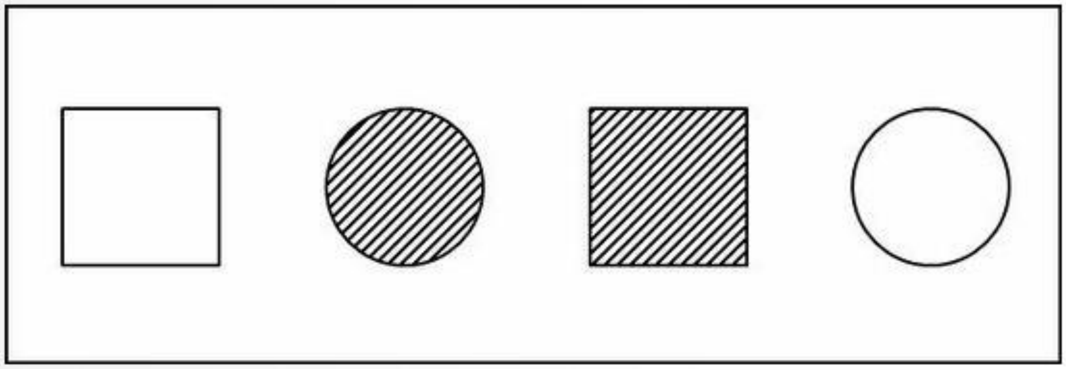
2. It is small or it is round or both.



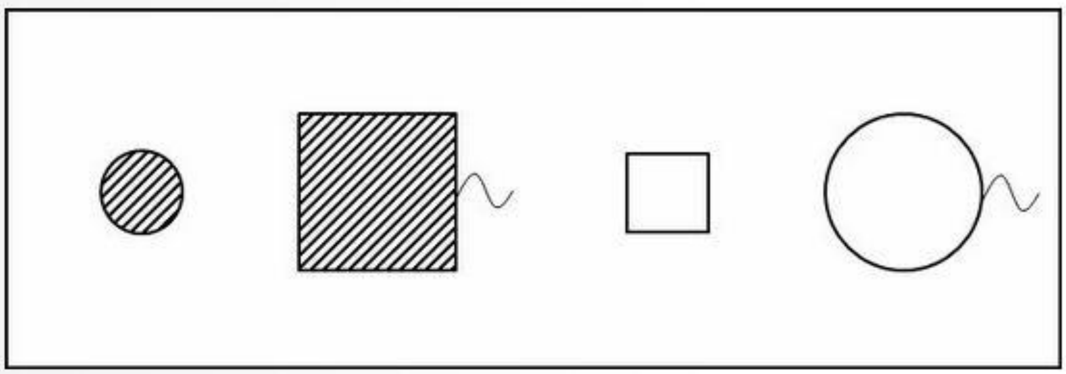
3. If it is large then it is round.



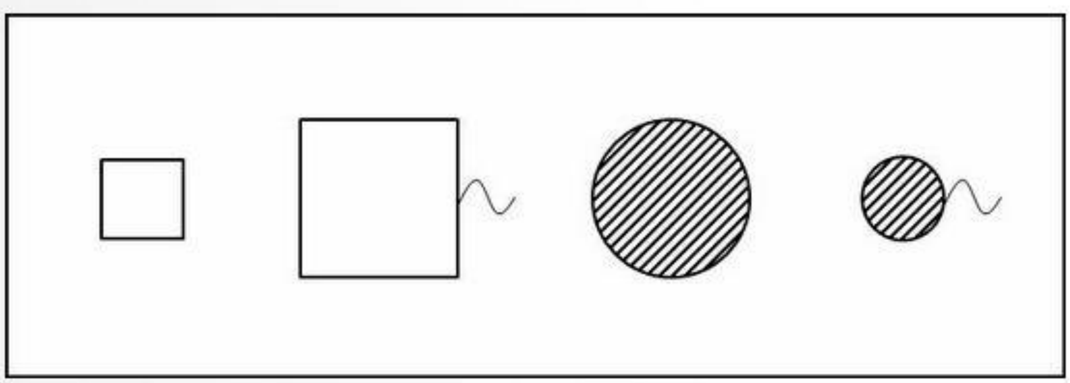
4. If it is round it is striped and if it is striped it is round.



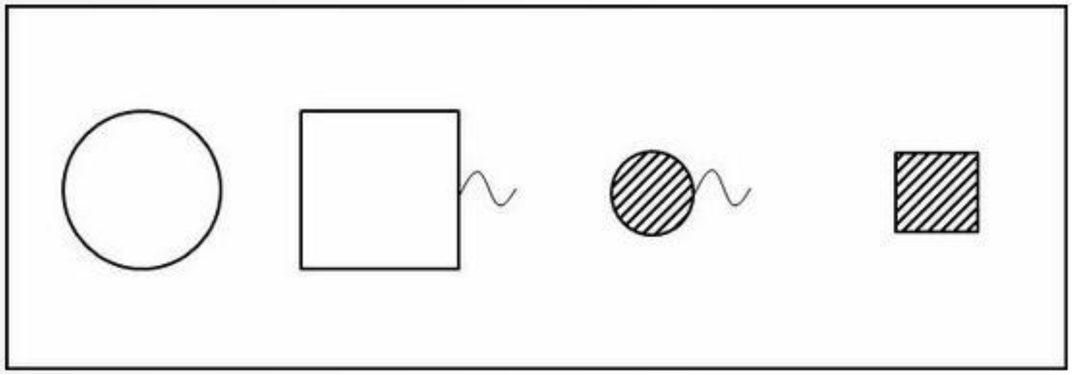
5. It is tailed or it is square or both.



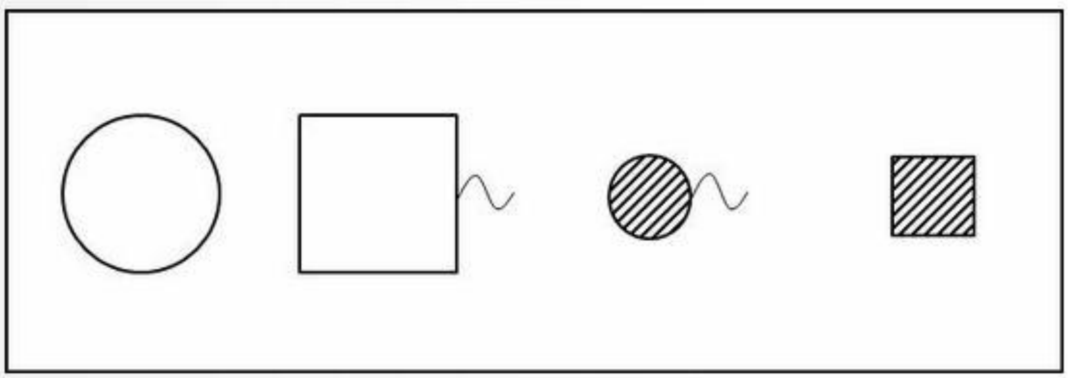
6. It is striped and it is large.



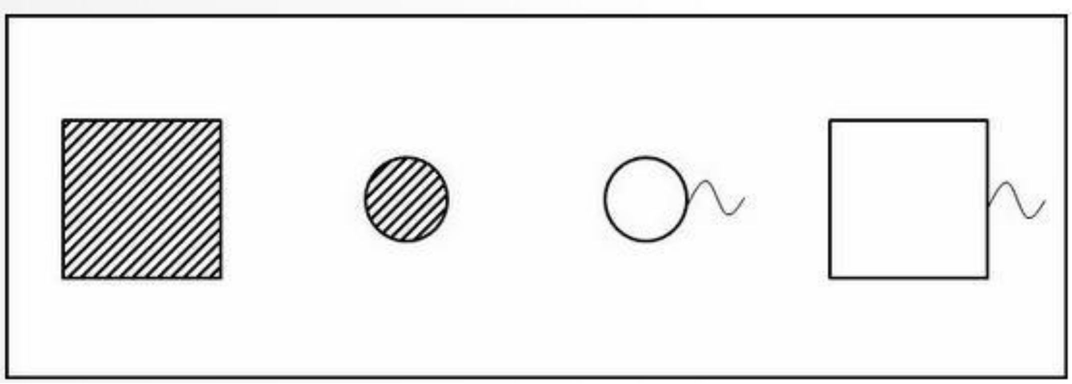
7. If it is large it is square and if it is square it is large.



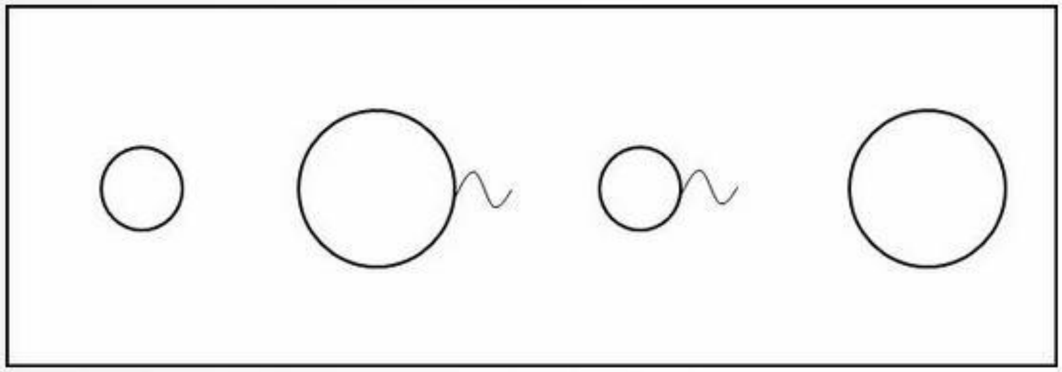
8. If it is small then it is square.



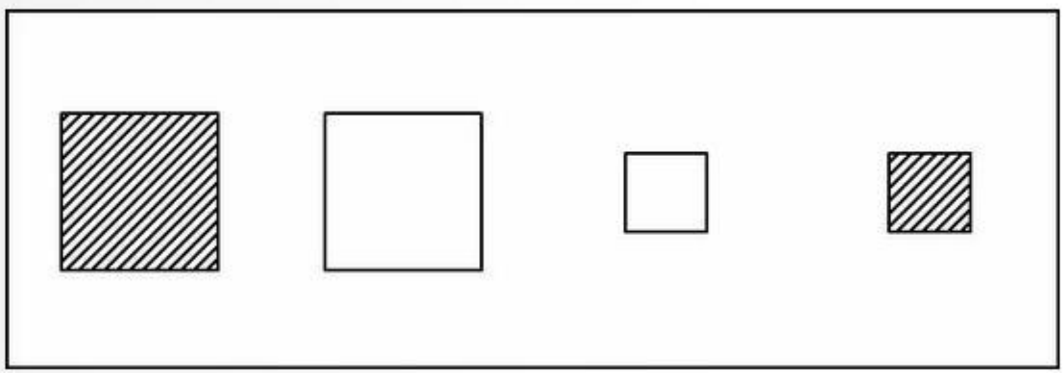
9. It is large and it is tailed.



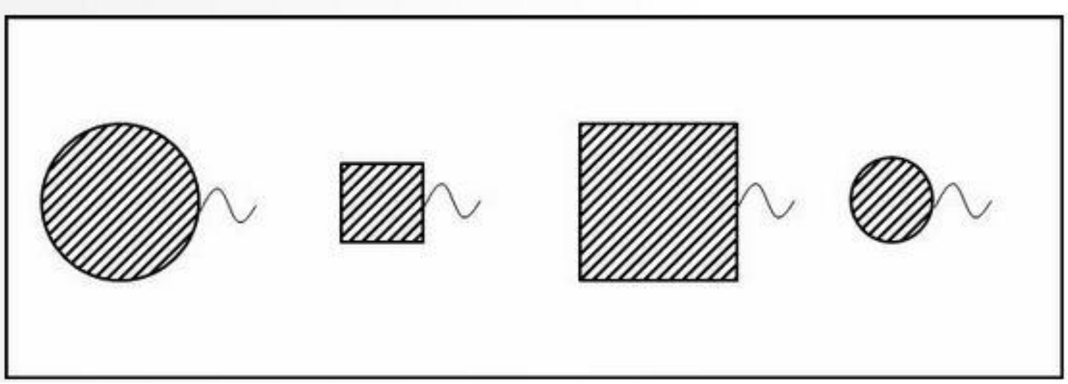
10. It is large or it is untailed or both.



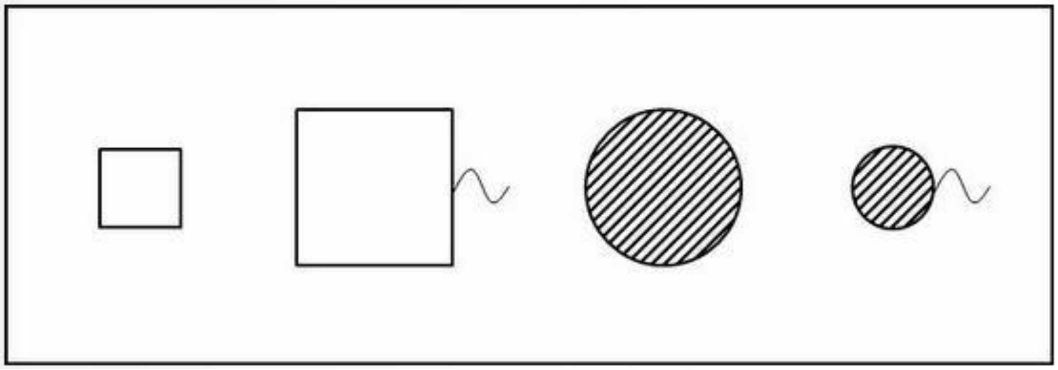
11. If it is white then it is large.



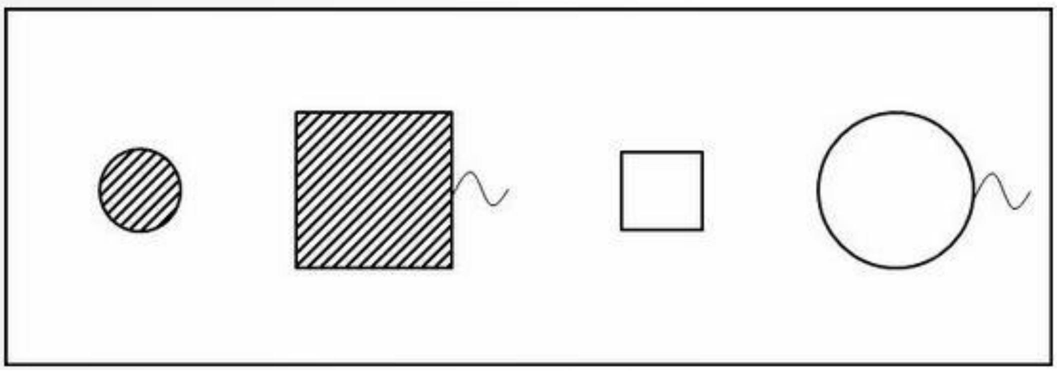
12. If it is small it is square and if it is square it is small.



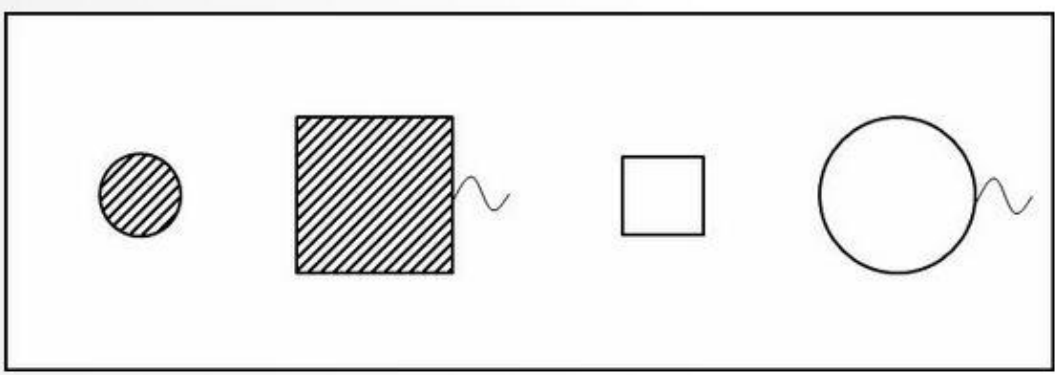
13. It is striped or it is small or both.



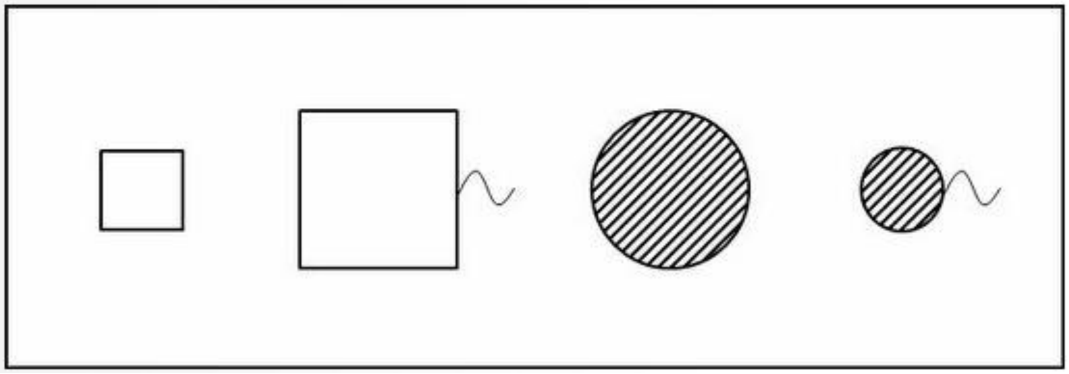
14. It is tailed and it is round.



15. If it is square it is white and if it is white it is square.



16. If it is striped then it is large.



Appendix C

Belief Bias Syllogism Items

Participants are given the following instructions:

In the following problems, you will be given two premises which you must assume are true. A conclusion from the premises then follows. You must decide whether the conclusion follows logically from the premises or not. You must suppose that the premises are all true and limit yourself only to the information contained in the premises. This is very important. Decide if the conclusion follows logically from the premises, assuming the premises are true, and select your response.

After each item the participant is given the choices "Conclusion follows logically from premises" and "Conclusion does not follow logically from premises."

1. PREMISES:
All guns are dangerous.
Rattlesnakes are dangerous.
CONCLUSION:
Rattlesnakes are guns.

2. PREMISES:
All living things need water.
Roses need water.
CONCLUSION:
Roses are living things.

3. PREMISES:
All things made of wood can be used as fuel.
Gasoline is not made of wood.
CONCLUSION:
Gasoline cannot be used as fuel.

4. PREMISES:
All African countries are hot.
Canada is not an African country.
CONCLUSION:
Canada is not hot.

5. PREMISES:
All bats have wings.
Hawks are not bats.
CONCLUSION:
Hawks do not have wings.

6. PREMISES:
All birds have feathers.
Robins are birds.
CONCLUSION:
Robins have feathers.

7. PREMISES:
All fish can swim.
Tuna are fish.
CONCLUSION:
Tuna can swim.

8. PREMISES:
All large things need oxygen.
Mice need oxygen.
CONCLUSION:
Mice are not large things.

9. PREMISES:
All mammals walk.
Whales are mammals.
CONCLUSION:
Whales walk.

10. PREMISES:
All nuts can be eaten.
Rocks cannot be eaten.
CONCLUSION:
Rocks are not nuts.

11. PREMISES:
All things that are alive drink water.
Televisions do not drink water.
CONCLUSION:
Televisions are not alive.

12. PREMISES:
All things that are smoked are good for the health.
Cigarettes are smoked.
CONCLUSION:
Cigarettes are good for the health.

13. PREMISES:
All things that have a motor need oil.
Automobiles need oil.
CONCLUSION:
Automobiles have motors.

14. PREMISES:
All things that move love water.
Cats do not love water.
CONCLUSION:
Cats do not move.

15. PREMISES:
All things with four legs are dangerous.
Poodles are not dangerous.
CONCLUSION:
Poodles do not have four legs.

16. PREMISES:
All unemployed people are poor.
Rockefeller is not unemployed.
CONCLUSION:
Rockefeller is not poor.

References

Ackerman, P.L, Beier, M.E., & Boyle, M.O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131*(1), 30-60.

Allen, C., & Hand, M. (2001). *Logic primer* (2nd Edition). Cambridge, MA: MIT Press.

Almstrum, V. (1999). The Propositional Logic Test as a diagnostic tool for misconceptions about logical operations. *Journal of Computers in Mathematics and Science Teaching, 18*(3), 205-224.

Beall, J.C. (2010). *Logic: the basics*. New York, NY: Routledge.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17*, 10 –17.

Bentler, P. M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software.

Bentler, P. M. (1989). EQS structural equations program manual. Los Angeles: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indexes in structural equation models. *Psychological Bulletin*, *107*, 238-246.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136 –162). Beverley Hills, CA: Sage Publishing.

Burgess, J.P. (2009). Philosophical logic. Princeton, NJ: Princeton University Press.

Byrne, B.M. (1998), *Structural Equation Modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, applications and programming*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Chater, N., & Oaksford, M. (2004). Rationality, rational analysis and human reasoning. In K. Manktelow & M. C. Chung (Ed.), *Reasoning: History and philosophy*. Hove, UK: Psychology Press.

Conway, A.R.A., Kane, M.J., & Engle, R.W. (2003). Working memory capacity and its relation to general intelligence. *TRENDS in Cognitive Sciences, 7*(12), 547-552.

De Neys, W., d'Ydewalle, G., Schaeken, W., & Vos, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica, 42*, 177–190.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3*, 412-423.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*, 309–331.

Evans, J.St.B.T., Newstead, S.E., & Byrne, R.M.J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Erlbaum.

Evans, J.St.B.T. & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11*(4), 382-389.

Evans, J.St.B.T. & Frankish, K. (2009). *In two minds: Dual processes and beyond*. New York, NY: Oxford University Press.

Evans, J.St.B.T. & Over, D. (2004). *If*. New York, NY: Oxford University Press.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition, 11*(3), 285-306.

Finney, S. J., & DiStefano, C. (in press). Nonnormal and categorical data in structural equation models. In G.R. Hancock & R.O. Mueller (Eds.). *A second course in structural equation modeling* (2nd Edition). Greenwich, CT: Information Age.

Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the achievement goal questionnaire in a general academic context. *Educational and Psychological Measurement, 64*, 365–382.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25-42.

Garnham, A., & Oakhill, J. (1994). *Thinking and reasoning*. Cambridge, MA: Blackwell Publishing.

George, C. (1995). The endorsement of the premises: Assumption-based or belief-based reasoning. *British Journal of Psychology, 86*, 93-111.

Gerbing, D.W., & Anderson, J.C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research*, *25*(2), 186-192.

Green, S.B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155-167.

Griggs, R.A., & Cox, J.R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73*, 407–420.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Humbleton, L. (2011). *The connectives*. Cambridge, MA: The MIT Press.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. Memory, 18, 394 – 412.

Johnson-Laird, P.N. (2006). *How we reason*. New York, NY: Oxford University Press.

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 99*, 418-439.

Jöreskog, K. G., & Sörbom, D. (1981). LISREL V: Analysis of linear structural relationships by the method of maximum likelihood. Chicago: National Educational Resources.

Jöreskog, K.G. & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

Jöreskog, K. G., & Sörbom, D. (2007) LISREL 8.80: Structural equation modeling with the SIMPLIS command language. Chicago: Scientific Software International.

Kim, Y. (1995). *The reasoning ability and achievement of college level students enrolled in a logic class in computer science*. University of Texas, Austin.

King, P., & Shapiro, S. (1995). The history of logic. In T. Honderich (Ed.), *The Oxford companion to philosophy* (pp. 496-500). New York, NY: Oxford.

Kirchner, W.K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352-358.

Kleene, S. (1967). *Mathematical logic*. Hoboken, NJ: Wiley Publishing.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd Edition). New York: Guilford.

Leighton, J.P., & Sternberg, R.J. (Eds.). (2004). *The nature of reasoning*. New York, NY: Cambridge University Press.

Manktelow, K.I., & Chung, M.C. (Eds.). (2004). *Psychology of reasoning: Theoretical and historical perspectives*. New York, NY: Psychology Press.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Newstead, S.E., & Evans, J.St.B.T. (Eds.). (1995). *Perspectives on thinking and reasoning*. Hove, UK: Erlbaum.

Nye, C.D., & Drasgow, F. (2011). Assessing goodness of fit: simple rules of thumb simply do not work. *Organizational Research Methods*, 14 (3), 548-570.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. New York, NY: Oxford University Press.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, & WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, *7*, 557-595.

Olsson, U. H., Troye, S. V., & Howell, R. D. (1999). Theoretic fit and empirical fit: The performance of maximum likelihood versus generalized least squares estimation in structural equation modeling. *Multivariate Behavioral Research*, 34, 31-58.

Owens, B.B., & J. Seiler. (1996). A study of the relationship between performance in propositional logic and computer science. *Proceedings of the Seventh Annual South Central Small College Computing Conference*, *The Journal of Computing in Small Colleges*, *11*(7), 90-93.

Peirce, C.S. (1992) *Reasoning and the logic of things: The Cambridge conferences lectures of 1898*. K. L. Ketnet (Ed.). Cambridge, MA: Harvard University Press.

Perdue, D. (1992). *Debate in Tibetan Buddhism*. Ithica, NY: Snow Lion Publications.

Perdue, D. (in press). *The course in Buddhist reasoning and debate*. Boston, MA: Shambala Publications.

Piburn, M.D. (1989). Reliability and validity of the Propositional Logic Test. *Educational and Psychological Measurement, 49*, 667-672.

Piburn, M.D. (1990). Reasoning about logical propositions and success in science. *Journal of Research in Science Teaching, 27*(9), 887-900.

Priest, G. (2008). An introduction to non-classical logic: From if to is (2nd edition). New York, NY: Cambridge University Press.

Priest, G. (2010). *Logic*. New York, NY: Sterling Publishing Co., Inc.

Ripley, D. (2009). Contradictions at the boarders. In R. Nouwen, R. van Rooij, U. Sauerland, and H.-C. Schmitz (Eds.), *Vagueness in communication: Lecture notes in computer science Vol. 6517* (pp.169-188). New York, NY: Springer Publishing.

Rips, L.J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT Press.

Rogers, K.M. (2009). *Tibetan logic*. Ithica, NY: Snow Lion Publications.

Sá, W., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 91*, 497 – 510.

Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors on covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis* (pp. 399–419). Thousand Oaks, CA: Sage.

Sevenants, A., Dieussaert, K., & Schaeken, W. (2011). Truth table tasks: Irrelevance and cognitive ability. *Thinking and Reasoning, 17*(3), 213-246.

Schaeken, W., Vandierendonck, A., Schroyens, W., & d'Ydewalle, G. (2007). *The mental models theory of reasoning: Refinements and extensions*. Hove, UK: Erlbaum.

Shapiro, S. (2009, November 2). Classical Logic. E. N. Zalta (ed.). Retrieved from http://plato.stanford.edu/entries/logic-classical/

Smullyan, R.M. (1961). *Theory of formal systems*. Princeton, NJ: Princeton University Press.

Stanovich, K.E. (2010). *Decision making and rationality in the modern world*. New York: Oxford University Press.

Stanovich, K.E. (2011) *Rationality and the reflective mind*. New York: Oxford University Press.

Stanovich, K. E., & West, R. F. (1998). Cognitive ability and variation in selection task performance. *Thinking and Reasoning, 4*, 193–230.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645 – 726.

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioural Research, 25*, 173-180.

Stegner, J.H. (2009). R2 [Software]. Available from http://www.statpower.net/Software.html

Steiger J. H., & Lind J. M. (1980). *Statistically based tests for the number of factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Stenning, K. (2002). *Seeing reason*. New York, NY: Oxford University Press.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*, 1275–1289.

Wason, P.C. (1966). Reasoning. In B. M. Foss (Ed.). *New horizons in psychology*. Harmondsworth, UK: Penguin.

Wason, P.C., & Johnson-Laird, P.N. (1972). *Thinking: Readings in cognitive science*. Cambridge, UK: Cambridge University Press.

Yu, C. (2002) *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcome*. (Doctoral dissertation). University of California, Los Angeles.