INFORMAL REASONING WITH AND WITHOUT THE INTERNET: AN INDIVIDUAL DIFFERENCES APPROACH

A Dissertation
Presented to
The Academic Faculty

By

Victor J. Ellingsen

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Psychology

Georgia Institute of Technology

May, 2017

Informal Reasoning With and Without the Internet:
An Individual Differences Approach

Approved by:

Dr. Phillip L. Ackerman, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Margaret E. Beier
Department of Psychology
Rice University

Dr. James S. Roberts
School of Psychology
Georgia Institute of Technology

Dr. Rick P. Thomas
School of Psychology
Georgia Institute of Technology

Dr. Paul Verhaeghen
School of Psychology
Georgia Institute of Technology

Date Approved: March 29, 2017

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Phillip Ackerman, for his guidance on this project and over the course of my graduate school career. I also appreciate the advice provided by the other committee members during this project.

I also would like to thank the many people whose support was perhaps less tangible, but just as important. Erin Marie Conklin, Samuel Posnock, and Charles Calderwood provided invaluable insight into how to follow in their footsteps. Elnora Kelly, Laura Barg-Walkow, Sadaf Kazi, Chelsea Vance, Mike Morrison, and Ilya Gokhman were with me to celebrate the highs and commiserate on the lows that are inevitably part of graduate school. Franklin Rea, Kyle Jones, and my exceptionally patient mother were always willing to listen despite not always knowing what I was talking about or why it seemed important.

The many friends I met through Georgia Tech Hapkido and Black Knight Martial Arts were a wonderful and much-needed community outside of the lab. In particular, Dr. Brian Ritchie and Dr. Matthieu Bloch generously offered their views of grad school from the “other side,” which helped me keep things in perspective.

I am also indebted to the 2017 federal hiring freeze, which left me with nothing to do for three months except to sit in New Hampshire and finish this dissertation.

Finally, this project would not have been possible without assistance from Chelsea Vance, Andrew Helbling, Soundarya Kanthimathinathan, and Sadaf Kazi, who helped code the more than 10,000 arguments generated by participants in these two studies.

TABLE OF CONTENTS

LIST OF TABLES

SUMMARY

Informal reasoning is used when people reason about complex issues for which there is not a single, agreed-upon correct answer (Perkins, 1985a; Sadler & Zeidler, 2005). Accordingly, an individual's ability and willingness to consider arguments on both sides of an issue is a key component of successful informal reasoning. However, people typically do not explore arguments contrary to their own position unless specifically instructed to do so (Perkins, 1985a). A major limitation of previous research is that participants usually have been required to reason with no access to outside sources of information, which does not reflect the reality of reasoning in the age of the Internet. In addition, the relationships between informal reasoning and various individual-differences factors have not been explored thoroughly. In this set of studies, I used hierarchical linear modeling in order to assess both item-level and person-level predictors of informal reasoning. I also manipulated Internet access in order to assess the effect of outside information during a standard argument generation task (Toplak & Stanovich, 2003). Strength of prior opinion and exposure to the issue described in the prompt emerged as the most salient predictors of reasoning performance, and access to outside information via the Internet increased the number of otherside arguments generated in the task. Like many previous research efforts, these studies failed to identify robust person-level predictors of informal reasoning performance. However, the non-ability traits of typical intellectual engagement and anti-intellectualism predicted a greater amount of time spent on reasoning items, which in turn predicted reasoning outcomes. Investigating person-

and situation-level predictors of the decision to stop searching for additional arguments may be a fruitful direction for future research.

CHAPTER 1
INTRODUCTION

Informal reasoning is a process in which a person generates and evaluates evidence with respect to a truth claim (Means & Voss, 1996; Perkins, 1985a). This process is used in response to *ill-structured* problems: “genuinely vexed” (Perkins, 1985a), complex issues for which there is no clear solution and for which reasonable arguments exist on both sides of the issue (Means & Voss, 1996; Sadler & Zeidler, 2005). In such problems, not all the information necessary to solve the problem is provided. Rather, the problem-solver must search for the relevant information, usually in his or her memory (Galotti, 1989). In addition, ill-structured problems involve some uncertainly about the criteria for evaluating solutions, which means that such solutions generally must be defended by the problem-solver rather than be compared to a single correct answer (Jonassen, 1997). Although most problems encountered by adults in everyday life require informal reasoning (Evans & Thompson, 2004; Galotti, 1989), a relatively small proportion of research on human reasoning has examined this type of reasoning

Instead, most of the problem-solving and reasoning literature to date has focused on formal reasoning. Formal reasoning involves applying a set of rules or a series of steps to a problem for which all relevant information is provided at the outset (Galotti, 1989). This type of reasoning is often studied using deductive reasoning tasks, such as logical syllogisms, and statistical reasoning tasks (Evans & Thompson, 2004). These problems are considered to be *well-structured*: they have a single correct solution, preferred processes for obtaining that solution, and they include all necessary information for reaching the solution in the problem presentation (Jonassen, 1997). Well-structured problems are often encountered in school settings or on cognitive ability tests, but are relatively uncommon in real life (Jonassen, 1997).

High-quality informal reasoning is different from high-quality formal reasoning. Whereas good formal reasoning often can be identified by observing whether an examinee arrived at the correct answer, good informal reasoning is somewhat more nebulous and difficult to measure (see, for example, the various reasoning outcomes employed by Furlong, 1993; Macpherson & Stanovich, 2007; Perkins, 1985; Sadler & Zeidler, 2005; and Wolfe, 2012). Perkins (1985b) identified several characteristics of informal reasoning. Primary among them were the requirements that in order to demonstrate good informal reasoning, a person must consider both sides of an issue, and must include multiple lines of reasoning (i.e., many different arguments that address different aspects of the problem). Multiple lines of reasoning are required for informal reasoning situations because no single line of reasoning can definitively resolve the issue. For example, in response to a question about legalizing recreational marijuana use, a given argument may address health concerns, but other arguments must be used in order to address the proper role of the state in placing limits on personal freedom. In multifaceted issues such as this one, no one facet provides a clear answer to the question. Instead, a reasoner must address several different angles as he or she builds a case in favor or in opposition to the issue.

Perkins described informal reasoning as a type of situation modeling, in which a reasoner must build a mental model of the situation presented in the reasoning prompt, and explore the resulting problem space (Perkins, Farady, & Bushey, 1991). Poor informal reasoning indicates a failure to build or explore complete situation models, resulting in biased and incomplete responses to informal reasoning items. This can be contrasted with formal reasoning, for which one-sidedness is not problematic, as in the development of a proof (Perkins et al., 1991). In accord with the different mental processes involved in the two types of reasoning, there is some evidence (Shin, Jonassen, & McGee, 2003) that different abilities are important for solving ill-structured problems (associated with informal reasoning) and well-structured problems (associated with formal reasoning). This disconnect between the types of problems used in school and assessment settings on the one hand, and the types of problems encountered in everyday life on the other hand, is at the core of many criticisms of intelligence tests, especially for adults (Ackerman, 2000; Sternberg, 1984).

The domain of informal reasoning has captured the interest of many researchers seeking to understand the processes and predictors of reasoning performance in the real world. The question of how people make complex decisions has implications for fields ranging from consumer behavior (Kerstetter & Cho, 2004), to politics (Voss, 1991), to medicine (Christensen & Elstein, 1991), to law and justice (Lawrence, 1991). However, informal reasoning presents substantial challenges for researchers who wish to study the topic scientifically. First, informal reasoning performance is considerably more difficult to assess than formal reasoning performance, due to the fact that multiple solutions and paths to those solutions exist; researchers are faced with the problem of rating the quality of a wide variety of open-ended responses, rather than determining whether one single correct solution was reached. Second, it is difficult to empirically separate informal reasoning from the related construct of argumentation[1] (Sadler & Zeidler, 2005), though some researchers have tried (e.g., Wolfe & Britt, 2008).

Considering that informal reasoning arguably represents the majority of real-world reasoning carried out by adults, the first question that arises is: What is the quality of reasoning that can be expected from a typical adult? Overall, results have been surprisingly disappointing, even among college students and graduates (Baron, 1995; Furlong, 1993; Macpherson & Stanovich, 2007; Nussbaum & Kardash, 2005; Perkins, 1985a; Perkins et al., 1991; Sadler & Zeidler, 2005; Wolfe, 2012; Wolfe & Britt, 2008). Researchers report substantial bias (i.e., generating more arguments that support one's position on an issue than oppose it) and a general failure to thoroughly explore the many lines of reasoning relevant to the issue in question (Perkins, 1985a). This is particularly worrisome given that many educators emphasize critical thinking and informal reasoning as important skills for citizens of the 21st century (Sadler, 2004).

Given that the typical level of informal reasoning is not optimal, the next question that arises is: What individual factors are predictive of good or poor informal reasoning? Several individual-level variables have been assessed in the literature, including general ability (Klaczynski & Robinson, 2000; Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003), prior knowledge (Furlong, 1993; Sadler & Zeidler, 2005), need for

[1] Sadler and Zeidler (2005) noted that informal reasoning refers to the cognitive and affective processes that are required for engaging with complex issues, while argumentation is generally understood to be a learned skill that is used to express the results of informal reasoning. As a result, in most cases, it is only possible to measure informal reasoning by way of argumentation. However, Sadler and Zeidler argued that although good argumentation must reflect good informal reasoning, weak argumentation does not necessarily indicate weak informal reasoning. One major goal of the interview-based research protocols they use is to circumvent argumentation skill by prompting participants in a more informal, adaptive way.

cognition (Furlong, 1993; Macpherson & Stanovich, 2007), prior involvement in the issue (Furlong, 1993), and prior opinion (Wolfe & Britt, 2008). Results have been inconsistent, likely due in part to the variety of informal reasoning tasks used (see next section), and in part to study design and analysis approaches that have at times been underpowered or relatively unsophisticated. These results will be reviewed in detail in a later section.

Overall, two things seem clear. First, informal reasoning is an important element of intellectual activity in adults. Second, little is known about the ability and non-ability trait correlates of informal reasoning. As a result, psychologists are extremely limited in their ability to predict a substantial component of everyday adult intellectual functioning. In the following sections, I will review the literature on informal reasoning, including the various measurement approaches, the effect of task instructions on informal reasoning performance, and the state of knowledge regarding individual differences correlates of informal reasoning. I also will identify some limitations in the current body of literature, which the studies in this project were designed to address.

Measurement Approaches

Prompt types. Different researchers have used widely different informal reasoning prompts. The key features common to all informal reasoning prompts is that they must reflect ill-structured, unresolved issues about which reasonable people could disagree (King & Kitchener, 2002), and for which multiple lines of argument could be generated for both sides (Perkins, 1985b). Within these very general constraints, prompt topics have varied substantially in terms of content domain, familiarity to participants, and the degree to which the issue is a “hot-button” issue in society at large.[2]

Perkins (1985a), who was interested in the general processes of reasoning separate from domain knowledge, selected prompts that had current relevance at the time the data were collected, and that were accessible to a wide range of people regardless of domain knowledge. He asked 320 participants, ranging from high school students to working adults, questions such as whether a bottle bill would help to reduce litter, or whether restoring the military draft would make the United States more influential in world affairs. Performance was measured using variables including the number of lines of argument generated, the number of objections the participant raised to his or her own position, and how well the participant responded to follow-up questions from the interviewer. In contrast, Sadler and Zeidler (2005) were interested specifically in the impact of domain knowledge on college students’ informal reasoning about socioscientific issues (issues such as cloning and genetic manipulation, which include both scientific and societal/moral considerations; Sadler & Zeidler, 2005). They selected issues that allowed knowledgeable participants (*N* = 15 undergraduates selected for high domain knowledge) to draw heavily on their domain knowledge, to the degree that the prompts were presented along with basic background information so that a low-knowledge group of participants (*N* = 15) could understand the questions. A representative item from this program of research is: “If science found a single gene that produced nearsightedness, should gene therapy be used to eliminate that gene from sex

[2] An even wider variety of topics has been employed in studies that have used tasks such as argument evaluation (e.g., Brem & Rips, 2000) and identification of reasoning fallacies (e.g., Ricco, 2007), which are not included in this review.

cells (egg cells or sperm cells) that will be used to create human offspring?" (Sadler & Zeidler, 2005, p. 91). Good informal reasoning was operationalized as avoiding any of a number of informal reasoning fallacies.

Another research program (Christenson, Chang Rundgren, & Höglund, 2012; Christenson, Chang Rundgren, & Zeidler, 2014) also assessed socioscientific issues, but focused specifically on the types of reasons that people give to support their arguments (e.g., from scientific, economic, sociological, or moral domains). Accordingly, their questions were particularly multidimensional, in order to elicit reasons from a variety of domains. A sample item from this program of research asked whether global warming is due to natural processes or to human activities; other prompts asked whether genetically modified organisms should be produced and sold, and whether Sweden (where the study was conducted) should invest in nuclear energy. Christenson and colleagues have carried out their studies using samples of Swedish high school students (*N* = 208 in Christenson et al., 2014; *N* = 80 in Christenson et al., 2012; in both samples, students were 18-19 years old).

Finally, some researchers have used prompts that are intended to be specifically relevant to college student samples. For example, Macpherson and Stanovich (2007) asked 195 undergraduates to generate arguments about taxpayer subsidies of tuition at publicly funded universities and about file sharing on the Internet. Wolfe and Britt (2008; see also Wolfe, 2012) asked 84 undergraduates to write essays about a fictional proposal to institute a rigorous math requirement at their university. The research programs of both Stanovich (e.g., Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003) and Wolfe (Wolfe, 2012; Wolfe & Britt, 2008) have focused on bias in informal reasoning. In particular, myside bias (Perkins, 1985a; see also Baron, 1995; Stanovich & West, 2007) refers to the tendency to consider or generate arguments that support one's stance on an issue, and to ignore arguments that oppose it (i.e., otherside arguments). Therefore, these researchers have selected issue prompts that are likely to elicit relatively strong opinions from participants in their undergraduate samples.

In summary, reasoning prompts have varied substantially across studies, depending on the focus of individual research programs. All of the studies just described have employed genuinely open questions for which multiple arguments could be generated in support of each side. However, investigators have differed in the degree to which they have selected items intended to rely more (e.g., Sadler & Zeidler, 2005) or less (e.g., Perkins, 1985a) on prior knowledge. In addition, most studies have used only a few reasoning prompts (typically two to four, although Sadler & Zeidler, 2005, used six), which makes it impossible to determine whether informal reasoning is trait-like or idiosyncratic depending on the particular issue being considered. As a result, it is difficult to assess the degree to which various individual differences factors (e.g., general cognitive ability, personality traits, or domain knowledge) influence reasoning. This limitation will be discussed further in the Individual Differences in Informal Reasoning section below.

Tasks. Informal reasoning has been investigated using a variety of methods, which can be distilled into three general categories: interviews, essays, and argument generation.[3] It is important to note that instructions play a vital role in each of these three

[3] Some researchers have also attempted to assess informal reasoning by using experiment evaluation or argument evaluation tasks (e.g., Brem & Rips, 2000; Stanovich & West,

types of tasks, and in fact are often the source of various manipulations in informal reasoning studies. Because the effect of instructions is relatively consistent across task types, the three tasks will be described first, and the effect of instructions will be discussed in the following section.

Interviews. One common approach to studying informal reasoning is to conduct one-on-one interviews with participants, with varying degrees of structure or prompting from the interviewer. Perkins (e.g., Perkins et al., 1991) described an extensive interview-based research program in which a participant is presented with an open-ended problem, given a few minutes to consider the issue, and then asked to state his/her conclusion and the reasoning that led to it. The interviewer then provides metacognitive scaffolding (e.g., prompting the participant to look for reasons on both sides of the issue, asking how the participant would rebut a counterargument proposed by the interviewer, or inquiring about how a reason supports a claim) intended to help the participant exercise his or her full reasoning powers (see also Furlong, 1993). Similarly, Sadler (e.g., Sadler & Donnelly, 2006; Sadler & Zeidler, 2005) used interviews in an explicit attempt to assess informal reasoning separately from argumentation skill. In his research program, interviewers used a set list of follow-up questions in an attempt to identify the quality of reasoning that leads participants to their conclusions. Sadler (Sadler & Donnelly, 2006; Sadler & Zeidler, 2005) typically used an extreme-groups design, selecting undergraduate

1997), statistical reasoning tasks (e.g., Klaczynski & Robinson, 2000), and fallacy identification tasks (Ricco, 2007; Weinstock, Neuman, & Glassner, 2006). Although these tasks do seem to tap cognitive components of informal reasoning, they are also relatively impoverished indicators that lack the complexity of genuine informal reasoning. As such, they are of only tangential interest to the current project and will not be considered further.

participants for the interview protocol who score very high or very low on tests of prior knowledge relevant to the socioscientific reasoning prompts used in Sadler's studies.

Interview protocols generate a rich dataset that can be analyzed in many different ways. Participant responses can be scored based on the number of arguments, counter-arguments, rebuttals, and/or supporting reasons for each (Furlong, 1993; Perkins, 1985a), the degree of bias (usually the number of arguments generated that are in favor of a participant's position, minus the number of arguments that are opposed to the participant's position; Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003; Wolfe, 2012), overall reasoning quality (Furlong, 1993; Nussbaum & Kardash, 2005; Perkins, 1985a), existence of flawed reasoning (Sadler & Zeidler, 2005), or the type of arguments offered (such as arguments based on personal experience or on relevant domain knowledge; Chang Rundgren & Rundgren, 2010).

The major benefit of interview studies is that they allow for a more complete probing of participants' thinking than is possible in more constrained methods, such as written essays and argument generation tasks (described below). In addition, as Sadler and Zeidler (2005) argued, interviews can allow researchers to tease apart informal reasoning and argumentation skill. However, interview protocols have been criticized on the grounds that follow-up questions may lead participants to consider positions and approaches that they would not have considered otherwise, resulting in an overestimation of the degree to which people successfully represent and spontaneously explore the problem space (Hofer, 2004).

A major drawback of interview protocols is that they are immensely resource-intensive. The sessions must be conducted one-on-one, and then must be transcribed before being coded and scored. As a result, many interview studies involve relatively small samples, given the number of groups (e.g., 10-15 cases per group; Means & Voss, 1996; Sadler & Zeidler, 2005) and some employ extreme groups designs (e.g., Sadler & Zeidler, 2005). Both of these features present problems for exploring the influence of individual differences, above and beyond the problems that these designs pose for an experimental context (Cronbach, 1957).

Essays. Some researchers (Nussbaum & Kardash, 2005; Wolfe, 2012; Wolfe & Britt, 2008) have opted to have participants write essays instead of engage in one-on-one interviews. Like interviews, essays can be scored using a variety of metrics, such as the number of arguments and counterarguments, bias, and overall quality (Nussbaum & Kardash, 2005; Wolfe & Britt, 2008). Research studies involving essay writing have the benefit of using a task with which many participants are generally familiar, because essay writing is a common task in high school and college courses. However, essay writing protocols are not pure measures of informal reasoning. As noted above, essays confound reasoning with several other factors, including verbal ability, argumentation skill (Sadler & Zeidler, 2005), and participants' ideas of what a good essay entails (for example, whether discussing counterarguments weakens one's case; Baron, 1991; Wolfe & Britt, 2008). Because adaptive, interview-style prompts would be difficult to implement in a written format, determining the source of bias or the reason for an (apparently) incomplete exploration of the problem space is more difficult with essay protocols than with interviews. No studies have directly compared results from essay and interview formats, although Perkins et al. (1991) claimed not to have found substantial differences between interview and essay formats across several studies.[4]

One approach to circumventing the non-adaptive nature of the essay method is to analyze participants' information search behavior and notes, in addition to their final essays, in attempt to gain a more complete understanding of participants' reasoning processes (Wolfe & Britt, 2008). Using this method with a sample of 84 undergraduates, Wolfe and Britt attempted to locate the point at which bias is introduced into the process. They found that nearly all participants accessed arguments from both sides of an issue in their initial search for information, but that bias appeared in the notes the participants took and in the final essays they wrote. Thus, participants apparently were not biased in the sources that they examined, but they selected information from those sources in a biased way. However, it is important to note that Wolfe and Britt presented participants with a set of arguments supporting both sides of the issue. As a result, it seems feasible that participants felt compelled to examine all of the arguments presented to them. A weaker situation—such as one in which participants have access to more information and sources than they could possibly consult in the time allotted—may not result in such complete exploration of the issue. To date, no experiments have examined informal reasoning in such a context. Additionally, no experiments, including those by Wolfe, have manipulated access to outside sources in order to examine the effect of such outside resources on informal reasoning.

[4] Perkins et al. (1991) made this statement in a review chapter and did not provide data to support it. They also did not provide in-depth descriptions of the samples involved in the research program they reviewed.

Argument generation. A third approach is an argument generation task, used primarily by Stanovich and his colleagues (Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003). In this task, participants are presented with a prompt and asked to list as many relevant arguments as they can. Participants' pre-existing positions on the issue are also assessed (via items embedded in a questionnaire completed before the reasoning task). Generated arguments are then scored as being "myside" or "otherside" based on each participant's response to the prior opinion item. Myside bias is calculated by subtracting the number of otherside arguments from the number of myside arguments, and this difference score is analyzed in relation to experimental manipulations or to other variables.

Perhaps the main draw of the argument generation approach is that it is relatively simple to score. The researcher need only score each of the arguments as being pro or con for the issue, and then translate these classifications into "myside" or "otherside" based on the participant's response to the previously presented opinion item. In addition, asking participants simply to generate arguments as opposed to defending a position avoids a problem encountered with essays, namely that some participants may erroneously believe that presenting otherside arguments weakens one's case (Baron, 1991). However, generating relevant arguments is only one component of informal reasoning (Means & Voss, 1996); the argument generation task gives no indication of how well a participant would integrate the various arguments in order to present a case supporting his or her side of the issue. One could argue that generating a set of arguments free of myside bias is a prerequisite for using those arguments to build a case or defend a position without displaying myside bias. However, performance on this task was only modestly correlated

(r = .25) with performance on an essay task in a sample of 84 undergraduates (Wolfe, 2012), indicating that the two tasks certainly are not interchangeable.

Summary of task types. The advantages of each of the three major methods for measuring informal reasoning must be weighed against their disadvantages. Although the argument generation task is attractive for its relative simplicity, it does not capture all of the complexities of informal reasoning. The interview and essay formats are much more difficult to score, but also provide a richer picture of participants' reasoning processes, especially when paired with follow-up prompts or analyses of notes and search behavior. Of note, the vast majority of informal reasoning studies have not allowed participants to access outside information. (Wolfe & Britt, 2008, and Wolfe, 2012, are exceptions, but they did not manipulate information access.) Instead, participants have been limited to the relevant knowledge that they already possess. This method dates back to the 1980s (e.g., Perkins, 1985a), and thus pre-dates the Internet. However, given the ubiquity of the Internet today, the situation used in most informal reasoning studies is rather artificial. In many instances of real-world informal reasoning, a person faced with an informal reasoning problem would be able to seek out additional information before making a decision on the issue. (See Clark & Chalmers, 1998, for an extended discussion of the related notion of the extended mind.) By ignoring this reality, researchers have examined informal reasoning under only one set of conditions, and certainly not the only one that occurs in the real world. This limitation will be discussed in greater depth in a later section.

Instructions

For any of the three tasks described in the previous section, the associated instructions exert considerable influence over participant performance. Indeed, several studies (e.g., Macpherson & Stanovich, 2007; Nussbaum & Kardash, 2005; Wolfe & Britt, 2008) have been dedicated to determining the effect of directive vs. non-directive instructions on various reasoning outcomes.[5] In general, directive instructions (or directive follow-up questions, in the case of interviews; Furlong, 1993; Perkins, 1985a) encourage participants to consider both sides of an issue and often explicitly ask for counterarguments, while non-directive instructions simply ask participants to write or say as much as they can (Furlong, 1993; Macpherson & Stanovich, 2007; Wolfe & Britt, 2008).

Consistently across studies, participants generate very few otherside arguments under non-directive instructional conditions (Furlong, 1993; Macpherson & Stanovich, 2007; Nussbaum & Kardash, 2005; Perkins, 1985a; Wolfe & Britt, 2008), resulting in substantial myside bias. Directive instructions reduce this bias by increasing the number of otherside arguments offered. The impact is substantial: Macpherson and Stanovich (2007) reported a very large effect ($d = 1.09$) of instruction type on myside bias in their argument generation task with a sample of 195 undergraduates. Nussbaum and Kardash (2005) reported an effect size of $d = .78$ for the number of counterarguments generated after a similar instructional manipulation for an essay-writing task in their sample of 107 undergraduates.

[5] Different investigators have used different terms to describe the two instruction types, but the actual content of the instructions have been remarkably similar across studies. For clarity, I will use “directive” and “non-directive” throughout.

Clearly, participants are capable of considering both sides of a complex issue, but do not do so (or do not verbalize it) spontaneously. One explanation is classic myside bias—people are simply unwilling to seek out or entertain positions inconsistent with their own, unless specifically instructed to do so (Perkins, 1985a; Wolfe, 2012). Alternatively, Baron (1995; see also Wolfe, 2012) has suggested that myside bias is partly due to faulty beliefs about what constitutes a good argument or good thinking. Some people, Baron proposed, believe that including counterarguments weakens one's case and makes one appear less confident in one's knowledge. This leads such people to leave otherside arguments out of their essays or interview responses. Another explanation, offered by Perkins and colleagues (1991), is that the gap between spontaneous and directed reasoning performance is caused by metacognitive deficits. That is, participants are able to reason better when provided with scaffolding (by the interviewer), but apparently are not able to "scaffold themselves" (Perkins et al., 1991, p. 97). Regardless of the precise cause of the differences in performance under different instructional conditions, it is clear that people approach the task very differently depending on what exactly they have been asked to do.

Interim Summary and Theoretical Framework

Biased and incomplete informal reasoning among adults is a highly consistent finding in the literature. Equally consistent is the observation that people are capable of reasoning better when they are given specific instructions as to what good reasoning entails. A framework for conceptualizing this difference from a measurement perspective is the distinction between maximal and typical performance (Cronbach, 1990; Stanovich & West, 2008). Measures of maximal performance seek to elicit the best possible performance from examinees, in part by giving clear instructions as to what good performance entails. Such measures are designed to determine how well a person *can* perform; performance is limited by the examinee's ability to perform the task, at least among sufficiently motivated examinees. In contrast, measures of typical performance are intended to assess the behavior that a person is *most likely* to exhibit, in the absence of a strong or highly constrained situation. Typical performance may be best predicted by non-ability traits that influence the amount of effort a person is *generally willing* to put forth, in the absence of strong incentive to do so.

Although the directive instructions used in informal reasoning studies are unlikely to induce a high-pressure maximal performance situation akin to the SAT or GRE, they do exert an effect on informal reasoning performance, particularly with regard to the number of otherside arguments generated. Perkins and colleagues (1991) argued that most people reason according to a "makes-sense epistemology" (Perkins et al., 1991, p. 99), in which people think about complex issues only to the degree necessary in order to create a situation model that is internally consistent and "makes sense." As Perkins and colleagues point out, this epistemology is adequate for many real-life scenarios, and is much less effortful than the ideal "critical epistemology" (Perkins et al., 1991, p. 99), which requires the careful consideration of alternative accounts. (See also Gigerenzer & Goldstein, 1996, for the similar concept of "fast and frugal" reasoning.) It is possible that the generally poor reasoning quality observed in research participants is due to a default adherence to the quick-and-easy makes-sense epistemology, unless they are told to do otherwise. In this respect, the non-directive instructions may elicit what amounts to typical behavior from participants, while the directive instructions yield something closer to maximal (or at least better-than-typical) behavior.

Because the general goal of informal reasoning research is to better understand how people reason in *everyday* situations, such as when voting (Perkins, 1985a), the discrepancy between typical and maximal levels of reasoning is particularly troubling. In the everyday context, as opposed to in an educational or assessment context, the question is not who *can* reason well when faced with an ill-structured problem. Rather, the relevant question is who *will* do so when faced with such a situation in real life, at least for situations that have some degree of personal relevance or importance for the reasoner.[6] As discussed in the next section, researchers have been largely unsuccessful in identifying individual-level predictors of reasoning quality. This may be due in part to a failure to consider the fact that the predictors of typical, unconstrained behavior often are different from the predictors of behavior under more constrained situations.

To summarize, the informal reasoning literature has two substantial shortcomings related to individual differences questions. First, researchers have generally presented only a few informal reasoning prompts (often in the same topic domain; Sadler & Zeidler, 2005; Zeidler & Schafer, 1984). This has made it difficult to assess the effect of content knowledge, beliefs, and other domain-specific influences that vary within individuals. Second, the majority of tasks used to study informal reasoning have been artificial in that they require participants to use only the knowledge and information that

[6] A limitation of informal reasoning research is that the experimental situation itself likely represents a more constrained and demanding scenario than many real-world reasoning situations. As a result, reasoning performance observed in the laboratory, even under non-directive instructional conditions, may overestimate the quality of reasoning that occurs in many real-world, low-stakes situations.

they already possess (Nussbaum & Kardash, 2005; Perkins, 1985a; Sadler & Zeidler, 2005; Zeidler & Schafer, 1984). In real-life informal reasoning, people have access to a wealth of information from various sources, including the Internet. Whether or not people use this information when making real-life decisions is an open question. However, access to outside information is a more accurate representation of situations in which people make complex decisions in the real world, and is a necessary component of a more complete understanding of who reasons well.

The two studies in this dissertation were designed to address both of these shortcomings in the literature. In the next section, I will review the literature related to individual differences correlates of informal reasoning and note some possible reasons it has been largely inconclusive.

Individual Differences in Informal Reasoning

To date, there has not been a concerted effort to study individual differences correlates of informal reasoning, and the limited results in the literature have not been entirely consistent. The most commonly assessed individual differences variables are general ability, prior domain knowledge, prior beliefs, and the non-ability factors need for cognition and actively open-minded thinking. First, I will review the limited evidence for within-person consistency in reasoning performance across different item prompts. Then, I will review findings related to potential correlates of reasoning performance.

Within-person consistency of informal reasoning performance. To date, evidence for the consistency of informal reasoning performance across item prompts is relatively limited. (See Perkins & Salomon, 1989, for an extended discussion of the domain-generality versus domain-specificity of reasoning skills). Researchers have tended to employ only a few prompts in any given study—generally four or fewer (Furlong, 1993; Macpherson & Stanovich, 2007; Nussbaum & Kardash, 2005; Toplak & Stanovich, 2003; Wolfe, 2012). To the extent that researchers have compared results across items, they often have focused on overall equivalences of mean-level responses such as the number of arguments generated (e.g., Christenson et al., 2012). Relatively few researchers have reported the degree to which individuals' performance on different items is consistent. Toplak and Stanovich (2003) reported pairwise correlations between the number of myside and otherside arguments generated for three prompts, in their sample of 112 undergraduates. Most of the correlations ranged from about $r = .30$ to $r = .46$ (Toplak & Stanovich, 2003, Table 2), although notably, the correlations between the number of myside and otherside arguments generated for a given issue were non-significant ($r = .07$ to $r = .16$). Wolfe (2012) reported a correlation of $r = .51$ between the number of otherside arguments generated in response to two different reasoning prompts. Thus, prior results suggest that performance is relatively consistent across individual reasoning items. However, this has not been investigated in-depth, and the degree to which reasoning performance is consistent within people remains relatively unexplored. This question was addressed in the current research project.

Cognitive ability. General cognitive ability would seem to be a primary predictor of informal reasoning performance. However, results have been mixed and depend on the reasoning task being used. Some researchers (Means & Voss, 1996; Perkins, 1985a) have reported significant relationships between informal reasoning and cognitive ability in samples of schoolchildren and adolescents (Means & Voss, 1996) and of adolescents and adults (Perkins, 1985a). In a sample that included high school students, undergraduates, graduate students, and non-student adults (total *N* = 320), Perkins (1985a) found that IQ (as measured by the Slosson Intelligence Test; Slosson, 1981) was the dominant predictor of informal reasoning performance to emerge from a multiple regression analysis, with standardized regression weights ranging from β = .32 to β = .48 in the student sample (*N* = 240).[7] IQ was not significantly predictive for two of the six reasoning outcomes in the non-student adult sample (*N* = 80).

Meanwhile, other researchers have found small or non-significant relationships between informal reasoning and cognitive ability in college students (*r* = .08 between IQ and myside bias on an argument generation task; Macpherson & Stanovich, 2007). For studies that have used samples of college students, it is possible that there is not enough variability in cognitive ability in the samples to detect a relationship. It should be noted that similar findings have been reported for more constrained informal reasoning tasks such as argument assessment and statistical reasoning that are not being reviewed here (e.g., see Klaczynski & Gordon, 1996; Klaczynski & Robinson, 2000; Stanovich & West, 1997, 1998), but that formal reasoning tasks often correlate substantially with general cognitive ability (*r* = .41 for identifying inconsistent syllogisms; Macpherson & Stanovich, 2007).

Stanovich and West (1997, 1998) have argued that the weak or nonexistent relationship between general cognitive ability and informal reasoning performance can be explained in a levels-of-analysis context. In their view, general cognitive ability

[7] It is unclear from Perkins's (1985a) description of his analysis whether the predictors were entered in a specified order, or whether they were entered together in a single step. The other predictors were age, years of education, and self-reported amount of prior thought given to the topic.

corresponds to the algorithmic level of analysis, which is associated with the computational processes required to complete a cognitive task. In contrast, informal reasoning largely depends on the intentional or rational level of analysis, which refers to the goals of the person (that is, what he or she is trying to compute, and why) along with his or her knowledge. This explanation parallels the typical/maximal distinction described earlier. In both of these accounts, the limiting factor of informal reasoning under non-directive instructional conditions is a choice not to put forth one's maximal effort on the task. Certainly, it is less cognitively demanding to note the main reasons one holds the views that one does, as opposed to exploring the issue from both sides. That is, developing and exploring a full situation model of the problem space is a far more effortful process (Perkins et al., 1991). Whether a person chooses to put forth this additional effort in the absence of directive instructions may be due to dispositional and motivational factors, not ability level.

One study has examined this possibility, at least obliquely. Macpherson and Stanovich (2007) divided their college sample (*N* = 195) into quartiles based on cognitive ability, and examined the effect of directive instructions on myside bias on an argument generation task. They reported a trend toward the instructions being more beneficial (i.e., they were more successful at reducing myside bias) to participants in the lowest ability quartile, who scored in the average range of the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999). However, this trend did not reach significance [*F*(3, 181) = 1.82, *p* = .15 for a three-way interaction between instructional set (non-directive vs. directive), cognitive ability quartile, and argument type (myside vs. otherside), in a sample of 195 undergraduates]. However, Macpherson and Stanovich grouped participants by cognitive ability rather than treating ability as a continuous covariate, which may have reduced their power to detect an effect.

Prior knowledge. Of all of the possible individual differences correlates of informal reasoning, content knowledge has received perhaps the most attention. The general assumption is that a larger store of relevant domain knowledge and conceptual understanding is helpful for reasoning (Sadler & Zeidler, 2005). Clearly, having a minimal amount of knowledge is necessary—it would be difficult for a person to reason effectively about the morality of genetic manipulation, for example, if he or she had absolutely no knowledge of genetics or heredity (Sadler & Zeidler, 2005). Beyond that most basic level of knowledge, however, support for a substantial relationship between content knowledge and informal reasoning performance generally has been nonexistent in both schoolchildren (Kortland, 1996) and in samples representing a wider range of adolescents and adults (Kuhn, 1991; Perkins et al., 1991). There are two exceptions. The first exception is Furlong (1993), who assessed specific topic knowledge by asking participants an open-ended response question about the factors that they thought contribute most to the federal budget deficit. In his reasoning task, participants were asked to propose solutions to the federal budget deficit, on both the revenue and spending sides. Given the similarity between these two tasks, it is perhaps not surprising that he found performance on the knowledge test to be strongly correlated with performance on the reasoning task (r = .55 for number of premises; r = .54 for overall rating).

Aside from Furlong (1993), Means and Voss (1996) provide the only other study that has reported the relationship between prior knowledge and reasoning outcomes, such as the number of arguments generated and their quality. In a sample of 90 students in

Grades 8, 10, and 12, knowledge about the effects of drugs and alcohol was a significant predictor of the number of arguments generated in response to reasoning prompts on the same topic (Kendall's $\tau = .56$ for the number of supporting reasons generated and $\tau = .42$ for the number of qualifiers, which set limits on the conditions in which a conclusion applies and are considered evidence of a high-quality argument). However, the soundness of the supporting reasons and the quality of the supporting reasons was predicted by student ability level (gifted, average, or below average), not by knowledge (that is, knowledge did not remain a significant predictor after ability level was partialled out; for all reason types, $\boldsymbol{\tau} < .16$ after ability level was partialled out). Means and Voss concluded that knowledge, by itself, does not produce sound arguments, although it does produce *more* arguments.

The research program of Sadler (e.g., Sadler & Zeidler, 2005) deserves mention because it has focused on content knowledge and reasoning. However, due to design differences, it is difficult to compare his findings with those of other researchers. Sadler has attempted to separate informal reasoning from argumentation skill by specifically prompting participants for arguments, counterarguments, and rebuttals. As a result, these outcomes cannot be used as dependent variables. Instead, in a study by Sadler and Zeider (2005), responses were scored based on intra-item and inter-item response coherence and overall ratings for counterarguments and rebuttals. The researchers found that they could not reliability separate the four criteria, and ultimately scored the items dichotomously, based on whether or not it exhibited any of four specific reasoning flaws. The fact that Sadler and Zeidler found a large effect of domain knowledge ($d = 1.23$, comparing the high-knowledge group to the low-knowledge group) using this relatively impoverished indicator of reasoning is notable, but must be considered in the context of their extreme-groups design. Participants were selected based on their very high or very low scores on a test of genetics knowledge, and it is possible that these groups differed in other factors (e.g., general ability, motivation, or personality traits associated with knowledge acquisition) in addition to content knowledge. Between this design issue and the fact that the dependent variable was different from those used by other studies, it is difficult to integrate these findings into those of other research programs.

Sadler and Zeidler (2005) and Means and Voss (1996; see also Zeidler & Schafer, 1984) assessed prior knowledge using multiple-choice items drawn from the relevant academic domain. Such tests are subject to the same limitations of any domain knowledge test, in that they necessarily only sample a small number of possible items from the domain. It is unclear whether a 20-item multiple-choice test, consisting mostly of isolated facts, can measure the kind of integrated domain knowledge that would be expected to be most beneficial to informal reasoning. Furlong (1993) used an open-ended question to assess domain knowledge, but chose one so similar to his reasoning prompt that it is perhaps surprising that it did not correlate more than the reported $r = .5$ with reasoning outcomes.

An alternative approach to assessing topic-related knowledge is via self-report. Perkins (1985a) asked participants to indicate the amount of time they had spent considering each issue in the past. Furlong (1993) asked his participants about their prior involvement in the topic (e.g., reading about the issue or raising the issue in conversation). Furlong also asked his participants to rate their overall knowledge of the issue on a scale of 0 to 9. This self-report approach has been used in the wider ability research as a proxy for objectively assessed knowledge (e.g., Rolfhus & Ackerman, 1996; although correlations between self-report and objectively assessed knowledge vary at the domain level; see Ackerman, Beier, & Bowen, 2002). The predictive power of, and relationships between, such prior knowledge and issue engagement variables have not been studied in an informal reasoning context, but they may reveal useful information beyond what can be obtained from relatively short multiple choice tests of content knowledge. In addition, these predictors may impact the quality of information search, if participants are allowed to consult outside sources during an informal reasoning task (Willoughby, Anderson, Wood, Mueller, & Ross, 2009).

Prior opinion. A major question in the reasoning literature is the extent to which people can reason independently of their prior opinion about a topic. Considering that being able to reason independently of prior beliefs is a key component of many definitions of critical thinking (see Stanovich & West, 1997, for a review), it should not be surprising that prior beliefs are commonly measured in informal reasoning research. Prior beliefs influence people's interpretations of information about controversial issues (Kardash & Scholes, 1996), as well as judgments about the quality of evidence, even among scientists (Koehler, 1993). Prior beliefs are often invoked when high school biology students reason about controversial socioscientific issues, and play a prominent role in some students' judgments of argument persuasiveness (Sadler, Chambers, & Zeidler, 2004).

Prior beliefs are of critical importance in investigations of myside bias—the very definition of myside bias involves prior beliefs (that is, to exhibit myside bias is to generate more arguments that support one's pre-existing position on the issue than support the other side). Typically, prior belief is assessed using items embedded in a questionnaire that participants complete before engaging in the reasoning tasks (e.g., Toplak & Stanovich, 2003). The direction of the prior opinion is then used to determine which side of the issue (pro or con) represents “myside” for that participant. Typically, myside bias is observed when participants are not given specific instructions to consider both sides of an issue and to set aside prior beliefs (Baron, 1995; Macpherson & Stanovich, 2007; Perkins et al., 1991; Wolfe & Britt, 2008). Although a common assumption is that myside bias arises from inability or unwillingness to entertain otherside arguments, Wolfe and Britt (2008) have shown that myside bias is present even when participants have been assigned to write an essay supporting a stance with which they do not agree. That is, people present more arguments in support of the position they are arguing, whether or not they agree with it. This suggests that myside bias may be rooted at least partially in people’s beliefs about what constitutes a persuasive argument (Baron, 1991), rather than being due to prior opinion alone.

In addition to the direction of prior beliefs, the impact of the strength of prior beliefs on reasoning has also been studied (Wolfe, 2012). Wolfe hypothesized that stronger prior opinions on an issue would be associated with a greater degree of bias. This hypothesis was partially supported by Wolfe (2012), in that he found a relationship[8] between opinion strength and bias in notes that participants took in preparation for writing essays, but not in the essays themselves. However, Wolfe used an essay prompt that was designed to elicit very little variance in opinion direction (he described scores as “essentially dichotomous,” which presumably means that most participants rated their

[8] This result is not reported in such a way that an effect size can be computed.

opinion as either “disagree” or “strongly disagree,” although the explanation of these results is not entirely clear). In addition, Wolfe’s myside bias criterion was dichotomous (essays and notes that did not contain any mention of the opposite side of the issue were scored as exhibiting myside bias, whereas any mention at all of the other side resulted in a “not biased” rating). Finally, Wolfe provided his participants with a booklet containing an equal number of arguments on both sides of the issue, which participants may have interpreted as something they must consider in full. For these three reasons, Wolfe’s study is rather different from most investigations of the relationship between prior opinion and reasoning, and it remains unclear how strength of belief would be related to reasoning outcomes under less restrictive conditions.

Non-ability factors. Two dispositional factors have been investigated in relation to informal reasoning: actively open-minded thinking (Stanovich & West, 1997) and need for cognition (Cacioppo & Petty, 1982; Cacioppo, Petty, & Kao, 1984).

Actively open-minded thinking. Stanovich (e.g, Macpherson & Stanovich, 2007; Stanovich & West, 1997) investigated the relationship between informal reasoning and his construct of actively open-minded thinking (AOT) in multiple studies. Items in the AOT measure were borrowed from several pre-existing questionnaires of various constructs, including the Openness—Values facet of the NEO, a categorical thinking scale, a flexible thinking scale, and three dogmatism scales (see Stanovich & West, 1997, for a full description). Stanovich’s program of research in relation to informal reasoning has focused on myside bias, so most of the actively open-minded thinking results available are related to that outcome. Overall, the scale has shown negligible correlations with myside bias in argument generation tasks (r = -.03 in a sample of 195 undergraduates; Macpherson & Stanovich, 2007). Thus, despite its attractiveness on a conceptual level as an identifier of individuals who would be more likely to consider multiple sides of an issue, attempts to relate it empirically to reasoning quality (as measured by myside bias) have not been successful.

Need for cognition. Need for cognition (Cacioppo & Petty, 1982; Cacioppo et al., 1984) is another conceptually appealing construct, as it represents people's tendency and desire to engage in effortful cognitive activity. However, it generally has not been found to correlate substantially with informal reasoning performance. For example, Furlong (1993) reported significant but moderate zero-order correlations (r = .26 range; N = 61 college students and adult nonstudents) between need for cognition and two of his five reasoning outcomes, but the variable ultimately was not retained in the final multiple regression analysis.

Summary of non-ability factors. Neither actively open-minded thinking nor need for cognition have been shown to have substantial relationships with informal reasoning. Stanovich's research has largely focused on argument generation tasks, so actively open-minded thinking has not been investigated in the context of the more complex essay and interview tasks. However, need for cognition, a similar construct, had a significant but rather modest (r = .26) relationship with the interview task used in Furlong's (1993) study. Furlong did not differentiate between performance under directive versus non-directive instructions when calculating the correlation between reasoning performance and his predictors. It is possible that the correlations differ in the two conditions. The typical/maximal account would lead to a prediction that a stronger relationship between personality factors and reasoning should be observed in the non-directive instructions condition, compared with the directive instructions condition. To date, no studies have examined the relationship between any personality factors and informal reasoning under different instructional conditions.

Program of Research

In light of the findings and limitations of the extant literature, I designed the current research program to address two overarching questions:

1. What is the effect of unrestricted access to information (via the Internet) on informal reasoning performance?

2. What intra-individual and inter-individual differences variables are related to reasoning performance outcomes, and do instructional manipulations and access to the Internet reduce or enhance their effects?

In order to address these questions, I conducted two experiments. In the first experiment, I examined potential individual differences correlates of reasoning performance. This first experiment was also used to select topics for the second experiment, which addressed both research questions using experimental manipulations of Internet access and instruction set.

Study 1. Study 1 served two purposes. The first purpose was to inform the selection of the prompts that were used for the second study. The second purpose was to assess within-person and between-person predictors of informal reasoning. I anticipated that the most important predictors of reasoning performance would be item-specific knowledge and involvement. This is based on the findings of Furlong (1993), who reported significant correlations between these two predictors and reasoning performance. This prediction is also consistent with the principle of Brunswik symmetry

(Wittmann & Süβ, 1999), which states that predictors will have the greatest predictive power if they match the criterion variables in terms of level of specificity. In the present context, this means that item-specific knowledge should be a better predictor of reasoning performance (assuming that domain knowledge is in fact relevant to reasoning quality) than more general characteristics of the person, such as general intelligence, personality, or a broad domain knowledge test. Conversely, general traits should be better predictors of performance only when the performance criteria are aggregated; it is difficult to predict specific incidents of behavior from general traits (Wittmann & Süβ, 1999). Considering the limited number of reasoning prompts that have been used in most previous studies, it is possible that the generally low correlations between trait factors (ability, personality) and reasoning performance are due, in part, to violations of Brunswik symmetry. The present study is unique in that it allows aggregation across a larger pool of items than has been possible in past studies.

I therefore made the following predictions:

Hypotheses 1a & 1b: For individual reasoning items, item-level knowledge and involvement factors will correlate with the number of (a) myside arguments and (b) otherside arguments (anticipated r *= .35).*

*Exploratory Hypotheses 1a & 1b: Correlations between person-level traits (ability, personality) and (a) myside arguments and (b) otherside arguments generated are expected to be negligible at the item level (*r *< .20).*

Study 2. The second study was designed to address two sets of hypotheses. The first set involves experimental manipulations of instructions and access to the Internet. The second set involves individual- and item-level predictors of reasoning performance.

Specifically, this study examines whether the predictors of reasoning performance differ in response to the experimental manipulations.

Experimental hypotheses. Study 2 involved two within-subjects manipulations designed to address both of the research questions. In the first manipulation, access to the Internet was manipulated in order to investigate the effects of unrestricted access to information on reasoning performance. In the second manipulation, instructions were manipulated in order to replicate and extend previous findings regarding the impact of directive versus non-directive instructions on reasoning performance. The main effect of instructional conditions is well-established in the literature. The effect of free access to outside information during informal reasoning has not been investigated previously, nor has the relationship between reasoning performance and individual differences predictors under these various conditions.

Several previous studies using undergraduate samples (Furlong, 1993; Macpherson & Stanovich, 2007; Nussbaum & Kardash, 2005; Wolfe & Britt, 2008) investigated the effect of directive versus non-directive task instructions on various reasoning performance indicators including myside/otherside arguments or their equivalents. Without exception, they reported medium-large to large effects of directive instructions on the number of counterarguments (i.e., otherside arguments) generated ($1.12 < d < 3.33$[9] and $d = 1.24$ in Furlong, 1993, and in Nussbaum & Kardash, 2005, respectively). Previous studies have not found significant effects of instructions (or

[9] Furlong (1993) does not report the correlation between performance on his two reasoning items, making it impossible to compute the within-person effect size precisely. The low end of the range reported here represents the result for a correlation of $r = .10$; the top of the range corresponds to a correlation of $r = .90$.

directive follow-up prompts, in the case of interviews) on the number of myside arguments (Nussbaum & Kardash, 2005).

Therefore, in light of previous research, I proposed the following hypothesis:

Hypothesis 2: The main effect of instruction condition will result in a large effect on the number of otherside arguments (expected effect size: d = *0.8) such that more otherside arguments will be generated in response to the directive instructions condition.*

The impact of unrestricted Internet access on reasoning performance has not been studied. However, on a general level, social commentators (Pariser, 2011) have argued that despite initial hopes that the Internet would broaden people's exposure to new ideas and information, in practice it has merely made it easier for people to seek out others who share their views. In addition, Perkins et al. (1991) noted that more intelligent undergraduates (with no access to outside information during reasoning) seemed to use their intelligence to generate more arguments supporting their own position, rather than to more fully explore the problem space. To the extent that outside information might make people "more intelligent," it might also exert its main influence in the number of myside arguments generated, but may not change the number of otherside arguments in the absence of directive instructions. Because there is no existing research that can inform an estimate of the magnitude of this effect, I assumed a medium-sized effect.

Hypothesis 3: *The main effect of Internet access will result in more myside arguments being generated in the Internet access condition than in the no-Internet condition (*d = *0.5).*

In addition to the two main effects, I also examined the interaction between Internet access and instructional condition. As with the main effect of Internet access, there is no directly relevant extant literature on which to base predictions about the interaction effect. However, I expected an interaction between Internet access and instructional condition, such that performance would be enhanced in the Internet/directive instructions condition. That is, after having been told to include otherside arguments in their lists of arguments, participants would use the Internet in order to generate more otherside arguments than they would have generated spontaneously. In the absence of any literature on which to base an anticipated effect size for the interaction, I proposed a medium-sized effect on otherside arguments. I was agnostic as to whether there would be an interaction effect on the number of myside arguments.

Hypothesis 4: *The interaction between Internet access and instructional conditions will result in significantly more otherside arguments in the Internet/directive instructions condition, compared with the other conditions (anticipated effect size:* d = *0.5).*

Individual differences hypotheses. The overarching question addressed by the individual differences component of this study is whether Internet access and directive reasoning instructions attenuate or exacerbate the effects of individual differences on reasoning outcomes. In general terms, there are two overall patterns of results that I expected could emerge in the results of this study. In the first overall pattern, Internet access and directive instructions both serve to mitigate the effects of cognitive individual differences on reasoning performance. The substantive interpretation of this pattern is that access to outside information allows people to compensate for cognitive shortcomings. According to this argument, allowing people to access information would level the playing field between people who possess greater or lesser amounts of prior knowledge. At the same time, this pattern of results would suggest that the directive instructions serve as scaffolding (Perkins, 1985a) for the reasoning task, and that this scaffolding is especially beneficial to those of relatively lower cognitive ability, who may not perform as well in the absence of directive instructions (Macpherson & Stanovich, 2007). Alternatively, a second potential overall pattern could show that Internet access and directive instructions strengthen the relationship between individual differences and reasoning performance. In this scenario, more able (or knowledgeable) people would be in a better position to capitalize on their pre-existing advantages when given access to the Internet and/or when given directive instructions. This scenario is in line with the Matthew Effect (Stanovich, 1986) in which the "rich get richer"—or, in this case, the better-prepared get better at generating arguments when given access to outside information.

In previous research, non-ability traits such as actively open-minded thinking generally have exhibited low or non-significant correlations with reasoning performance (e.g., Macpherson & Stanovich, 2007). However, researchers typically have collapsed performance across instructional conditions when computing correlations with non-ability traits, leaving open the possibility that such traits may be differentially predictive under different conditions. That is, when participants are given non-directive instructions, non-ability traits may indeed predict reasoning outcomes. However, when participants are specifically told to generate arguments on both sides of an issue, the directive instructions may wash out most or all of the influence of non-ability traits. Therefore, an aim of Study

2 was to examine the relationship between the non-ability traits and reasoning performance under each of the two instructional conditions separately.

I expected that the Internet manipulation might also yield different relationships between individual differences and reasoning performance. It is possible that Internet access would increase the influence of personality factors that are positively associated with reasoning performance. One mechanism by which this effect may operate is through removing content knowledge as a limiting factor. For example, a person high in typical intellectual engagement (Goff & Ackerman, 1992) may be quite willing to consider both sides of an issue, but may not be familiar enough with the issue to be able to generate more than one or two arguments on each side. When given Internet access, such a person may use the opportunity to more thoroughly explore the problem from multiple perspectives.

The individual differences variables examined in this set of studies can be classified into two groups: knowledge/abilities and non-ability traits. The two groups are associated with different hypotheses for the experiment.

Knowledge and abilities. Measures of domain knowledge (Furlong, 1993; Sadler & Donnelly, 2006) and cognitive ability (Macpherson & Stanovich, 2007; Perkins et al., 1991; Toplak & Stanovich, 2003) have been collected frequently in reasoning research, but results have been mixed. Although relevant knowledge and cognitive ability would seem to be necessary for good informal reasoning performance, it appears that neither is sufficient.

A study by Macpherson and Stanovich (2007) is the only one to have examined the impact of an instructional manipulation on the reasoning performance of people of different cognitive ability levels. They noted a trend toward directive instructions reducing the degree of myside bias, primarily by increasing the number of otherside arguments, in the lowest cognitive ability quartile in their sample [$F(3, 181) = 1.82$, $p = .145$ for a three-way interaction between instructional set (non-directive vs. directive), cognitive ability quartile, and argument type (myside vs. otherside), in a sample of 195 undergraduates]. Means and Voss (1996) have suggested that informal reasoning is, at its core, a specific way of using language that is learned gradually over time, and that more intelligent people learn it better than less intelligent people. According to this line of thinking, directive instructions would be expected to be more helpful to people with lower cognitive ability, because they have not internalized this specific way of using language to the same degree as people who are more intelligent. This would result in a weaker relationship between cognitive ability and the number of otherside arguments in the directive instructions condition.

Extrapolation from the typical/maximal account of the instructional manipulation provides an alternative prediction. According to this account, ability should become a stronger predictor of otherside arguments under the directive instructional set, insofar as the directive instructions create a more constrained situation that elicits performance that is less influenced by participants’ personalities. According to this line of thinking, ability would be the limiting factor on performance under directive instructions, which would lead to a stronger correlation between ability and the number of otherside arguments generated. I predicted that the typical/maximal account would prevail.

Hypothesis 5: The relationship between cognitive ability and the number of otherside arguments will be significantly stronger in the directive instruction conditions than in the non-directive conditions (anticipated difference between correlations: Δr = *.3).*

Unlike cognitive ability, limitations in reasoning performance due to lack of domain knowledge seem unlikely to be mitigated substantially through instructions, because the limitation is not due to failure to understand the task goals. However, domain knowledge may interact with Internet access condition to affect reasoning performance. The form that this interaction may take is uncertain. One prediction is that having all of the information on the Internet at one's fingertips would serve as a great equalizer, reducing the impact of prior knowledge on reasoning performance. However, research on Internet search behavior has found that higher domain knowledge is associated with more successful Internet searches and better responses to search prompt questions, at least for relatively constrained search prompts (Desjarlais & Willoughby, 2007; Willoughby et al., 2009). In fact, undergraduates (*N* = 150) who were allowed to access the Internet while writing an essay in a topic area in which they had low domain knowledge, did not produce significantly better essays than similarly low-domain respondents who were not provided access to the Internet (Willoughby et al., 2009; see also Symons & Pressley, 1993, for a review of similar research regarding the impact of domain knowledge on searching for information in print resources such as textbooks). Desjarlais and Willoughby (2007) suggested that individuals with low domain knowledge become "lost in hyperspace" (Desjarlais & Willoughby, 2007, p. 5) when the non-linear organization of the information on the Internet imposes excessively high cognitive demands on the low-domain-knowledge learner. Thus, based on the information search literature, I expected to observe a "Matthew effect" (Stanovich, 1986) in the interaction between Internet access and prior knowledge, such that individuals with greater domain knowledge would benefit more from Internet access than people with lower domain knowledge.

Hypothesis 6: The relationship between prior knowledge and the number of myside arguments will be stronger in the Internet conditions than in the no-Internet conditions (anticipated effect size: Δr = *.3).*

Non-ability traits. Given that previous studies have not consistently identified salient non-ability trait predictors of informal reasoning, I sought to include traits that may be expected to be related to reasoning, but that had not necessarily been investigated previously. In accordance with the typical/maximal framework of performance, I expected that all of the non-ability traits would exert their greatest influence under the weaker situation (i.e., the non-directive instructions).

Typical intellectual engagement (TIE; Goff & Ackerman, 1992) reflects an interest in intellectual pursuits and activities. It is related to the personality trait need for cognition (Cacioppo et al., 1984; Woo, Harms, & Kuncel, 2007), which has been used in informal reasoning research (Furlong, 1993). In the present context, TIE is expected to be positively related to the number of arguments generated, and particularly to the number of otherside arguments.

Dogmatism is "relatively unchangeable, unjustified certainty" in one's beliefs (Altemeyer, 2002, p. 713). Such certainty may influence informal reasoning by making it difficult to entertain opinions other than one’s own, and as a result may be particularly related to the number of otherside arguments generated.

Need for closure refers to a desire to find an answer to settle on “*an* answer on a given topic, *any* answer, . . . compared to confusion and ambiguity” (Kruglanski, 1990, p. 337, italics in original). Individuals high in need for closure tend to terminate the decision-making process early in order to escape what they consider to be uncomfortable uncertainty; individuals low in need for closure tend to be willing to seek additional information and to reexamine currently held beliefs (Kruglanski & Webster, 1996). Need for closure may be negatively related to the number of arguments, particularly the number of otherside arguments, because of its association with a desire for certainty.

As noted above, I expected that personality factors would have the greatest effect under the non-directive instructions, which represent a weaker situation than the directive instructions. Therefore, the expected results for all of the non-ability trait variables used in Study 2 were as follows:

Hypothesis 7: The relationship between non-ability traits and the number of otherside arguments will be weaker in the directive instruction conditions than in the non-directive instruction conditions (anticipated effect size: Δr *= .3).*

I did not have specific hypotheses about the relationship between non-ability traits and myside arguments. Neither previous research nor the relevant theoretical frameworks provide compelling predictions about such an effect in the context of the manipulations in this experiment. Therefore, I was agnostic about changes to the relationship between non-ability traits and reasoning across conditions.

Prior opinion. I anticipated that directive instructions would reduce the influence of prior opinion on the number of otherside arguments generated, compared to the non-directive instructions. The rationale for this prediction is the same as the rationale for the non-ability trait factors. That is, I expected non-cognitive factors to become less important in the more constrained directive instruction condition.

Hypothesis 8: The relationship between prior opinion and the number of otherside arguments will be significantly weaker in the directive instructions condition than in the non-directive instructions condition (anticipated difference between correlations: Δr = *.3).*

Interim Summary and Overview of Method

The two studies in this project were designed to address gaps in the literature. Specifically, I expected to replicate previous findings that directive instructions would lead to a large increase in the number of otherside arguments generated. In addition, I anticipated that the novel manipulation of allowing participants to look up information online would increase the number of myside arguments generated in response to informal reasoning prompts, but that Internet access would interact with instructional condition to produce the greatest number of otherside arguments in the Internet/directive instructions condition. In addition, this study was designed to examine a wide range of individual differences correlates of informal reasoning. Regarding individual differences, I anticipated that the predictive power of ability and knowledge would increase under directive instructions and Internet access conditions, respectively, but that non-ability traits would be relatively better predictors under non-directive instructions.

CHAPTER 2
STUDY 1

The first study served two purposes in the research program. The main purpose was to inform selection of items to be used in Study 2. Study 2 required four items approximately matched in terms of the number of arguments that are generated for each side, the amount of bias that can be expected, and the range of relevant experience and prior knowledge. The second purpose of Study 1 was to examine the relationships between informal reasoning performance and potential individual differences correlates.

Method

Power analysis. I conducted a power analysis for Hypotheses 1a and 1b using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009). A sample size of 84 would enable me to detect a correlation as small as $r = .3$ with the standard level of power ($1 - \beta = .80$). Therefore, I aimed to recruit 100 participants with the goal of having 84 participants complete the full protocol and provide complete and useable data. Note that the exploratory hypotheses were not considered for this power analysis. Testing the exploratory hypotheses at the conventional power level ($1 - \beta = .80$) would require more than twice the sample size required for testing Hypotheses 1a and 1b. The exploratory hypotheses were not of great theoretical importance for the current project, but were assessed because they have the potential to provide some useful information. The underpowered tests would not allow me to assert the absence of a relationship between person-level traits and item-level reasoning performance if the results relevant to those hypotheses were not significant. However, if the relevant correlations were found to be significant, then this would suggest that there is not in fact a dissociation between person-level and item-level predictors, in terms of their relationships with item-level reasoning performance. Although this approach is not in line with the standard null hypothesis significance testing framework, it nevertheless has the potential to provide some information that can guide future work in this area.

Sample. Georgia Tech undergraduates were recruited through a posting on SONA and flyers placed around the School of Psychology. Non-student participants were recruited via postings on Amazon's Mechanical Turk (MTurk). The post made clear that the project was a research study. In both samples, participants were required to be fluent in English. MTurk participants were required to have a 95% or greater approval rating of their submissions.

Thirteen Georgia Tech students completed the study in group sessions in a computer lab. In the MTurk sample, 186 participants started the study. Of these, 93 completed the entire protocol. The other 93 either closed the browser themselves before finishing the full protocol, or failed one of the attention checks embedded in the questionnaires and were redirected out of the study. Between the two samples, 106 participants completed the full study and provided at least one pro or con argument for at least eight of the 10 prompts. Six of the MTurk participants failed an additional attention check embedded in the instructions for the argument generation task and were removed prior to analysis. This resulted in a final sample of 100, including the 13 Georgia Tech students.

The mean age of the full sample was 34.1 years (*SD* = 12.7, range: 18—67). Forty participants were male, 58 were female, and two did not report their gender. Most participants (81 out of 100) had at least some college education. Fourteen were pursuing or had attained a graduate degree. For analysis purposes, participants were grouped into five educational status groups [high school only (*N* = 19); pursuing or attained associate's (*N* = 21); pursuing bachelor's (*N* = 20); attained bachelor's (*N* = 26); pursuing or attained graduate degree (*N* = 14)]. For analysis purposes, the education variable was represented using a single ordinal variable coded 0 (high school only) through 4 (graduate degree).

MTurk participants were compensated $2.50 deposited to their MTurk account upon completion of the study, and Georgia Tech students were compensated with 3.0 SONA credits.

Measures. Study 1 included self-report measures, ability tests, and the informal reasoning task.

***Self-report questionnaires*.** Participants completed a series of self-report questionnaires assessing a variety of personality and item-relevant factors.

Typical intellectual engagement. Typical intellectual engagement (TIE) was assessed using the 12-item short form of the scale of the same name (Goff & Ackerman, 1992). The internal consistency reliability of the 12-item short form TIE scale in a large sample of high school students was α = .86 (P.L. Ackerman, personal communication, March 24, 2012). The response scale is 1 (strongly disagree) to 6 (strongly agree). A representative item is "Thinking is not my idea of fun" (reverse scored).

Dogmatism. Dogmatism was assessed using the 20-item DOG Scale (Altemeyer, 2002). The original measure has a 9-point response scale, but a 6-point scale was used for this study to maintain consistency with the other personality questionnaires. An example item is "If you are 'open-minded' about the most important things in life, you will probably reach the wrong conclusions." (α = .88 to .93; Crowson, 2009)

Conservatism. Because some complex informal reasoning items have implied political components, the Social and Economic Conservatism Scale (Everett, 2013) was used to assess conservatism. The primary motivation for including this scale was to identify items for which performance was strongly related to political orientation, which was not considered to be a desirable item characteristic. This scale consists of 12 concepts that are characteristic of modern conservatism in the United States, and respondents rate how positively or negatively they feel about each topic using a “feeling thermometer” scale ranging from 0 (extremely negative) to 100 (extremely positive). An advantage of this scale is that it allows respondents to indicate their feelings about an issue without requiring that they have any particular policy knowledge about the issue (Everett, 2013). The original item “traditional marriage” was replaced with “same-sex marriage” and reverse-coded. All other items appeared as in the original scale. (α = .88; Everett, 2013)

Need for closure. The 15-item short form of the Need for Closure scale (Roets & Van Hiel, 2011) includes items such as “I don’t like situations that are uncertain.” The questionnaire used a 6-point response scale. (α = .87)

Self-reported knowledge. Participants rated their knowledge of each of the topics that were presented as reasoning prompts. The instructions and 8-point response scale for this questionnaire were based on those used by Rolfhus and Ackerman (1996) for their self-reported knowledge scales.

Issue involvement. Furlong (1993) measured prior topic involvement using items such as “I read or listen to news stories about this issue.” Furlong’s scale used a 5-point response scale, but I used a 6-point scale to maintain consistency with other personality measures. (α = .84; Furlong, 1993)

Prior opinion. Items probed participants' opinions about the topic of each item prompt to be used in the argument generation task using a 1 (strongly disagree) to 6 (strongly agree) scale.

Cognitive ability. Three tests assessed cognitive ability, specifically crystallized intelligence. Participants were instructed not to look up any answers to the test.

Extended-range Vocabulary Test. In this test (Ekstrom, French, Harman, & Dermen, 1976), examinees must choose the word whose definition most closely matches a target word. The original test consists of two parts, but only the first part was used for this study. Participants had 6 minutes to complete 24 items. Score is the number of correct items, minus 0.25 times the number of incorrect items. (α = .78; Ackerman & Beier, 2007)

Verbal Analogy Test. This test (Ackerman & Kanfer, 1993) consisted of analogies in the format "Light : Dark :: Pleasure : _______." Participants had 12 minutes to complete this 50-item multiple-choice test. Score is the number of correct items, minus 0.25 times the number of incorrect items. (α = .78; Ackerman & Kanfer, 1993)

General knowledge. A general current-events knowledge test was created for this study. The 40 multiple-choice items in this test referred to national and international events from approximately the past five years (2010 to 2015). Items covered a range of difficulty. There was no time limit for this test. The score is the number of correct items.

Argument generation task. The reasoning task consisted of 10 items, and participants generated arguments for each, as described below.

Items. Reasoning prompts were drawn from the literature and additional prompts were generated for this study. They represented a variety of topic domains, and pilot testing indicated that multiple lines of reasoning could be generated in response to them. For example, in response to a prompt asking whether marijuana should be legalized in the United States, participants could generate arguments related to economics, medicine/science, public health, the role of government and personal freedom, and morality/ethics. Table 1 presents the text of the 10 prompts. Items were presented to participants in one of four orders, to distribute possible order effects.

Table 1: Argument Prompts

Code	Prompt Text
*TV	Should the amount of violence on television be restricted?
*GMO	Should genetically modified foods (e.g., corn, fish) be produced and sold in the USA?
*LDS	Should learning-disabled students be integrated into regular school classrooms?
*TRI	Should college instructors provide "trigger warnings" when they are about to present class material or assign reading that might upset some students (e.g., material related to sexual assault, discrimination, child abuse), and allow students to leave the lecture or avoid the reading if they feel they might be upset by the material?
GG	Should the US government take additional steps to reduce greenhouse gas emissions?
MJ	Should the recreational use of marijuana be legalized in the USA?
SB	Should the USA have a national mandatory seatbelt law?
KND	Should all-day kindergarten (that is, kindergarten that has the same length school day as the higher elementary school grades) be provided free-of-charge in all public school districts in the USA?
ALC	Should the drinking age in the USA be lowered to 18 years?
RX	Should a pharmaceutical company be allowed to charge any amount it wants for a drug?

*Prompt included in Study 2.

Task procedure. The argument generation task was based on a task of the same name used by Stanovich (Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003). In this task, participants are asked to list arguments relevant to a given issue. Arguments were scored using the procedure described in Toplak and Stanovich (2003). Specifically, each argument was scored as pro or con for the issue. For example, one item was “Should the recreational use of marijuana be legalized in the United States?” An argument such as “The government should not restrict individual freedoms” would be a pro argument for this item, whereas an argument such as “Marijuana is often a ‘gateway’ drug to other, more harmful drug use” would be a con argument.

Scoring. Two raters (another graduate student and I) coded all arguments as pro, con, or exclude. The exclude category was used for statements that were not arguments (e.g., “I don’t know much about this topic”), that were conditional (e.g., “Marijuana should be allowed only if it can be proven that it is not harmful to society”), or whose pro/con directionality was not clear. Across all prompts, 8.2% of arguments were excluded (343 excluded out of 4,175 total arguments in the final sample), which is in line with the percentage of arguments excluded in previous studies (Macpherson & Stanovich, 2007).

Pro and con arguments were then translated into myside or otherside arguments for each individual participant, based on responses to the prior opinion items in the self-report questionnaire. To do this, participant responses to the opinion items were dichotomized into agree or disagree. For participants who agreed with an issue statement, pro arguments were myside arguments, and con arguments were otherside arguments. For participants who disagreed with an issue statement, con arguments were myside arguments, and pro arguments were otherside arguments. The numbers of myside and otherside arguments for each item serve as the two main dependent variables in this study. The amount of time spent on each prompt's screen was also recorded.

Procedure. Participants completed the protocol online via Qualtrics. Georgia Tech participant pool participants accessed Qualtrics during a group session in a computer lab, and MTurk participants accessed it via posting on MTurk.

After providing informed consent, participants completed the questionnaires and ability tests. They were offered a break, and then were shown the instructions for the reasoning task. Participants were asked to think about each question and generate as many arguments as possible about each issue. They were told that they could take up to 10 minutes per question. To encourage participants not to rush though the items, the instructions stated: "Prior research with questions like these shows that people often come up with more ideas after they've been thinking about a question for a few minutes, so please take your time to think." Participants were also asked not to look up any information or consult with other people in generating their list of arguments. The program auto-advanced to the next prompt after 10 minutes, but participants were allowed to advance themselves when they were out of ideas, so in reality few participants spent the full 10 minutes on any prompts. Note that there was no instructional manipulation in this study, and all prompts were presented using non-directive instructions. The full text of the instruction screen is presented in Appendix A.

Results

As noted above, this study served two main purposes. First, it functioned as an extended pilot study for Study 2, and was designed to inform the selection of prompts to be selected for Study 2. I was also interested in leveraging the unusually large number of prompts administered in this study in order to examine the consistency of reasoning performance across prompts, as defined by otherside and myside arguments, and bias. The second main purpose of Study 1 was to examine the predictors of reasoning performance under non-directive instructional conditions.[10]

The main dependent variables of interest for this study are the numbers of myside and otherside arguments. Some previous researchers (e.g., Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003; Wolfe & Britt, 2008) have focused on myside bias as a reflection of reasoning ability, despite the problems associated with analyzing difference scores (Cronbach & Furby, 1970; Lord, 1963). In order to avoid these problems, myside bias is not the primary outcome of interest in this project. However, some results related to myside bias are reported in order to facilitate comparison with previous studies, or to make methodological points.

Results are presented in two main sections. In the first section, I describe the results of the argument generation task. This section includes issues of methodological importance such as inter-rater reliability and the distribution of pro/con arguments and prior opinions for individual items. I also present results related to the consistency of performance across the 10 items. In the second section, I assess the predictors associated with reasoning performance in three different ways: zero-order correlations, multiple regression and hierarchical linear modeling (HLM).

[10] All results reported here are for the full sample, but the pattern of results is the same when the 13 Georgia Tech undergraduates are excluded.

Assessments of abilities and non-ability factors. Scores on the topic involvement scale and the self-reported knowledge item for the same prompt topic were substantially positively correlated ($.45 \leq r \leq .76$), so the self-reported knowledge item was added to the topic involvement scale as a fifth item, to create a "topic exposure" variable. The three ability measures (Extended Range Vocabulary, Analogy, and the current events knowledge test) were combined into a single crystallized intelligence factor by summing their *z*-scores. Table 2 contains descriptive statistics for the person-level predictors, and Table 3 presents the correlations between them.

Argument generation task. Descriptive statistics for the individual reasoning prompts are presented in Table 4. Inter-rater agreement (i.e., percentage of arguments that received the same pro, con, or exclude classification from both of the raters) exceeded 90% for eight of the 10 prompts; the lowest agreement was 88%. Most of the disagreements involved decisions to exclude arguments. There were few instances of disagreements between pro/con classifications (110 total across all prompts, or 2.6% of all arguments), and many of these reflected coding errors rather than actual disagreements. All disagreements were reconciled through discussion between the raters.

Turning to the distribution of arguments generated for each prompt, first note that prior opinion was measured on a 1—6 scale, so the midpoint is 3.5. The first five prompts listed in the table have prior opinions that are closest to 3.5 (range: 3.40 to 3.85; see the second column from the right in Table 4) and also a relatively equal mean numbers of pro and con arguments (percent of pro arguments generated by an average participant for these four prompts: TV: 46.0%; GMO: 49.5%; LDS: 46.8%; TRI: 52.0%; ALC: 46.8%). The other five prompts exhibit a greater discrepancy between the number of pro and con

Table 2: Descriptive Statistics for Person-Level Predictors (Study 1)

			Range	
Scale	Mean (SD)	α	Potential	Actual
Analogy	27.89 (8.93)	.74	-12.50—50	0—42.50
Vocabulary	11.43 (5.10)	.65	-6—24	-2.25—21.50
CE Knowledge	20.51 (6.20)	.72	0—40	8—36
TIE	52.88 (10.20)	.91	12—72	22—72
NFC	54.96 (10.90)	.86	15—90	22—79
Dogmatism	55.04 (16.29)	.93	20—120	26—112
Conservatism	639.70 (220.56)	.85	0—1200	50—1116

Note. N = 100. CE Knowledge = current events knowledge. TIE = typical intellectual engagement. NFC = need for closure.

Table 3: Correlations Between Person-level Predictors (Study 1)

Predictor	1	2	3	4	5	6
1. CE Knowledge						
2. Vocabulary	**.56**					
3. Analogy	**.56**	**.51**				
4. TIE	**.20**	**.25**	**.25**			
5. NFC	-.04	-.17	-.01	**-.36**		
6. Dogmatism	**-.31**	**-.36**	**-.26**	**-.24**	**.31**	
7. Conservatism	-.19	-.20	**-.23**	-.17	.14	**.34**

Note. N = 100. CE Knowledge = current events knowledge. TIE = typical intellectual engagement. NFC = need for closure. Bold text indicates correlations significant at $p < .05$

arguments, as well as more polarized opinions: Of the four that resulted in more pro than con arguments, the percentage of pro arguments for an average participant ranged from 71.1% (for the MJ prompt) to 81.7% (for GG). The one prompt with more con than pro arguments (RX) had 23.6% pro arguments.

Table 5 presents mean, standard deviations, and ranges for myside and otherside arguments for each of the 10 prompts. Note that the myside/otherside distribution is quite different from the pro/con distribution presented in Table 4, especially for the first five prompts. Also included in Table 5 are the results of paired *t*-tests comparing the number of myside and otherside arguments for each prompt. All of the differences are very large (smallest $t = 4.05$, $p < .05$, $d = 0.67$), replicating previous findings that many more otherside arguments than myside arguments are generated when instructions are non-directive.

Although the total number of arguments and myside bias are not primary outcomes of interest in this project (and also not independent of the myside and otherside dependent variables), I also examined predictors of these indicators in order to provide a fuller picture of the results, and to facilitate comparisons with previous studies. As in previous studies, the mean myside bias is positive ($M = 2.22$ across all participants and all prompts). However, it should be noted that across all participants and all prompts, 14.5% of cases (140 out of 967 cases in which at least one codable argument was provided for a given prompt) had negative bias (i.e., more otherside arguments than myside arguments) and an additional 7.8% of cases (76 out of 967) had zero bias (same number of myside and otherside arguments). That is, although the mean level of bias is positive and most cases exhibited bias, over 22% of cases did not exhibit this pattern.

Table 4: Descriptive Statistics for Reasoning Items (Study 1)

		M (*SD*) Range		*M* (*SD*)	
Prompt	Pct. Agreement	Pro	Con	Prior Opinion[a]	Topic Exposure[b]
TV[c]	91.5	1.73 (2.42) 0—14	2.03 (1.86) 0—7	3.40 (1.55)	17.24 (6.55)
GMO[c]	90.8	1.92 (2.12) 0—8	1.96 (2.10) 0—10	3.55 (1.55)	19.02 (6.51)
LDS[c]	90.4	1.73 (1.80) 0—7	1.97 (2.41) 0—14	3.85 (1.37)	14.28 (6.93)
TRI[c]	89.7	1.80 (1.85) 0—9	1.66 (1.88) 0—7	3.75 (1.40)	13.88 (6.49)
ALC	93.4	1.93 (2.20) 0—10	2.39 (2.76) 0—15	3.06 (1.81)	17.01 (6.23)
GG	93.4	3.00 (2.24) 0—11	0.67 (1.22) 0—6	4.89 (1.29)	20.52 (5.91)
MJ	96.0	3.54 (2.73) 0—13	1.44 (2.35) 0—13	4.31 (1.67)	21.03 (6.37)
SB	96.8	2.58 (1.73) 0—8	0.82 (1.34) 0—6	4.88 (1.31)	18.87 (5.94)
KND	92.4	2.82 (2.08) 0—10	0.86 (1.61) 0—9	4.69 (1.20)	14.32 (7.45)
RX	88.0	0.82 (1.38) 0—6	2.65 (1.87) 0—8	1.91 (1.17)	18.10 (6.01)

Note. *N* = 100. TV = TV violence. GMO = GMOs; LDS = learning-disabled students; TRI = trigger warnings; GG = greenhouse gas emissions; MJ = marijuana; SB = seatbelts; KND = all-day kindergarten; ALC = alcohol drinking age; RX = prescription drug costs.

[a]Possible range: 1—6. [b]Possible range: 5—32. [c]Prompt selected for use in Study 2.

Table 5: Myside and Otherside Arguments by Prompt (Study 1)

	Mean (SD) Range			
Prompt	Myside	Otherside	*t*	Cohen's *d*
TV	3.01 (2.36) 0—14	0.75 (1.10) 0—4	8.04*	1.23
GMO	3.08 (2.18) 0—10	0.80 (1.23) 0—5	8.44*	1.29
LDS	2.65 (2.19) 0—14	1.05 (1.73) 0—8	5.30*	0.81
TRI	2.32 (1.93) 0—9	1.14 (1.59) 0—7	4.05*	0.67
ALC	3.35 (2.40) 0—15	0.97 (1.99) 0—15	6.92*	1.08
GG	3.10 (2.17) 0—11	0.57 (1.14) 0—6	9.55*	1.46
MJ	4.02 (2.79) 0—13	0.96 (1.63) 0—6	8.32*	1.34
SB	2.80 (1.63) 0—8	0.60 (1.12) 0—6	10.14*	1.57
KND	2.80 (2.18) 0—10	0.88 (1.49) 0—9	6.49*	1.03
RX	2.74 (1.90) 0—8	0.73 (1.20) 0—6	8.40*	1.26
Total	29.87 (15.21) 3—96	8.45 (8.75) 0—33	13.24*	1.73

Note. $N = 100$.
**p* < .05.

Therefore, as always, it is important to keep in mind that some individual results vary substantially from aggregated results.

Consistency of reasoning performance. Prior researchers typically have used multiple items without much explicit consideration of the consistency of reasoning performance across items. Consistency would indicate that the items tap an underlying ability, while large inconsistencies across items would suggest that individual-level differences are not the main driver of performance on these items. That is, inconsistency would suggest that item-level predictors are a major source of variability in performance. In order to investigate this, I treated the 10 items as a scale and computed the internal consistency reliability. Cronbach's α for the total number of arguments generated for the 10 prompts was α = .95; for myside arguments, α = .88; for otherside arguments, α = .81; for myside bias, α = .75.

To provide a clearer picture of the relationship between items, correlations between myside and otherside arguments for the 10 prompts are presented in Table 6. There are two things to note about the pattern of correlations. First, there was a significant negative correlation between the number of myside and otherside arguments for seven of the 10 prompts (the remaining three were negative, but did not reach the threshold for significance). This suggests that to some degree, participants "traded off" myside and otherside arguments within a given prompt. Second, in general, the most substantial correlations in Table 6 are between myside-myside and otherside-otherside pairings; there are relatively few significant myside-otherside correlations (though ALC, the drinking age prompt, is an exception). This shows that the satisfactory internal consistency reliability of myside and otherside "scales" is not due to general positive

Table 6: Myside and Otherside Argument Correlations by Prompt (Study 1)

Arguments	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. TV-M																			
2. TV-O	***-.22***																		
3. GMO-M	**.59**	.04																	
4. GMO-O	.10	**.33**	*-.19*																
5. LDS-M	**.41**	-.06	**.35**	.15															
6. LDS-O	**.25**	**.34**	.04	**.38**	*-.17*														
7. TRI-M	**.43**	-.03	**.40**	.11	**.32**	.18													
8. TRI-O	.16	**.35**	.08	**.24**	.11	**.39**	***-.36***												
9. GG-M	**.54**	.06	**.73**	.14	**.36**	.12	**.40**	.06											
10. GG-O	**.26**	.18	.03	**.22**	.03	**.40**	.08	**.31**	***-.20***										
11. MJ-M	**.49**	-.06	**.55**	.05	**.47**	-.09	**.23**	.14	**.44**	.02									
12. MJ-O	.10	**.34**	.15	**.28**	-.13	**.48**	.16	.11	.14	**.33**	***-.34***								
13. SB-M	**.53**	.06	**.55**	.14	**.36**	.04	**.24**	**.23**	**.57**	-.12	**.41**	.07							
14. SB-O	.14	**.27**	.06	**.26**	.05	**.53**	.16	**.28**	.00	**.60**	-.05	**.33**	***-.22***						
15. KND-M	**.54**	.07	**.57**	.08	**.39**	.10	**.43**	.11	**.56**	.02	**.32**	.12	**.47**	-.01					
16. KND-O	.08	**.27**	-.06	**.28**	-.02	**.59**	.19	**.22**	-.04	**.49**	-.06	**.32**	-.08	**.52**	***-.27***				
17. ALC-M	**.41**	-.01	**.40**	.09	.16	.18	**.20**	**.24**	**.38**	.10	**.33**	**.25**	**.37**	.01	**.37**	.11			
18. ALC-O	**.41**	**.30**	**.34**	**.26**	**.42**	**.27**	**.38**	.10	**.39**	.18	**.29**	.13	**.22**	**.24**	**.37**	**.21**	***-.23***		
19. RX-M	**.59**	-.09	**.45**	.05	**.41**	.18	**.45**	.09	**.47**	.15	**.43**	.00	**.46**	.16	**.50**	.19	**.46**	.18	
20. RX-O	-.01	**.35**	.14	**.24**	.16	**.24**	-.01	**.25**	.15	**.29**	-.12	**.46**	-.04	**.38**	.06	**.30**	.03	**.32**	*-.15*

Note. N = 100. M = myside; O = otherside; TV = television violence; GMO = GMOs; LDS = learning-disabled students; TRI = trigger warnings; GG = greenhouse gas emissions; MJ = marijuana; SB = seatbelts; KND = all-day kindergarten; ALC = alcohol drinking age; RX = prescription drug costs. Bold indicates $p < .05$. Italics indicates correlation between myside and otherside arguments for the same prompt.

correlations between both types of arguments across all items, but rather to across-item consistency in the number of myside arguments on the one hand, and otherside arguments on the other. However, most of these correlations are not especially high (only 15 out of 190 correlations exceeded *r* = .50), indicating that there is substantial variability in the rank ordering of participants across items. (Note that it is not possible to correct these correlations for unreliability because there is no way to compute reliability at the item level. The items cannot reasonably be considered alternate forms, and they were only administered once, so there is no test-retest information available.)

Predictors of reasoning performance. Hypotheses 1a and 1b were specified in terms of zero-order correlations between predictors and individual reasoning items. I had predicted that item-level predictors would be more salient predictors of individual item-level performance (the number of myside and otherside arguments) than person-level predictors. Results are presented in Table 7 (myside) and Table 8 (otherside). Although the results do not show a clear pattern that supports the hypothesis across items, it is the case that opinion strength is significantly correlated with the number of myside arguments generated for five of the 10 prompts, and education level is correlated with otherside arguments for four of the prompts. Aside from a handful of cases, person-level predictors (the personality and ability measures) are not significantly related to either myside or otherside arguments at the item level.

Regression analyses predicting scale results. The acceptable internal consistency reliability indicates that it is reasonable to aggregate the results from the items into "scales" and examine the relationship between the available predictors and reasoning

Table 7: Correlations Between Myside Arguments and Predictors (Study 1)

Predictor	TV	GMO	LDS	TRI	GG	MJ	SB	KND	ALC	RX
Opinion strength	**.29**	**.23**	.06	**.21**	**.37**	.16	.16	.14	.15	**.27**
Topic exposure	-.01	.08	-.05	.08	.13	.04	-.04	.14	.02	-.07
Education	.04	.13	.07	**.24**	.18	.01	.17	.11	.16	.13
TIE	**-.30**	-.07	-.01	-.12	-.05	-.12	-.07	-.05	**-.27**	**-.28**
Dogmatism	.07	-.05	-.02	-.04	-.04	-.02	-.04	.02	.02	-.01
NFC	.19	.07	-.04	.05	.11	-.08	.17	.08	.02	.16
Conservatism	.02	-.09	.00	-.05	-.10	-.09	-.10	-.17	.06	-.04
g_c	-.02	**.20**	.11	**.22**	.13	.16	.18	.19	.07	.06

Note. N = 100. TIE = typical intellectual engagement. NFC = need for closure. g_c = crystallized intelligence . TV = television violence; GMO = genetically modified foods; LDS = learning-disabled students; TRI = trigger warnings; GG = greenhouse gas emissions; MJ = marijuana; SB = seatbelts; KND = all-day kindergarten; ALC = alcohol drinking age; RX = prescription drug costs. Entries in bold are significant at $p < .05$.

Table 8: Correlations Between Otherside Arguments and Predictors (Study 1)

Predictor	TV	GMO	LDS	TRI	GG	MJ	SB	KND	ALC	RX
Opinion Strength	-.08	-.13	.08	-.19	**-.27**	-.15	-.12	**-.31**	-.08	-.17
Topic Exposure	-.03	-.12	.07	-.07	-.12	-.11	**-.26**	-.11	-.04	.08
Education	**.25**	**.21**	.18	.08	.13	.16	**.26**	**.26**	.11	.10
TIE	.03	-.08	**-.27**	-.12	-.18	-.07	-.18	-.20	-.09	-.03
Dogmatism	-.03	-.04	-.01	.07	**-.24**	-.08	-.11	-.08	.07	-.14
NFC	.01	.02	.13	.09	-.09	-.04	-.05	.03	.17	.04
Conservatism	-.01	.13	.03	.13	.09	.11	.03	.11	-.07	-.04
g_c	.11	.09	.03	.01	.13	-.04	.12	.07	.05	**.21**

Note. N = 100. TIE = typical intellectual engagement. NFC = need for closure. g_c = crystallized intelligence . TV = television violence; GMO = GMOs; LDS = learning-disabled students; TRI = trigger warnings; GG = greenhouse gas emissions; MJ = marijuana; SB = seatbelts; KND = all-day kindergarten; ALC = alcohol drinking age; RX = prescription drug costs. Entries in bold are significant at $p < .05$.

performance aggregated across items. Aggregated results typically provide more stable estimates of the underlying trait being measured, increasing the relationship with other variables. Zero-order correlations between the five person-level predictors (TIE, dogmatism, need for closure, education level, and the crystallized intelligence composite) and the myside, otherside, total, and bias "scales" are presented in Table 9. The most salient relationship is a positive correlation between education level and the number of otherside arguments (r = .28; with total arguments, r = .26). TIE is also correlated with otherside arguments (r = -.20; with total arguments, r = -.25) although this relationship is in the opposite direction from what would be expected. The significant correlation between crystallized intelligence and the number of myside arguments (r = .18; with total arguments, r = .20) is in line with the report by Perkins et al. (1991) that more intelligent participants tended to generate more myside arguments. I also used multiple regression to assess the ability of more than one predictor to account for variance in the reasoning outcomes. This is consistent with the approach of researchers such as Furlong (1993),

Table 9: Person-level Predictors of Argument "Scales" (Study 1)

Predictor	Myside	Otherside	Total	Bias
TIE	-.19	-.20*	-.25*	-.07
Dogmatism	-.02	-.08	-.05	.03
NFC	.12	.07	.13	.08
g_c	.18*	.11	.20*	.11
Education	.17	.28*	.26*	.01

Note. N = 100. TIE = typical intellectual engagement. NFC = need for closure. g_c = crystallized intelligence.
*p < .05.

although for data with this structure it is not preferred (see HLM results, below). For these analyses, I included the three individual-level predictors with significant zero-order correlations (i.e., education, TIE, and crystallized intelligence).[11] I also included the mean amount of time spent per reasoning prompt in the multiple regressions.

Although time per item is not a true predictor in itself, it is important to consider this variable, given that most of participants completed the study in an unproctored setting. Because MTurk participants were compensated for completion, many likely were motivated to finish the study quickly. It is important to note that it is impossible to determine the direction of any relationship between time and number of responses in this dataset. A positive relationship between time spent and arguments generated may be due to participants hurrying through the protocol and not taking the time to think of more arguments, but it may also be the case that participants who would not come up with many arguments (even given an infinite amount of time) spend less time on the prompts because they run out of ideas more quickly. Despite this ambiguity, I included mean time as a predictor in order to address the methodological concern associated with the unproctored setting. In effect, mean time serves as a control variable that provides a more stringent test of the other person-level variables, in particular crystallized intelligence, which may also be influenced by the amount of time that participants are willing to spend on the study tasks. Including mean time in the model means that in order for other variables to be significant, they must account for variance when mean time is controlled for.

[11] I also ran multiple regression models including dogmatism and need for closure, which did not exhibit significant zero-order correlations with any argument outcomes. They were never significant predictors, so they are not included in the models reported here.

Results of the hierarchical regression analyses are presented in Table 10. As is evident in the table, mean time spent per prompt is the primary predictor of all four dependent variables. (Note that the overall model for bias was not significant, $R = .28$, $R^2 = .08$, *ns*, although the coefficient for time met the threshold for significance, $\beta = .25$, $t = 2.33$, $p < .05$.) Education was a significant predictor for all dependent variables except for bias. TIE and crystallized intelligence were not significant predictors in any model.

Hierarchical linear modeling to predict myside and otherside arguments. Hypotheses 1a and 1b stated that item-level predictors would correlate with the number of (H1a) myside and (H1b) otherside arguments. In two exploratory hypotheses, I predicted that person-level traits would be negligibly correlated with (EH1a) myside and (EH1b) otherside arguments. Taken together, these four hypotheses can be summarized conceptually as stating that item-level predictors are more important than person-level predictors when considering reasoning performance at the item level. This prediction can be assessed using HLM. HLM is preferable to standard multiple regression analyses used by previous researchers (e.g., Furlong, 1993) because each participant responds to multiple items, and therefore the item-level responses are not independent. In repeated-measures HLM, Level 1 is the item level, and Level 2 is the person level. These analyses were carried out using the MIXED command in SPSS Version 24.

In building the HLM models, I used the five-step, bottom-up exploratory approach outlined by Hox (2010). The empty model (no predictors) is computed in Step 1 to serve as a baseline. The fixed part of the model is then created by adding item-level predictors (Step 2) and person-level predictors (Step 3), and retaining predictors that are significant. Then, the random part of the model is built by testing random slopes for the

Table 10: Multiple Regression for Argument "Scales" (Study 1)

Model	*B* (*SE*)	β	*t*	*R*	R^2	Adj. R^2
Myside				.50*	.25*	.19*
Mean Time	6.20 (1.44)	.41	4.30*			
Education	1.21 (1.06)	.11	4.30*			
g_c	0.60 (0.64)	.10	0.93			
TIE	-2.10 (1.49)	-.14	-1.40			
Otherside				.41*	.17*	.14*
Mean Time	2.24 (0.87)	.26	2.57*			
Education	1.54 (0.64)	.24	2.41*			
TIE	-1.17 (0.90)	-.13	-1.30			
g_c	0.11 (0.37)	.03	0.31			
Total				.59*	.34*	.32*
Mean Time	8.44 (1.67)	.45	5.06*			
Education	2.75 (1.23)	.20	2.24*			
TIE	-3.26 (1.73)	-.17	-1.89			
g_c	0.71 (0.71)	.10	-1.89			
Bias				.28	.08	.04
Mean Time	3.96 (1.70)	.25	2.33*			
Education	-0.33 (1.25)	-.03	-0.27			
TIE	-0.09 (0.17)	-.06	-0.53			
g_c	0.49 (0.73)	.08	0.67			

Note. *N* = 100. TIE = typical intellectual engagement. g_c = crystallized intelligence.
*p < .05.

item-level fixed effects (Step 4) and cross-level interactions (Step 5). To control familywise Type I error rate at each step, I used Holm's (1979) procedure, recommended by Hox (2010), which is based on Bonferroni's correction but retains greater power than Bonferroni's method.[12] Only significant predictors are included in the tables, with the coefficients that were obtained after the non-significant predictors had been removed. I specified an unstructured covariance matrix in order to obtain intercept—slope covariances in Step 4 (Heck, Thomas, & Tabata, 2014), because forcing the covariance term to zero by leaving it out of the model is not recommended (Hox, 2010).

Power is difficult to determine in HLM (Hox, 2010; Snijers & Bosker, 2012). Hox (2010) suggested that a sample size of 100 Level 2 (in this case, person-level) entities with 10 observations each is sufficient for investigating the random part of the model. Therefore, it is reasonable to follow all steps of Hox's model-building method given the size of my final sample (N = 100, each responding to 10 prompts).

To facilitate the interpretation of the HLM results, the predictors were transformed to yield a meaningful intercept (Snijers & Bosker, 2012). For opinion strength (a 1—3 scale), one was subtracted from each score so that a score of zero corresponded to the "slightly agree/disagree" option. Education was represented using an ordinal variable in which zero represented a high school education. All other predictors were transformed to z-scores. This means that the coefficient for prior opinion reflects the amount of change in the number of arguments associated with endorsing "agree/disagree"

[12] In this procedure, the individual tests in a set of n tests are ordered by their p-values, from smallest to largest. The smallest p-value is compared against a threshold of α/n, the second-smallest value is compared against $\alpha/(n\text{-}1)$, and so on until a test fails to meet the threshold for significance (Holm, 1979).

as opposed to “slightly agree/disagree.” The coefficient for education represents the change associated with being in the next higher education category. For all other predictors, the coefficient reflects the amount of change associated with a 1-*SD* increase in the predictor when all other predictors are held constant. The arguments were left in their raw metric (i.e., counts of the number of arguments generated). Therefore, the coefficients can be interpreted directly in terms of the increase or decrease in the number of arguments generated (e.g., a coefficient of 1.00 indicates that the variable in question is associated with one additional argument being generated, compared to when that variable is at its zero/mean point).

Table 11: HLM for Otherside Arguments (Study 1)

Parameter	Model 1	Model 2	Model 3
	Fixed effects		
Intercept	0.85* (.09)	1.13* (.10)	0.79* (.15)
Item level			
Opinion Strength		-0.28* (.05)	-0.29* (.05)
Time		0.28* (.06)	0.28* (.06)
Person level			
Education			0.18* (.06)
	Random parameters		
Person level			
Intercept var.	0.61* (.11)	0.54* (.10)	0.49* (.09)
Item level			
Residual	1.51* (.07)	1.44* (.07)	1.44* (.07)
-2 log likelihood	3,411.48	3,359.81	3,351.18
R^2_1		.07	.09

Note. *N* = 100.
**p* < .05.

Table 12: HLM for Myside Arguments (Study 1)

Parameter	Model 1	Model 2
Fixed effects		
Intercept	2.99* (.15)	2.46* (.16)
Item level		
Opinion Strength		0.51* (.08)
Topic Exposure		0.24* (.07)
Time		0.54* (.08)
Random parameters		
Person level		
Intercept var.	1.99* (.32)	1.63* (.27)
Item level		
Residual	2.97* (.14)	2.64* (.12)
-2 log likelihood	4,129.99	4,006.59
$R^2{}_1$		.14

Note. N = 100
**p* < .05.

The appropriate effect size statistic for HLM is a matter of debate (Hox, 2010; Snijers & Bosker, 2012). In this dissertation, I follow the recommendation of Snijers and Bosker (2012) in using $R^2{}_1$ to indicate the percentage of Level 1 (item-level) variance that a model accounts for. $R^2{}_1$ is obtained by the equation $R^2{}_1 = 1 – (\sigma^2{}_{\text{model}} + \tau^2{}_{\text{model}})/(\sigma^2{}_{\text{empty}} + \tau^2{}_{\text{empty}})$, where σ^2 is the item-level residual variance, τ^2 is the person-level residual variance, the subscript “model” indicates the model in question, and the subscript “empty” indicates the empty model.

HLM results are presented in Table 11 (otherside) and Table 12 (myside). In both tables, Model 1 is the empty model with no predictors. This model provides a point of comparison for the fit of subsequent models, which is necessary because a model’s fit can only be assessed by comparing the fit of nested models (Snijers & Bosker, 2012). For otherside arguments (Table 11), prior opinion strength and time spent on each prompt were significant predictors (for Model 2, overall $R^2{}_1$ = .07). A 1-point increase in opinion strength was associated with a decrease in the number of otherside arguments (β = -0.28, *SE* = .05), whereas a 1-*SD* increase in the amount of time spent on a prompt (*SD* = 118 seconds, or about two minutes) predicted an increase in the number of otherside arguments generated (β = 0.28, *SE* = .06). Education status was the only significant person-level predictor (β = 0.18, *SE* = .06; see Model 3; overall $R^2{}_1$ = .09). The fit for Model 3 was significantly better than the fit for Model 2 (Δ-2LL = 8.63 with one additional parameter, $p < .05$; Δ $R^2{}_1$ = .02) and therefore is retained as the final model.

The final model for myside arguments is somewhat different. Topic exposure (β = 0.24, *SE* = .07) joins opinion strength (β = 0.51, *SE* = .08) and time per item (β = 0.54, *SE* = .08) as a significant predictor of myside arguments. Note that the sign of the opinion strength coefficient is reversed from the otherside model, so that there is a positive relationship between opinion strength and the number of myside arguments. There were no significant person-level predictors of myside arguments. Model 2 fit the data significantly better than Model 1 (Δ-2LL = 123.4 with three additional parameters, $p <$ .05; $R^2{}_1$ = .14), and serves as the final model.

Discussion

The main purpose of Study 1, aside from informing item choice for Study 2, was to examine item- and person-level predictors of reasoning performance. First I will briefly discuss the psychometric properties of the argument generation task, which influence the options for analysis and are important for interpreting the results. Then, I will discuss the results of the three different ways I examined the predictors of reasoning performance. I will focus primarily on the HLM results, because these analyses take into account the nested nature of the data.

Reasoning performance across items was fairly consistent, as evidenced by the acceptable internal consistency reliability of the myside, otherside, total, and bias "scales" and by the moderate positive pairwise correlations between performance on individual items. This is promising for research using this task, because it indicates that the 10 items tap the same construct. However, the pairwise correlations between items show substantial variability in responses at the item level. That is, performance is not so consistent that the items are interchangeable. Future researchers using this task are therefore advised to use multiple item prompts when possible.

There were few item- or person-level predictors that were significantly correlated with myside or otherside arguments on individual items. Opinion strength and education level were related to the number of myside and otherside arguments (respectively) for about half of the items. Other predictors did not show consistent relationships with item-level results. This means that neither item-level nor person-level factors were particularly good predictors of reasoning performance on individual items when the items were considered separately.

The acceptable internal consistency reliability meant that the items could be aggregated into scales, which could then be assessed for relationships with predictor variables. This reflects the analytical approach of most prior researchers who have examined individual differences in informal reasoning, albeit with fewer items. Using these scales, I computed pairwise correlations with the person-level predictors, and also conducted multiple regression analyses. Results from both of these analyses are similar to many previous reports (Furlong, 1993; Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003), in that ability and personality factors account for minimal variance in the number of myside and otherside arguments generated. The amount of variance accounted for by these predictors becomes non-significant when education level and mean item time are included in the regressions. In particular, these predictors account for almost no variance in bias, which has been the main dependent variable of interest for some researchers, particularly those using the argument generation task (e.g., Toplak & Stanovich, 2003). This supports the notion that bias, a difference score, perhaps should not be the primary outcome of interest in studies of informal reasoning, even though it is conceptually attractive. Use of bias as the primary outcome may contribute to the inconsistent results in previous literature regarding person-level predictors of informal reasoning performance.

The HLM approach provides a more robust picture of the predictors of reasoning performance, because it accounts for the nested data structure and allows simultaneous consideration of item-level and person-level predictors. Notably, different predictors were associated with myside and otherside arguments. Together, the predictors in the myside model accounted for a greater proportion of variance ($R^2{}_1 = .14$) than in the otherside model, ($R^2{}_1 = .09$). At the item level, topic exposure was associated with a greater number of myside arguments, but not otherside arguments, indicating that people who are more knowledgeable about a topic tend to preferentially list a greater number of myside arguments. Perkins et al. (1991) observed that their more intelligent participants seemed to use their intelligence to generate more myside arguments rather than explore the issue more thoroughly; to the extent that participants with a greater amount of relevant topic knowledge are similar to more intelligent participants, a similar effect may be occurring here. That is, participants may use the cognitive resources at their disposal (whether topic-specific knowledge or general intelligence) to provide as much support as possible for their own opinions. Another possibility is that the relationship reflects bias in the information that more knowledgeable participants have been exposed to. To the degree that people seek out or pay attention to information that supports their own opinions while they are learning about a topic, people who have gained more exposure to a topic may have been exposed to an increasingly biased subset of information regarding the issue. Again, the current dataset cannot address these questions, and these are speculations. However, it is notable that increased exposure to a topic is associated with a greater number of myside arguments, but not otherside arguments.

Also at the item level, opinion strength was a significant predictor of both myside and otherside arguments, but in different directions. For otherside arguments, a 1-point increase in prior opinion strength (on a 3-point scale) is associated with 0.29 fewer otherside arguments being generated. For myside arguments, a 1-point increase in opinion strength predicted approximately one-half additional myside argument. This result is consistent with Toplak and Stanovich’s (2003) finding that opinion strength is correlated with myside bias, but clarifies the magnitude of the effect on the number of otherside and myside arguments specifically.

The importance of opinion strength for argument generation is apparent when one considers the impact of opinion strength for both dependent variables simultaneously. For illustration, imagine two hypothetical participants. Participant A *slightly* agrees/disagrees with a prompt (score of 0 on the 0—2 scale), and Participant B *strongly* agrees/disagrees with the prompt (score of 2 on the 0—2 scale). Both hypothetical participants score at the mean for the other predictors (recall that the other predictors are *z*-scores, so their terms drop out of the equations). Participant A is expected to generate 2.46 myside arguments and 0.79 otherside arguments (i.e., the intercept term for the final model for each outcome). Participant B is expected to generate 3.48 myside arguments (2.46 + 2 x 0.51) and 0.21 otherside arguments (0.79 – 2 x 0.29). This translates into a mean myside bias of 1.67 for Participant A, and 3.27 for Participant B. This example illustrates the impact of opinion strength on the number of myside and otherside arguments generated under non-directive instruction conditions.

At the person level, there were no significant predictors of myside arguments. The only significant person-level predictor of otherside arguments was education status. Some prior researchers (Toplak & Stanovich, 2003), but not all (Furlong, 1993; Perkins, 1985a), have reported a relationship between education level and various outcomes on similar reasoning tasks. Increasingly higher levels of education may provide people with practice in dealing with complex topics like the ones used in this study. Education may also expose people to the notion that a strong set of arguments is one that considers rather than ignores opposing views, which may contribute to the positive relationship between education and otherside arguments observed here. No other person-level predictors were significant, such as crystallized intelligence or TIE. It is possible that the low average number of otherside arguments (*M* = 0.85 per item) played a role in the failure of other predictors to account for a significant amount of variance. In any event, this result is in line with the many previous studies that have failed to find robust individual differences predictors of informal reasoning quality.

To summarize: Study 1 confirmed the ubiquitous finding that people generate more myside arguments than otherside arguments under non-directive instructional conditions. Results from the HLM analyses suggest that a somewhat different set of predictors is associated with myside arguments than with otherside arguments. Education was a significant predictor of otherside arguments, topic knowledge was associated with myside arguments, and opinion strength was associated with both. Study 1 only addressed one of the two conditions used in the argument generation task (i.e., the non-directive condition), and was consistent with previous studies in requiring participants to use only the topic knowledge that they already possessed. However, results may be different if participants are given explicit instructions to consider both sides of the issue, and if they are allowed to refer to outside information when they are generating their responses. Study 2 addressed both of these possibilities by manipulating instruction type and Internet access.

CHAPTER 3
STUDY 2

Study 2 was designed to address two questions. First, what are the predictors of informal reasoning when participants are given explicit instructions to consider all sides of an issue? If Baron (1995; see also Stanovich & West, 2008) was correct in suggesting that participants generate few otherside arguments because they do not realize that otherside arguments are desirable, then instructions that specify the goal of the task may lead to additional person-level predictors being retained in the model. The second question is whether the ability to look up information on the Internet impacts reasoning in either of the instructional conditions. This provides a test of the suggestion that participants are biased because they have difficulty thinking of otherside arguments, and that they would list more otherside arguments if only they could think of them without expending substantial effort. It also provides a more realistic reasoning scenario in an age in which all of the information on the Internet is available to anyone with a smartphone. To address these two questions, Study 2 utilized the standard instructional manipulation that is typically used in the argument generation task, and also introduced a novel Internet access manipulation.

Method

In this study, four prompts (selected based on the results of Study 1) were used to assess hypotheses related to the two research questions: the effect of allowing access to outside information on reasoning outcomes, and changes in the relationships between predictors and reasoning outcomes under different reasoning conditions.

Power analysis. I conducted several power analyses using G*Power (Faul et al., 2009) in order to determine the sample size needed to detect the various hypothesized effects. The tests of dependent correlations (i.e., Hypotheses 6 and 8) required the greatest sample size (N = 100) in order to detect the anticipated effect of $\Delta r = .30$ (a change from about $r = .20$ to $r = .50$). Therefore, I aimed to recruit 110 participants with the goal of having at least 100 participants with complete and useable data.

Sample. One hundred nine Georgia Tech undergraduates (60 males, 49 females) from the School of Psychology participant pool participated in Study 2. Participants were required to be fluent in English and must not have participated in Study 1. Data from four participants were eliminated prior to analyses: one for obvious lack of effort, one for questionable English proficiency, one for questionable effort and English proficiency, and one for generating a very high percentage (63%) of uninterpretable arguments. This left a final sample of 105 (56 males, 49 females; 16 freshmen, 30 sophomores, 19 juniors, and 32 seniors). Participants were compensated with 3.0 hours of SONA credit.

Measures. Measures for Study 2 were similar to those used in Study 1, except for the reasoning task, which involved the two experimental manipulations already described.

Self-report questionnaires. Participants completed the same self-report questionnaires as in Study 1 (prior opinion, self-report knowledge, and involvement for each reasoning prompt; political conservatism, typical intellectual engagement, dogmatism, and need for closure). In order to obtain a finer-grained measure of opinion strength, the prior opinion measure was modified to a 0—100 point slider instead of a 6-

point Likert scale as was used in Study 1. Two additional measures were added for Study 2: anti-intellectualism and the Big Five personality traits.

Anti-Intellectualism. The Student Anti-Intellectualism Scale (Eigenberger & Sealander, 2001) was used to measure anti-intellectualism, which is a generally negative attitude toward intellectual activity. This 25-item scale was designed for use with a college student sample, and therefore was not used for the primarily non-student sample in Study 1. A sample item is "I don't like taking courses that are not directly related to my goals after college." (α = .91; Eigenberger & Sealander, 2001)

Big Five personality traits. Personality was assessed using the NEO-FFI (Costa & McCrae, 1992). This 60-item measure assesses the Big Five personality traits on a 6-point scale (1 = very untrue of me to 6 = very true of me). John and Srivastava (1999) reported Cronbach's α exceeding .70 for all five subscales in a sample of college students.

Cognitive ability tests. Cognitive ability was assessed using the three tests used in Study 1 (i.e., Extended Range Vocabulary, Analogy, and the general knowledge test). Two additional tests were added for Study 2 in order to assess logical reasoning, as described in the ETS Kit manual: "the ability to reason from premise to conclusion, or to evaluate the correctness of a conclusion" (Ekstrom et al., 1976, p. 141). These logical reasoning tests were not linked to any specific hypotheses, but were included in order to provide information about the nomological network of the argument generation task.

Diagramming Relations. The Diagramming Relations test (Ekstrom et al., 1976) consists of two parts of 15 items each, with a time limit of four minutes per part. Each item consists of a list of three classes of objects (e.g., "Animals, cats, dogs") and the respondent must select which one of five diagrams correctly represents the relationship between the three classes. Score is the number of items correct minus 0.25 of the number of items incorrect. (α = .90; Beier & Ackerman, 2005)

Inference Test. The Inference test (Ekstrom et al., 1976) consists of two, six-minute parts with 10 items each. Each item presents a scenario followed by five possible conclusions. Respondents must select which one of the conclusions can be drawn from the information in the scenario. Score is the number of correct responses minus 0.25 of the number of incorrect responses. (A prior study reporting reliability could not be located.)

Academic record data. SAT scores, college GPA, declared major, and number of college credits were obtained from the academic records of students who provided separate consent to the release of this information to the investigator.

Reasoning task. The argument generation task in this study was the same as the one used in Study 1, with two differences. First, there were only four prompts, selected from the 10 used in Study 1. Second, there were two instruction conditions (non-directive and directive instructions) and two Internet conditions (no Internet access and Internet access). The two manipulations were crossed to create four conditions. One prompt was presented in each condition. Instructions for each condition are presented in Appendix B. The order of the directive versus non-directive instruction manipulation was fixed (non-directive always was presented first), due to the very high likelihood that receiving the directive instructions first would result in carryover effects for non-directive instructions presented later (but see Furlong, 1993). The branching feature in Qualtrics was used to randomly direct participants into one of eight possible conditions which counterbalanced the order of the Internet conditions and the order of the prompts (and, by extension, the matching of the four prompts with each of the four conditions). The eight orders are provided in Appendix C.

At the beginning of the session, participants were told that they could advance to the next prompt when they were out of ideas for the current prompt. They were also told that the software would advance to the next prompt eventually to keep them on track to finish within the allotted time for the session, so they should feel free to take their time and not worry about falling behind. Participants were not told the specific amount of time after which the screen would advance, which was 18 minutes. Pilot testing indicated that very few participants spent this long on a single item. Nearly all participants in Study 2 self-advanced before 18 minutes had passed.

Access to the Internet was genuinely open (i.e., not constrained to an artificial set of sources as in many previous studies, e.g., Wolfe & Britt, 2008; Wolfe, 2012). The only limitation was that participants were required to use the search engine DuckDuckGo, which does not filter or otherwise adjust search results based on past activity (Wawro, 2013). It was important that participants use a search engine that did not “learn” the user’s preferences based on clickstream data, as is the case with many major search engines (e.g., Google). DuckDuckGo was set to the default search engine in Mozilla Firefox, and participants were told to use this search engine. Auto-complete suggestions were turned off, and browsing history was cleared before each session.

Items. The reasoning task used the following four prompts:

1. Should learning-disabled students be integrated into regular school classrooms?

2. Should college instructors provide “trigger warnings” when they are about to present class material or assign reading that might upset some students (e.g., material related to sexual assault, discrimination, child abuse), and allow students to leave the lecture or avoid the reading if they feel they might be upset by the material?

3. Should genetically modified foods (e.g., corn, fish) be produced and sold in the USA?

4. Should the amount of violence on television be restricted?

These items were selected based on results of Study 1. Three criteria were key in selecting these arguments from the original set of 10. First, the sample in Study 1 endorsed the full range of prior opinion regarding these prompts. Second, participants in Study 1 generated roughly the same number of pro and con arguments for these prompts, and approximately the same total number of arguments across the four prompts. A third requirement was that myside bias was evident, but in reality, all 10 prompts used in Study 1 resulted in substantial bias, so this requirement did not play a prominent role in the decision.

Instructions (e.g., whether Internet use was allowed) appeared along with the prompt and remained on the screen until the participant advanced to the next prompt. An embedded video containing an audio file of the instructions appeared at the top of each screen. Participants were told at the beginning of the session to play the video on each screen, and were reminded to do so by the experimenter when they did not. Each screen included 25 numbered text boxes. Participants were instructed to type each argument into a separate box. The instructions stated that they did not have to provide an argument for each box, but they were asked to come up with as many arguments as they could. No instructions were provided regarding the quality of the arguments.

Scoring. Two scoring schemes were used. The simpler scheme, in which arguments are scored as pro or con, has already been described for Study 1. The second, more complex scheme included a consideration of the quality of each argument and represents a novel contribution of this project.

Quality ratings were obtained from two raters using the following procedure. Arguments generated for each prompt were first sorted based on content to facilitate scoring by having all the similar arguments together.[13] For each of the four prompts, an initial calibration set, which contained approximately 50 representative arguments, was selected and scored by the two raters on a 60-99 scale. This scale was selected due to its intuitive relationship to the standard grading scheme in US high schools and universities. A rubric (Appendix D) was provided that described characteristics of arguments corresponding to each "letter grade" range (e.g., 90-99, 80-89, etc.). The descriptions of argument quality in the rubric were based on descriptions of scoring schemes used by previous researchers using reasoning tasks other than the argument generation task (Hahn

[13] Initially, arguments generated in Study 1 were sorted by two raters (an undergraduate research assistant and me) in order to develop a detailed coding scheme. The intent was to score each code, and assign the code's score to all of the arguments that had been given that code. However, I later determined that this step was at best unnecessary and at worst reducing the integrity of the scoring procedure due to the need to constantly make decisions about where to set category boundaries. Therefore, this step was discontinued, and the arguments in Study 2 were scored directly, as described in the main text above. However, the codebooks created from the Study 1 arguments were used to facilitate the content sorting of the Study 2 arguments, which helped the scorers by having the similar arguments presented together.

& Oaksford, 2007; Kuhn, Shaw, & Felton, 1997; Walton, Reed, & Macagno, 2008). The raters met to discuss the arguments on which they differed by greater than five points. The raters then completed a second calibration set of approximately 50 more arguments, and further discussed remaining points of disagreement before independently scoring the rest of the set for that prompt.

Table 13: Inter-rater Reliability for Argument Scoring (Study 2)

	Pro/Con	Sum		Average	
Prompt	Agreement (%)	Pro	Con	Pro	Con
GMO	91.9	.97	.97	.86	.88
TV	89.6	.99	.98	.95	.88
LDS	88.2	--	--	--	--
TRI	89.3	--	--	--	--

Note. Inter-rater reliability for quality was sufficient for GMO and TV prompts, so only one coder completed the final coding of the LDS and TRI prompts. Both coders completed the calibration part of the procedure for all items.

Inter-rater reliability for both scoring methods is presented in Table 13. The "Pro/Con Agreement" column is the percent of cases that received the same score of pro, con, or other across two the two pro/con raters (another graduate student and me). Discrepancies were resolved through discussion. Inter-rater reliability for the two quality raters (an undergraduate and me) are presented in the rest of Table 13. Because the dependent variable is the aggregated score of the myside and otherside arguments for each participant, the correlation between the aggregated scores obtained from the two raters is presented in the table. There are many possible ways to aggregate the quality scores. Prior literature has only used a count of the number of myside and otherside arguments, and therefore offers no guidance or precedent. I chose to use two aggregation methods: a sum of the quality ratings on each side, and an average quality rating for each side. As indicated in the table, the inter-rater reliability for the sum score was very high ($.97 < r < .99$). Inter-rater reliability for the average score also was acceptable ($.86 < r < .95$).

Procedure. As with Study 1, the main task of interest was the argument generation task. Instruction type (non-directive versus directive) and Internet access (access versus no access) were crossed in a 2 x 2 fully within-subjects ANOVA design. That is, all partcipants responded to all four prompts, one in each of the four experimental conditions. As noted above, the pairings of prompts and conditions were counterbalanced, as was the order of the Internet conditions.

Participants attended a single in-lab session, in groups of up to 11. After providing informed consent, participants completed the questionnaires, the reasoning task, and the ability tests. The entire protocol could be finished in under three hours. Participants were told that some components of the session were self-paced, but that the session would take the full three hours, so it was not to their benefit to rush through the tasks. They were also told that the reasoning task prompts would auto-advance eventually in order to keep them on track to finish on time, so they should not worry about falling behind. Nearly all participants completed the main protocol in under two hours and 45 minutes. Because it would be undesirable for participants to leave the session while others were still working, additional reasoning prompts were presented to fill the remaining session time for those who finished early. A few participants who were very fast completed all of these additional prompts, and were given a pencil-and-paper packet of filler tasks to work on until the session was over. Participants were given 5-minute breaks between the reasoning task and the ability tests, and between the ability tests and the additional reasoning prompts. All questionnaires and tests (except the filler task packet) were presented on the computer using Qualtrics. Screen capture video was recorded until the end of the reasoning task using Screenpresso.

Results

Study 2 was designed to address two general sets of hypotheses. The first set of hypotheses related to the experimental manipulations themselves. The main effect of the instructional condition might be considered a manipulation check rather than a true hypothesis, given the precedent for this result in the literature. However, the main effect of Internet access was examined for the first time in this experiment. The second set of hypotheses involved the interaction between individual differences variables and the experimental manipulations.

The dependent variables in this study were computed in three different ways. Counts of myside and otherside arguments were the main outcomes of interest and are reported first for all analyses. The quality ratings were aggregated two ways: by summing the quality ratings across myside and otherside arguments, and by taking the average quality rating of myside and otherside arguments. In most cases, the quality-based results mirror the Count results. Differences across these aggregation methods are noted where applicable.

In this section, I first describe overall results related to the argument generation task and the potential predictors. I then note an adjusted analysis plan due to missing data. In the third and fourth sections I present results from the experimental manipulations and the individual differences analyses. Finally, I describe the results of exploratory analyses aimed at further clarifying the relationships between individual differences and reasoning outcomes: HLM analyses similar to those used in Study 1; group-level comparisons of participants who did and did not use the Internet when it was offered; person-level predictors of the amount of time spent on each item; the relationship between argument generation task performance and two tests of formal reasoning; and the existence of "bias" in participants who report being neutral about the topic presented in the reasoning prompt.

Predictor variables. Descriptive statistics for the person-level predictors (including personality scales, ability tests, and SAT scores from the academic record) are presented in Table 14, and for item-level predictors in Table 15. Inspection of the top half of Table 15 shows variability in the means for the item-level variables across the four prompts. However, the bottom half of the table shows that these differences disappear when computed in terms of the experimental condition rather than the prompt topic, indicating that this variability is distributed across the four experimental conditions.

The SAT scores exhibited substantial restriction of range in this sample [SAT Verbal: *M* (*SD*) = 693.67 (78.15); SAT Math: *M* (*SD*) = 710.33 (64.03); SAT Writing: *M* (*SD*) = 666.29 (74.52)].[14] SAT Verbal scores showed moderate positive correlations with the ability tests administered in this study (.47 < *r* < .71), and a more modest correlation with the current events knowledge test (*r* = .36). These correlations provide support

[14] Means and standard deviations for all college-bound seniors in 2014 were: SAT Verbal: *M* (*SD*) = 497 (115); SAT Math: *M* (*SD*) = 513 (120); SAT Writing 487 (115) (The College Board, 2014)

Table 14: Descriptive Statistics: Personality and Ability (Study 2)

					Range	
Predictor	*M*	*SD*	# Items	α	Potential	Actual
Non-ability Questionnaires						
Conservatism	647.18	188.00	12	.81	0—1200	167—1088
TIE	53.65	8.85	12	.86	12—72	33—71
Dogmatism	48.70	13.81	20	.90	20—120	21—88
NFC	53.36	9.44	15	.81	15—90	27—74
Neuroticism	40.23	10.80	12	.86	12—72	13—69
Extraversion	47.62	8.61	12	.82	12—72	22—67
Openness	51.15	7.69	12	.72	12—72	32—66
Agreeableness	51.98	8.37	12	.81	12—72	26—71
Conscient.	51.64	9.74	12	.90	12—72	24—70
A-I	74.35	16.60	25	.89	12—72	28—129
Ability Tests						
Vocabulary	10.44	3.67	24	.80	-6—24	1.25—18
Analogy	32.59	4.67	50	.74	-12.50—50	13.25—42.25
Diag. Relations	23.33	5.60	30	.92	-7.50—30	3—30
Inference	14.72	3.21	20	.72	-5—20	4—22.50
CE Knowledge	20.03	5.41	40	.85	0—40	8—34
SAT Math[a]	710.33	64.03	[b]	[b]	200—800	520—800
SAT Verbal[a]	693.67	78.15	[b]	[b]	200—800	490—800
SAT Writing[c]	666.29	74.52	[b]	[b]	200—800	470—800

Note. *N* = 105 unless otherwise noted. Reliabilities for ability tests were computed using the correlation between two parts, corrected using the Spearman-Brown prophecy formula.

[a]*N* = 90. [b]Scores were obtained from academic records, so item-level statistics are not available. [c]*N* = 89.

Table 15: Descriptive Statistics: Item-level Predictors (Study 2)

					Range	
Predictor	*M*	*SD*	# Items	α	Potential	Actual
Reasoning Item Predictors by Prompt Topic						
Prior Opinion						
LDS	51.28	25.09	1	[a]	0—100	0—100
GMO	64.87	30.97	1	[a]	0—100	0—100
TRI	48.52	31.59	1	[a]	0—100	0—100
TV	35.59	29.19	1	[a]	0—100	0—100
Self-reported Knowledge						
LDS	3.27	1.65	1	[a]	1—8	1—8
GMO	4.46	1.82	1	[a]	1—8	1—8
TRI	4.00	1.72	1	[a]	1—8	1—8
TV	4.36	1.47	1	[a]	1—8	1—8
Involvement						
LDS	9.95	4.37	4	.82	4—24	4—24
GMO	14.59	5.20	4	.89	4—24	4—24
TRI	11.75	4.93	4	.87	4—24	4—24
TV	12.89	4.46	4	.85	4—24	4—24
Reasoning Item Predictors by Condition						
Prior Opinion						
Non-directive, No Internet	50.46	30.37	1	[a]	0—100	0—100
Non-directive, Internet	49.18	31.18	1	[a]	0—100	0—100
Directive, No Internet	52.18	31.52	1	[a]	0—100	0—100
Directive, Internet	48.44	31.30	1	[a]	0—100	0—100
Self-reported Knowledge						
Non-directive, No Internet	4.03	1.76	1	[a]	1—8	1—8
Non-directive, Internet	4.08	1.64	1	[a]	1—8	1—8
Directive, No Internet	3.87	1.77	1	[a]	1—8	1—8
Directive, Internet	4.10	1.75	1	[a]	1—8	1—8
Involvement						
Non-directive, No Internet	12.05	5.08	4	[b]	4—24	4—23
Non-directive, Internet	12.94	5.25	4	[b]	4—24	4—24
Directive, No Internet	12.51	4.77	4	[b]	4—24	4—24
Directive, Internet	11.68	4.98	4	[b]	4—24	4—24

[a] Single-item assessment. [b] Reliability not reported because items were counterbalanced across conditions.

regarding the validity of the ability tests. Normally when there is a restriction of range due to a selected sample, it is advisable to compute disattenuated correlations between the predictor (ability tests) and the criterion (reasoning task), via the test used for selection (SAT Verbal test). However, the SAT Verbal scores have extremely low correlations with the reasoning task performance ($r < .10$), which means that the difference between the raw and disattenuated correlations is almost zero. (The SAT Math and Writing tests have much smaller correlations with the other ability tests than does the SAT Verbal test, while also having correlations with the reasoning task that are not significantly different from zero.) Disattenuated correlations therefore are not presented here, but it must be noted that the sample was drawn from students at a selective university. As a result, these students may have performed better than an unselected young adult sample on the ability tests and/or on the argument generation task, which would reduce the observed

Table 16: Correlations Between Ability Tests (Study 2)

Predictor	1	2	3	4	5	6	7
1. Vocabulary							
2. Analogy	**.56**						
3. Diag. Relations	**.27**	**.55**					
4. Inference	**.35**	**.50**	**.51**				
5. CE Knowledge	**.28**	**.39**	.23	.06			
6. SAT Verbal	**.62**	**.48**	**.50**	**.49**	.18		
7. SAT Math	.07	.25	**.32**	.16	.08	**.35**	
8. SAT Writing	**.43**	**.28**	**.37**	**.40**	.01	**.72**	**.35**

Note. N = 105 except for SAT Math and Verbal (N = 90) and SAT Writing (N = 89). Diag. Relations = Diagramming Relations test. CE Knowledge = current events knowledge. Bold text indicates correlations significant at $p < .05$.

correlation between ability and reasoning performance. [The Diagramming Relations test in particular exhibited ceiling effects: *M* (*SD*) = 23.33 (5.60) on a test with a maximum score of 30]. Correlations between the ability tests, including the SAT scores, are displayed in Table 16.

As in Study 1, topic knowledge and involvement were combined into a single topic exposure variable. Extended Range Vocabulary, the Analogy test, and SAT Verbal score were combined into a single verbal factor composed of unit-weighted *z*-scores. The correlation between the two verbal tests administered in the study and the general current events knowledge test were notably lower than in Study 1,[15] so the knowledge test was not included in this factor.

Table 17: Descriptive Statistics for Myside and Otherside Arguments by Condition

	No Internet		Internet	
Condition	Myside	Otherside	Myside	Otherside
Non-directive	4.32 (2.81) 0—13	1.97 (2.06) 0—9	4.24 (3.09) 0—17	2.69 (2.65) 0—13
Directive	3.67 (1.97) 0—10	3.36 (2.33) 0—15	4.03 (2.58) 0—18	3.54 (2.47) 0—11

Note. Due to neutral responses on the prior opinion scale, the number of cases differs across cells. Non-directive, No Internet *N* = 92; Non-directive, Internet *N* = 90; Directive, No Internet *N* = 97; Directive, Internet *N* = 94.

[15] In Study 1, the correlations between the verbal tests and current events knowledge were *r* = .56 for both Extended Range Vocabulary and the Analogy test (which correlated *r* = .51 with each other). In Study 2, the correlations between the verbal tests and current events knowledge were *r* = .28 for Extended Range Vocabulary and *r* = .39 for Analogy (which correlated *r* = .56 with each other).

Argument generation task. Although myside bias is not the primary outcome of interest in this study, a general picture of the relative number of otherside and myside arguments will be helpful to frame the presentation of the results. Table 17 shows the mean, standard deviation, and range of the count of myside and otherside arguments in each condition. Substantial bias is evident, especially in the non-directive conditions. Table 18 shows the number of participants in each condition who listed more myside arguments than otherside arguments (positive bias), the same number of myside and otherside arguments (zero bias), and more otherside arguments than myside arguments (negative bias). A non- negligible number of participants in each condition displayed zero or negative bias. Overall, positive bias occurred in 56.0% of cases, while negative bias occurred in 28.7% of cases. Thus, although on average there were more otherside arguments than myside arguments generated in each of the four conditions, positive bias is by no means universal at the individual level.

Table 18: Direction of Bias by Condition (Study 2)

	Bias		
Condition	Positive	Zero	Negative
Non-directive, No Internet	67	8	17
Non-directive, Internet	52	11	27
Directive, No Internet	45	24	28
Directive, Internet	45	14	35
Total (Percent)	209 (56.0%)	57 (15.3%)	107 (28.7%)

Note. N = 105. Due to neutral responses to the prior opinion question, not all rows sum to 105.

Missing data and adjusted analysis plan. In the proposal, the main hypotheses were to be tested using repeated-measures ANOVAs. However, two factors limit the effective sample size in this study and demand an alternative analytical approach.

The first factor is that the instructions for the Internet condition allowed, but did not require, participants to look up information online when generating arguments. This was intentional, in order to avoid serious demand characteristics. Screen capture video from the argument generation task showed that 37 participants did not use the Internet for at least one of the two prompts for which Internet use was allowed. Of these, 15 did not use the Internet for either prompt. Further, six participants used the Internet for at least one prompt under no-Internet instructions.[16] Screen capture video was not available for one participant due to technical problems, so that participant's activity could not be verified. Thus, only n = 61 participants can be considered to have completed the experiment with full exposure to the experimental conditions.

The second factor limiting the effective sample size is related to an error in setting up the measurement of prior opinion. Prior opinion was assessed using a 0—100 slider scale. This was selected because Study 1 results indicated a relationship between otherside arguments and strength of prior opinion assessed on a 3-point scale (i.e., half of a 6-point scale) and a more nuanced scale was desired in order to investigate the effect further in this study. However, inexplicably it did not occur to me until after the data were collected that a 0—100 scale allowed participants to select the midpoint. This allowed them to effectively opt-out of the prior opinion question, because a response of

[16] Although the experimenter monitored participants' behavior, it was difficult to determine which condition participants were in when they were browsing the Internet on pages other than the Qualtrics site used for data collection.

50 meant that it was not possible to code arguments as myside or otherside. For the planned within-subjects ANOVA analyses, missing data requires listwise deletion. Thirty-eight participants (out of the full sample of 105) selected a response of 50 for at least one of the prior opinion items. Therefore, these participants would have to be excluded from a repeated-measures ANOVA, leaving a sample of $n = 67$.

When the effects of these two issues are combined (i.e., non-use of the Internet and selection of the neutral response for at least one of the prior opinion items), the sample size is reduced to $n = 36$ participants. Because this is a substantially smaller sample than intended, I assessed Hypotheses 2 through 4 using HLM instead of using the repeated-measures ANOVAs that were specified in the proposal. HLM does not require listwise deletion of cases with missing data, and allows a larger proportion of observations to be retained.

Table 19: Experimental Manipulations: Otherside Count (Study 2)

Parameter	Model 1	Model 2	Model 3	Model 4
		Fixed effects		
Intercept	2.89* (.17)	2.33* (.20)	2.12* (.21)	2.05* (.23)
Item level				
Instructions		1.09* (.20)	1.09* (.20)	1.21* (.25)
Internet			0.56* (.22)	0.74* (.30)
Inst.*Internet				-0.34 (.41)
		Random parameters		
Person level				
Intercept var.	2.04* (.44)	2.09* (.43)	2.04* (.42)	2.06* (.42)
Item level				
Residual	3.95* (.34)	3.57* (.31)	3.51* (.30)	3.50* (.30)
-2 log likelihood	1,679.09	1,650.38	1,643.63	1,642.98
R^2_1		.06	.07	.07

Note. $N = 105$.
$^*p < .05$.

Experimental manipulation results (Hypotheses 2 through 4). Hypotheses 2 through 4 were tested using HLM. As was the case with Study 1, Level 1 is the item level and Level 2 is the person level. To facilitate interpretation, the arguments were left in their original metric rather than transformed. Predictor variables were transformed to *z*-scores. (Note that unlike Study 1, prior opinion is a *z*-score due to the different scale used in this study.) In the interest of retaining as many observations as possible, while also reflecting actual participant behavior, I coded as Internet trials those trials in which the participants actually used the Internet, rather than those in which participants were told that they were allowed to use the Internet. As a result, there are fewer Internet trials than non-Internet trials.[17]

Results for otherside arguments (relevant to Hypotheses 2 and 4) are presented in Table 19. Model 1 is the empty model. Model 2 includes the main effect of instructions, which results in a significant improvement in model fit (Δ-2LL = 28.71, $p < .05$; R^2_1 = .06). This supports Hypothesis 2, which stated that the main effect of the instruction condition would result in a greater number of otherside arguments being generated in the directive instructions condition. The main effect of Internet use was added in Model 3, resulting in significantly better fit than Model 2 (Δ-2LL = 6.75, $p < .05$) but only a small increase in the proportion of variance accounted for ($\Delta R^2_1 = .01$). (I did not make a specific prediction about the effect of Internet condition on the number of otherside arguments.) The coefficients in Model 3 indicate that participants on average generated approximately one additional otherside argument in the directive instruction condition

[17] As noted above, in a few cases, participants accessed the Internet on non-Internet trials. Removing their data does not meaningfully change the results, so they are retained in the reported results.

compared with the non-directive instruction condition, and just over one-half of one additional otherside in the Internet condition compared to the no-Internet condition. Together, these manipulations accounted for 7% of the variance in the number of otherside arguments across conditions. Results for the sum of otherside arguments follow the same pattern as for count. For the average quality rating of otherside arguments, only the instructions term is significant. Models for sum and average otherside arguments are presented in Tables 20 and 21, respectively.

Hypothesis 4 also made a prediction about otherside arguments: that there would be an interaction between the instruction and Internet manipulations such that the Internet with directive instructions condition would yield the greatest number of otherside arguments. This hypothesis can be assessed in an HLM context by adding the instructions x Internet interaction to the model (Model 4 in Table 19). As can be seen in the table, the change in fit is minimal (Δ-2LL < 1) and the amount of variance accounted for by the interaction is not significantly different from zero (β = -0.34, *ns*; $\Delta R^2_1 < .01$).

Hypothesis 4 can also be assessed by using planned comparisons, since the nature of the interaction was specified *a priori*. As noted above, deleting cases with missing data (but retaining cases that did not use the Internet when it was allowed) resulted in a reduced sample of $n = 67$. In order to retain more participants in the sample, I conducted three paired *t*-tests with a Bonferroni correction, (directive, Internet condition compared to each of the others). This allowed for pairwise deletion rather than the listwise deletion that would be required for a Dunnett's test as a follow-up to a within-subjects ANOVA. For count data, the comparison with the non-directive, no Internet condition was significant ($t = 5.38$, $p < .05$, $d = 0.65$). No other comparisons were significant (non-

Table 20: Experimental Manipulations: Otherside Sum (Study 2)

Parameter	Model 1	Model 2	Model 3	Model 4
		Fixed effects		
Intercept	68.96* (4.03)	55.25* (4.86)	49.11* (5.25)	50.76* (5.68)
Level 1				
Instructions		26.69* (5.41)	26.66* (5.36)	23.48* (6.80)
Internet			16.54* (5.81)	12.12* (8.22)
Inst*Internet				8.61 (11.29)
		Random parameters		
Level 2				
Intercept var.	877.34* (239.36)	894.10* (232.6)	857.44* (225.43)	849.44* (224.52)
Level 1				
Residual	2,880.07* (246.62)	2,666.37* (228.23)	2,619.01* (224.25)	2,618.52* (224.25)
-2 log likelihood	4,106.08	4,082.61	4,074.59	4,074.01
R^2_1		.05	.07	.08

Note. N = 105.

**p* < .05.

Table 21: Experimental Manipulations: Otherside Average (Study 2)

Parameter	Model 1	Model 2	Model 3	Model 4
		Fixed effects		
Intercept	20.26* (0.63)	16.90* (0.84)	16.24* (0.94)	16.17* (1.04)
Item level				
Instructions		6.58* (1.10)	6.59* (1.10)	6.70* (1.39)
Internet			1.77 (1.16)	1.93 (1.65)
Inst*Internet				-0.32 (2.29)
		Random parameters		
Person level				
Intercept var.	6.24 (6.30)	8.62 (6.01)	8.22 (5.94)	8.24 (5.94)
Item level				
Residual	125.36* (10.73)	111.96* (9.58)	111.56* (9.55)	111.53* (9.55)
-2 LL	2,877.60	2,843.70	2,841.37	2,841.35
$R^2{}_1$		.08	.09	.09

Note. N = 105.

*p < .05.

directive, Internet: *t* = 1.36, *ns*; directive, no Internet: *t* = 0.40, *ns*). The same pattern of results was observed for the sum data. For average data, the comparison for the non-directive, Internet condition was significant as well. In short, Hypothesis 4 was not supported for any of the dependent variables.

Hypothesis 3 predicted a significant main effect of Internet condition on the number of myside arguments. Results for myside count data are presented in Table 22. Model 1 is the empty model. The main effect of instructions was added in Model 2, and the main effect of Internet was added in Model 3. The coefficient for instructions is significant, and denotes a decrease of nearly one-half of a myside argument in the directive instruction condition compared to the non-directive instruction condition. (I did not have a specific hypothesis about the effect of instruction type on the number of myside arguments.) The coefficient for Internet is not significant, and model fit is not significantly improved (Δ-2LL < 2) when Internet condition is added to the model. Note

Table 22: Experimental Manipulations: Myside Count (Study 2)

Parameter	Model 1	Model 2	Model 3	Model 4
Fixed effects				
Intercept	4.03* (.19)	4.26* (.22)	4.14* (.24)	4.31* (.25)
Item level				
Instructions		-0.44* (.22)	-0.44* (.22)	-0.77* (.28)
Internet			0.32 (.24)	-0.14 (.34)
Inst*Internet				0.90 (.46)
Random parameters				
Person level				
Intercept var.	2.42* (.52)	2.44* (.52)	2.46* (.52)	2.43* (.51)
Item level				
Residual	4.47* (.38)	4.41* (.38)	4.37* (38)	4.32* (.37)
-2 log likelihood	1,728.70	1,724.72	1,722.92	1,719.19
R^2_1		.01	.01	.02

Note. *N* = 105.
**p* < .05.

that all of these models account for a very small amount of variance ($R^2{}_1$ = .01 for Models 2 and 3; $R^2{}_1$ = .02 for Model 4). For sum and average myside data, neither instructions nor Internet is significant (see Tables 23 and 24). Therefore, Hypothesis 3 was not supported. For completeness, results associated with adding the instructions x Internet interaction are displayed as Model 4. The interaction is not significant for any of the dependent variables, and was not specified in a hypothesis.

Individual differences results (Hypotheses 5 through 8). Hypotheses 5 through 8 proposed interactions between certain covariates and the experimental manipulations. That is, the hypotheses involved testing whether the relationship between the predictor variables and reasoning performance varied, depending on the experimental condition. There were two general categories of predictor variables used in this study: those that vary both between and within individuals (i.e., prior knowledge, issue involvement, and prior opinion, all of which were assessed for each item prompt) and those that vary between individuals only (i.e., ability and personality measures).

Predictors varying between individuals only. Hypothesis 5 predicted that otherside arguments would be more strongly correlated with ability tests under directive instruction conditions (compared with non-directive instruction conditions), and Hypothesis 7 predicted that otherside arguments would be more strongly correlated with personality factors in under non-directive instructions (compared with directive instruction conditions). In other words, performance under directive instructions (maximal performance) would be determined by ability, whereas performance under non-directive instructions (typical performance) would be determined by personality. Results related to these two hypotheses are presented in Table 25. (Comparisons across Internet

Table 23: Experimental Manipulations: Myside Sum (Study 2)

Parameter	Model 1		Model 2		Model 3		Model 4	
Fixed effects								
Intercept	95.18*	(4.31)	99.64*	(5.26)	96.19*	(5.77)	100.61*	(6.21)
Item level								
Instructions			-8.71	(5.89)	-8.73	(5.86)	-17.22	(7.40)
Internet					9.27	(6.35)	-2.56	(8.95)
Inst*Internet							22.94	(12.29)
Random parameters								
Person level								
Intercept var.	1,031.22*	(276.32)	1,035.97*	(275.93)	1,053.11*	(277.04)	1,042.18*	(274.31)
Item level								
Residual	3,184.20*	(273.73)	3,159.89*	(271.65)	3,129.73*	(269.10)	3,101.38*	(266.65)
-2 log likelihood	4,146.87		4,144.69		4,142.57		4,139.11	
R^2_1			.01		.01		.02	

Note. N = 105.
**p* < .05.

Table 24: Experimental Manipulations for Myside Average (Study 2)

Parameter	Model 1	Model 2	Model 3	Model 4
		Fixed effects		
Intercept	23.25* (0.45)	22.68* (0.63)	22.13* (0.70)	22.13* (0.78)
Item level				
Instructions		1.12 (0.85)	1.12 (0.84)	1.12 (1.07)
Internet			1.47 (0.88)	1.47 (1.26)
Inst*Internet				0.01 (1.75)
		Random parameters		
Person level				
Intercept var.	2.47 (3.27)	2.74 (3.29)	2.36 (3.26)	2.36 (3.26)
Item level				
Residual	66.92* (5.73)	66.35* (5.69)	66.19* (5.69)	66.19* (5.69)
-2 log likelihood	2,639.35	2,637.60	2,634.83	2,634.83
R^2_1		<.01	.01	.01

Note. N = 105.
**p* < .05.

and no-Internet conditions are also presented for the sake of completeness, although they are not tied to specific hypotheses.) The hypotheses were tested using Williams' T_2 test of the difference between dependent correlations. None of these tests involving significant correlations were significant, and the hypotheses were not supported. Only count data are presented, but sum and average yield the same pattern of results.

Turning to the correlations themselves rather than the comparisons between them, only a few significant relationships were observed. Specifically, performance on the Extended Range Vocabulary test was significantly correlated with the number of otherside arguments. However, when this test was combined with the Analogy test and the SAT Verbal score to create a verbal composite consisting of unit-weighted *z*-scores, the relationship was no longer significant. It is unclear why this would be the case. One possibility is the difference in testing situations for the SAT Verbal test and the

Table 25: Comparisons of Correlations with Predictors Across Conditions

	Correlations			Correlations		
Predictor	General	Specific	T_2	No Internet	Internet	T_2
Ability Tests						
Verbal composite	.01	.11	0.77	.07	.05	0.17
Vocabulary	.23	.26*	0.26	.25*	.25*	0.00
Analogy	-.22	-.03	1.63	-.14	-.12	0.18
Diag. Relations	-.01	.20	1.80	.11	.09	0.18
Inference	.12	.20	0.68	.17	.16	0.09
CE Knowledge	-.10	-.06	0.33	-.04	-.12	0.73
SAT Math[a]	-.18	-.17	0.08	-.16	-.20	0.32
SAT Verbal[a]	.06	.09	0.23	.10	.06	0.31
SAT Writing[b]	.04	.06	0.17	.07	.04	0.25
Personality Questionnaires						
TIE	.34*	.34*	0.00	.43*	.26*	-1.70
Anti-intellectualism	-.28*	-.22	0.52	-.34*	-.17	1.64
Dogmatism	.01	.07	0.50	.10	-.03	-1.19
NFC	.08	.08	0.00	-.04	.20	2.27*
Neuroticism	-.02	.03	0.42	-.11	.13	2.26*
Extraversion	-.01	.03	0.33	.10	-.07	1.57
Openness	.11	.17	0.51	.16	.13	0.28
Agreeableness	.16	.22	0.51	.13	.27*	1.32
Conscientiousness	.24*	.21	0.26	.23	.23	0.00
Conservatism	-.07	-.11	0.33	-.10	-.09	0.09

Note. $N = 67$. Only participants with non-neutral opinions on all four prompts are included but using pairwise deletion instead does not substantially change the overall results. CE Knowledge = current events knowledge. Diag. Relations = Diagramming Relations test. TIE = typical intellectual engagement. NFC = need for closure.

[a] $N = 59$ because SAT scores were not available for all participants. [b] $N = 58$ because one participant had only SAT Math and SAT Verbal scores on file.

* $p < .05$.

Vocabulary test. The SAT Verbal score represents performance on a high-stakes test, whereas the Vocabulary test and the argument generation task were both administered in the lower-stakes experimental context. This may be why there is common variance between the Vocabulary test and the argument generation task, but not between the SAT Verbal score and the argument generation task. The divergence between results for the Vocabulary test and the Analogy test is more difficult to explain. It is possible that the Analogy test requires more cognitive effort than the Vocabulary test, and that under the relatively low-stakes experimental setting, performance on the simpler Vocabulary test better reflected verbal ability than did performance on the more complex Analogy test. Or, the relationship between the Vocabulary test and informal reasoning could be spurious. The existence of an actual relationship between otherside arguments and verbal ability therefore is questionable. Among personality traits, TIE and anti-intellectualism exhibited significant correlations ($.26 < |r| < .43$) with the number of otherside arguments in more than one condition. Conscientiousness was significantly correlated with otherside arguments only under the non-directive instructions condition ($r = .24$), while Agreeableness predicted the number of otherside arguments only when Internet access was allowed ($r = .27$).

Predictors varying between and within individuals. Hypotheses 6 and 8 involved item-level predictors. Hypothesis 6 predicted that the relationship between topic exposure and the number of myside arguments would be stronger in the Internet conditions than in the no-Internet conditions, because I predicted that higher-exposure participants would be better able to locate arguments online. Collapsing across conditions for item-level predictors does not make sense, so results for all four conditions are reported here: non-directive, no Internet (*r* = .23, *ns*), non-directive, Internet (*r* = .11, *ns*), directive, no Internet (*r* = .37, *p* < .05), directive, Internet (*r* = .21, *ns*). (Results for sum and average were the same, so only count data are presented here.) The only significant correlation is for a no-Internet condition, and none of the correlations are significantly different from each other (largest Steiger's $\bar{Z}_2*$ = 1.66, *ns*). Therefore, Hypothesis 6 is not supported.

Hypothesis 8 predicted a weaker relationship between the strength of prior opinion and the number of otherside arguments under directive instructions compared to non-directive instructions. As with Hypothesis 6, it does not make sense to collapse across conditions in this situation. The following correlations were observed: non-directive, no Internet (*r* = -.27, *p* < .05), non-directive, Internet (*r* = -.12, *ns*), directive, no Internet (*r* = -.05, *ns*), directive, Internet (*r* = -.11, *ns*). The only significant correlation was in the non-directive, no Internet condition, but this relationship was not significantly stronger than the correlations in the other conditions (largest Steiger's $\bar{Z}_2*$ = 1.22, *ns*). Hypothesis 8 strictly speaking is not supported, but the results are not inconsistent with the reasoning behind the hypothesis: that the strength of prior opinion exerts the greatest influence on otherside arguments in the least restrictive condition.

Exploratory Analyses. Given that few hypotheses were supported, I undertook a series of exploratory analyses to address some additional questions. First, I present results from a set of exploratory HLM analyses that predict myside and otherside arguments from the available predictors. These analyses extend the results reported above, moving beyond the experimental manipulations in order to consider the impact of both item-level and person-level predictors on reasoning outcomes. In the second set of exploratory analyses, I assess group differences between participants who did and did not use the

Internet when it was offered. This is a question of information-seeking behavior, and is not captured in the reasoning outcomes that were examined in the hypotheses or in the exploratory HLM results presented below. Identifying factors associated with not bothering to use outside information when allowed may indicate directions for interesting follow-up research on information-seeking behavior more generally. I also examine predictors of the amount of time spent per item in the four conditions, because this was a significant predictor in Study 1 and may have ability or personality correlates.

The last two sets of exploratory analyses address some methodological and interpretive concerns related to the argument generation task. The first concern is related to discriminant validity between formal and informal reasoning. I examine the relationship between the argument generation task and two tests of formal reasoning in order to check that the (lack of) relationship reported in previous studies was also the case in the present investigation. Finally, the methodological error of allowing a neutral response option on the prior opinion items provided an opportunity to examine the distribution of pro and con arguments generated by participants who claim a neutral stance on the reasoning items. Analysis of these cases provides additional context for interpreting the degree of myside bias reported in this and other studies.

HLM with person- and item-level predictors. As an alternative means of assessing the impact of individual differences and item-level variables on the number of myside and otherside arguments, I conducted additional HLM analyses using the Hox (2010) procedure (described in the explanation of the Study 1 results). For these exploratory analyses, I did not consider variables that did not exhibit any significant zero-order correlations with myside or otherside arguments, as these were unlikely to play a meaningful role in the model.

Results for otherside arguments are presented in Table 26 through Table 28. The final model for otherside count (Table 27) included three item-level predictors (instruction condition, time per item, and opinion strength; R^2_1 = .26). No person-level variables were significant. Random slopes were either non-significant, or models including them failed to converge in 1,000 iterations. There were no significant cross-level interactions. The final models for otherside sum (Table 27) and otherside average (Table 28) included the same predictors. Notably, in the otherside average model, the random term for the intercept is not significant, indicating that, when all predictors are at their mean, there is not a significant amount of between-person variability in the average argument quality score.

Table 26: Exploratory HLM: Otherside Count (Study 2)

Parameter	Model 1	Model 2
Fixed effects		
Intercept	2.89* (.17)	2.44* (.17)
Item level		
Instructions		1.04* (.19)
Time		1.03* (.12)
Opinion strength		-0.37* (.11)
Random parameters		
Person level		
Intercept var.	2.04* (.44)	1.19* (.30)
Item level		
Residual	3.95* (.34)	3.14* (.27)
-2 log likelihood	1,679.09	1,574.03
R^2_1		.28

Note. N = 105.
**p* < .05.

Table 27: Exploratory HLM: Otherside Sum (Study 2)

Parameter	Model 1	Model 2
	Fixed effects	
Intercept	68.96* (4.03)	57.94* (4.20)
Item level		
Instructions		25.67* (5.05)
Time		26.22* (2.96)
Opinion Strength		-10.03* (3.03)
	Random parameters	
Person level		
Intercept var.	877.34* (239.36)	440.60* (158.84)
Item level		
Residual	2,880.07* (246.62)	2,340.80* (200.38)
-2 log likelihood	4,106.08	4,005.77
R^2_1		.26

Note. N = 105.
*p < .05.

Table 28: Exploratory HLM: Otherside Average (Study 2)

Parameter	Model 1	Model 2
	Fixed effects	
Intercept	20.26* (0.63)	17.32* (0.82)
Item level		
Instructions		6.48* (1.07)
Opinion strength		-1.94* (.62)
Time		1.99* (.60)
	Random parameters	
Person level		
Intercept var.	6.24 (6.30)	7.43 (5.67)
Item level		
Residual	125.36* (10.73)	106.83* (9.16)
-2 log likelihood	2,877.60	2,824.01
R^2_1		.13

Note. N = 105.
*p < .05.

Results for myside arguments are presented in Table 29 through Table 31. The final model for myside count (Table 29) includes topic exposure and time per item at the item level. Agreeableness is the only person-level predictor (overall $R^2_1 = .21$). The effect of the instructional manipulation was not significant.[18] The final model for myside sum (Table 30) had the same item- and person-level predictors, and added a significant random effect of topic exposure. In keeping with the recommendation of Snijers and Bosker (2012), I retained the covariance term between the random intercept and the

Table 29: Exploratory HLM: Myside Count (Study 2)

Parameter	Model 1	Model 2	Model 3
	Fixed effects		
Intercept	4.03* (.19)	4.02* (.17)	4.01* (.16)
Item level			
Topic exposure		0.61* (.13)	0.59* (.13)
Time		0.89* (.14)	0.93* (.14)
Person level			
Agreeableness			-0.52* (.16)
	Random parameters		
Item level			
Intercept var.	2.42* (.52)	1.76* (.41)	1.50* (.37)
Person level			
Residual	4.47* (.38)	3.95* (.34)	3.96* (.34)
-2 log likelihood	1728.70	1670.45	1660.40
R^2_1		.17	.21

Note. N = 105.
$*p < .05$.

[18] When time per item was not included in the model, the Internet manipulation was significant for the otherside argument models, but not for the myside argument models. In other words, item time replaced the Internet manipulation as a predictor in the otherside argument models, but appears to be a predictor in its own right in the myside argument models. Given the significant mean difference between item time in the Internet and no-Internet conditions, it is not surprising that item time replaced Internet condition in the otherside models, but interesting that it appeared as a "new" predictor in the myside models.

Table 30: Exploratory HLM: Myside Sum (Study 2)

Parameter	Model 1	Model 2	Model 3
	Fixed effects		
Intercept	95.18* (4.31)	94.99* (3.84)	94.80* (3.69)
Item level			
Topic exposure		13.67* (3.35)	13.02* (3.31)
Time		21.39* (3.51)	22.38* (3.45)
Person level			
Agreeableness			-10.75* (3.70)
	Random parameters		
Person level			
Intercept var.	1031.22* (276.32)	714.63* (222.98)	596.94* (207.05)
Intercept—topic cov.			236.04 (150.62)
Topic var.			449.11* (205.50)
Item level			
Residual	3184.20* (273.73)	2892.88* (249.14)	2894.64* (249.14)
-2 log likelihood	4146.8	4097.00	4088.89
$R^2{}_1$		.14	.17

Note. N = 105.
*p < .05.

Table 31: Exploratory HLM: Myside Average (Study 2)

Parameter	Model 1
Fixed effects	
Intercept	23.25* (0.45)
Random parameters	
Person level	
Intercept var.	2.47 (3.27)
Item level	
Residual	66.92* (5.73)
-2 log likelihood	2639.35

Note. N = 105.
*p < .05.

random slope, even though the covariance term was not significant. The final model for myside average (Table 31) was the empty model itself (no item- or person-level predictors were significant).

To summarize the results of the exploratory HLM analyses: The item-level variables were the most salient predictors of both myside and otherside argument generation. Ability factors were never significant predictors, though this may be due in part to range restriction in the Georgia Tech sample. Results for the various aggregation methods were the same for otherside arguments, but differed for myside arguments. The only predictors of reasoning were instructions (otherside), opinion strength (otherside), topic exposure (myside), agreeableness (myside), and time per item (both).

Group differences between Internet users and non-users. To investigate possible differences between those participants who used the Internet both times it was offered (n = 61) and those who did not use it on one or both possible occasions (n = 37), I compared the two groups using independent t-tests. (The six participants who accessed the Internet during non-Internet prompts and the one participant whose screen capture video was lost to technical problems are not included in these analyses.) The Internet users scored significantly higher than non-users on Openness [users: M (SD) = 52.92 (7.41); non-users: M (SD) = 49.21 (7.87); $t = 3.71$, $p < .05$, $d = 0.49$]. The groups did not differ significantly on any other personality or ability measures.

Across conditions and dependent variables (myside and otherside count, sum, and average), the Internet users had higher scores than the non-users in a few cases. Internet users had higher otherside sum scores than non-users in the directive, Internet condition, and higher myside average scores in the directive, Internet and directive, no Internet conditions. This suggests that Internet use is associated with higher-quality arguments, although it is hard to draw a firm conclusion since the otherside results are only significant for the sum variable, and the myside results are only significant for the average variable. There is also one difference in a condition that would not be expected (the directive, no-Internet condition), which makes these results difficult to interpret. The groups did not differ in the total number of arguments generated in any of the conditions. As would be expected, the Internet-using group spent significantly more time on the Internet prompts than the non-Internet using group [Internet users: *M* (*SD*) = 651.19 (242.79) sec; non-users: *M* (*SD*) = 455.73 (213.50) sec; *t* = 4.04, *p* < .05], but the two groups did not differ on prompts for which Internet access was not allowed [Internet users: *M* (*SD*) = 437.71(225.88) sec; non-users: *M* (*SD*) = 445.73 (225.04) sec; *t* = 0.17, *ns*].

At the individual item level, cases in which participants did not use the Internet when it was allowed (*n* = 55) can be compared with cases in which they did (*n* = 153). Failure to use the Internet was not related to instructional condition (non-directive: *n* = 27; directive: *n* = 28). Strength of prior opinion was significantly higher in those instances in which participants did not use the Internet when it was offered [cases used: *M* (*SD*) = 24.52 (16.73) cases not used: *M* (*SD*) = 31.20 (16.56); *t* = 2.55, *p* < .05, *d* = 0.40]. No other item-level predictors were significant.

Predictors of time per item. The amount of time spent on individual reasoning items was a significant predictor of both myside and otherside arguments. This is not especially surprising, considering that people who spend a greater amount of time thinking about a prompt (or searching for more information about it on the Internet) are likely to come up with more arguments. If time per item is a major driver of argument generation, a reasonable question is: What predicts time per item? Table 32 presents correlations between predictor variables and time per item in each of the four conditions. Potential predictors that are not significantly correlated with time per item in any condition are not included in the table. TIE is positively correlated with time per item in all conditions, and anti-intellectualism is negatively correlated in all conditions (all correlations for these two predictors exceed $|r| = .25$, $p < .05$). Conservatism was significantly correlated with time per item in both of the no-Internet conditions. Performance on the Extended Range Vocabulary test is correlated with time spent on the items with non-directive instructions, but not the items with directive instructions (the difference between these correlations is not significant, however; largest $T_2 = 1.51$, $p < .05$). Overall, personality traits appear to be more robust and consistent predictors of time per item than the ability tests.

Table 32: Correlates of Time per Item by Condition

	General		Specific	
	No Internet	Internet	No Internet	Internet
Vocabulary	.21*	28*	.15	.19
Current Events Knowledge	.16	.22*	.17	.09
TIE	.37*	.29*	.31*	.29*
Anti-intellectualism	-.30*	-.29*	-.25*	-.29*
Conservatism	-.20*	-.12	-.22*	-.09

Note. $N = 105$. Only predictors with at least one significant correlation are included.
$^*p < .05$.

***Discriminant validity with formal reasoning*.** Two tests of formal reasoning, Diagramming Relations and Inference, were included in the test battery. Prior research (e.g., Macpherson & Stanovich, 2007) has not found informal reasoning to be substantially correlated with formal reasoning performance, so these tests were included for exploratory purposes related to clarifying the nomological network of the argument generation task, and were not related to any specific hypotheses. As expected, performance on the two formal reasoning tests did not correlate significantly with most indicators of informal reasoning performance. The sole exception was a significant correlation with the number of otherside arguments generated in the directive, Internet condition (Diagramming Relations: $r = .25$, $p < .05$; Inference: $r = .21$, $p < .05$). However, when Extended Range Vocabulary performance was partialled out of the relationship between the two inductive reasoning tests and the number of otherside arguments in this condition, the relationships were no longer significant. This means that the only observed relationship between formal reasoning performance and informal reasoning can be attributed to shared variance with a test of verbal ability, rather than to a direct relationship between formal and informal reasoning performance.

"Bias" in neutral participants. The methodological error of selecting a scale with a neutral option for prior opinion provided an opportunity to examine the directionality of arguments generated by participants who indicated neutrality on one or more of the prompts. Stanovich and colleagues (Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003) do not directly discuss expected results for people who claim to be neutral about a topic. Assuming, as Stanovich and colleagues do, that prior opinion is a primary driver of bias, a reasonable expectation is that neutral participants should generate approximately the same number of pro and con arguments. Across all participants and all prompts, there were 43 neutral ratings (i.e., ratings of 50 on a 0—100 point scale).

Table 33 presents the number of neutral ratings that occurred in each condition, along with the mean number of pro and con arguments generated by these participants. As is evident in the table, even neutral participants do not generate the same number of pro and con arguments, and in many cases show substantial "bias" (difference between pro and con arguments). In fact, the degree of "bias" displayed by these participants in the two directive instruction conditions is quite consistent with the actual bias (difference between myside and otherside arguments) shown by the rest of the sample in these conditions (all values between 0.30 and 0.49). For the non-directive instruction conditions, neutral participants show less "bias" in the non-directive, no Internet condition (-0.58 for the neutral participants vs. 2.35 for the rest of the sample), but somewhat more "bias" for the non-directive, Internet condition (-2.36 for the neutral participants vs. 1.55 for the rest of the sample).

Table 33: Descriptive Statistics of "Bias" Among Neutral Participants

	M (*SD*) Range			
Prompt	Pro	Con	Pro/Con "Bias"	Absolute "Bias"
Gen, No	3.00 (2.34)	3.58 (2.88)	-0.58 (1.62)	1.25 (1.14)
(*n* = 12)	1—10	1—11	0—4	0—4
Gen, Internet	3.21 (2.99)	5.57 (4.64)	-2.36 (3.88)	3.50 (2.79)
(*n* = 14)	0—11	-4—2	-10—4	0—10
Spec, No	3.14 (1.57)	2.71 (1.38)	0.43 (1.90)	1.29 (1.38)
(*n* = 7)	s1—5	1—5	-3—3	0—3
Spec, Internet	2.70 (2.40)	2.40 (2.07)	0.30 (1.89)	1.50 (1.08)
(*n* = 10)	0—5	0—6	-3—3	0—3

Note. Gen = general. Spec = specific instructions. No = no Internet condition.

It should be reiterated that this “bias” is simply pro/con bias, and is not tied to participants’ prior opinions, since they indicated that they had none. The mean bias therefore may be a conservative way of considering “bias” for these participants. One could also consider absolute bias (i.e., the difference without taking direction into account). Computed in that way, bias ranges from 1.25 in the non-directive, no-Internet condition to 3.50 in the non-directive, Internet condition. The argument here is not that absolute bias is a meaningful way to compare the neutral cases with the full sample—after all, the absolute bias assumes that all participants exhibit myside bias, and there is a not-insignificant number of cases of negative bias in the full sample. Instead, the point is that even participants who claim to be neutral do not generate an even number of arguments on both sides. While there are too few neutral cases to draw meaningful conclusions from statistical tests, these cases do suggest that prior opinion alone does not drive the tendency to generate more arguments for one side than the other.

Discussion

Study 2 contributed a novel manipulation to the informal reasoning literature based on the argument generation task: allowing participants to search for arguments on the Internet. This manipulation led to an increase in the number of otherside arguments generated by participants. The other manipulation, directive versus non-directive instructions, led to the expected increase in the number of otherside arguments in the directive instruction condition. In this section, I will review these results, along with results from the exploratory HLM analyses that included item- and person-level predictors of myside and otherside arguments. I also will consider the ramifications of the “bias” observed in neutral participants, and discuss the failure of the quality-based argument scoring to yield a clearer picture of reasoning performance than the standard scoring procedure of counting the number of arguments.

Based on prior research related to information search behavior (Whitmire, 2004; Willoughby et al., 2009; Wolfe & Britt, 2008), I anticipated that participants would use the Internet to find more myside arguments. This was not the case; Internet use was not a significant predictor of the number of myside arguments. Instead, use of the Internet was associated with approximately one-half additional *otherside* argument. Therefore, it does not appear that these participants used the Internet disproportionately to find additional support for their own positions. On the contrary, results suggest that, to a limited degree, participants used the Internet to further explore the area of the problem space (Perkins et al., 1991) that does not support their own opinion.

One interpretation of this result is that using the Internet made finding these arguments easier, allowing participants to locate more otherside arguments without expending more effort than they were willing to allocate to the task. However, the Internet by itself did not eliminate myside bias (which should not be surprising, given that explicit instructions do not eliminate it either, in this and previous studies; Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003). Although this study found a significant effect of Internet use on the number of otherside arguments, it should be noted that simply listing arguments is a relatively constrained and artificial task. The effect may be different for more complex tasks such as writing an essay or defending a position. In particular, tasks that require participants to integrate various lines of reasoning—a much more cognitively difficult task than simply listing relevant arguments—may exhibit a different effect than has been observed here with the argument generation task. It also must be noted that the effect size for Internet use was quite small, accounting for only 1% additional variance in otherside arguments beyond the instruction manipulation.

The second experimental manipulation, instruction type, was also a significant predictor of otherside arguments. Consistent with every relevant previous study (e.g., Furlong, 1993; Macpherson & Stanovich, 2007; Nussbaum & Kardash, 2005; Perkins, 1985a; Wolfe & Britt, 2008), directive instructions to consider both sides of an issue resulted in a less biased set of arguments being generated. In practice, this means that participants listed significantly more otherside arguments in response to directive instructions than to non-directive instructions. The present study also found that participants generated significantly fewer myside arguments in response to directive instructions.

At least two explanations are possible for this unexpected effect of instructions on myside arguments. The first is that participants actively reject some myside arguments in order to reduce the target number of otherside arguments that they must list in order to appear to consider both sides of the issue equally. The second possible explanation is that participants list the first few arguments that come to mind (which may be mostly myside arguments), and then as they begin a more effortful search for additional arguments, they specifically target otherside arguments. As a result, they “find” somewhat fewer myside arguments during this more effortful search phase. In either case, it appears that participants terminate their search prematurely, given that no condition has the highest number of both myside and otherside arguments, which suggests that in any condition, additional arguments could have been found. It is possible that people differ not only in their ability to generate or find arguments, but also in their willingness to continue searching for more. This question was addressed to a limited degree by the exploratory analyses related to the amount of time spent per item. Across the four conditions, the most consistent predictors of time per item were TIE (positive) and anti-intellectualism (negative). Although a full investigation of the factors leading to search termination for this type of task would require a separate study, TIE and anti-intellectualism are plausible candidates for playing a role in the amount of time spent per item.

The exploratory HLM analyses, which included item-level and person-level predictors of myside and otherside argument generation, present a somewhat fuller picture than the hypothesis-testing HLM analyses that included only the experimental manipulations. The only predictors of otherside arguments were item-level predictors: instruction condition, opinion strength, and time spent on the item. Collectively, these predictors accounted for 28% of the variance in the number of otherside arguments. It is not surprising that instruction condition remains significant in these models, given the large effect of instructions found in this and other studies (Furlong, 1993; Macpherson & Stanovich, 2007; Nussbaum & Kardash, 2005; Perkins, 1985a; Wolfe & Britt, 2008). The coefficient for time per item was nearly the same as that for instructions ($\beta = 1.04$ for instructions; $\beta = 1.03$, for item time), meaning that a 1-*SD* increase in time spent on an item (*SD* = 261 sec, or about 4 min and 20 sec) had about the same effect on the number of otherside arguments as directive instructions to think about the issue from opposite sides. The coefficient for opinion strength was smaller and in the opposite direction ($\beta = -0.37$), indicating that stronger opinions were associated with fewer otherside arguments. Toplak and Stanovich (2003) reported a small but significant positive correlation between opinion strength and myside bias for one of their three items; the present results clarify that the effect appears to be due to fewer otherside arguments being generated by those with stronger opinions. Models for the sum and average of quality ratings for otherside arguments included the same three predictors as the model for the count of otherside arguments.

The exploratory HLM results for myside arguments are somewhat different from the otherside models. Instruction condition and opinion strength are not significant predictors of myside arguments. Instead, topic exposure and time per item are the item-level predictors, and agreeableness is a person-level predictor. The interpretation of the time per item predictor is the same as for its role in the otherside model. Topic exposure (e.g., reading and talking about the topic) was associated with an increase of approximately one half of a myside argument per standard deviation increase in exposure score in the present study, which is consistent with results reported by Furlong (1993) using an almost-identical questionnaire. A relationship between exposure and myside arguments is not particularly surprising, considering that greater exposure to an issue likely leads to greater knowledge of arguments surrounding that issue. More interesting, though, is the fact that topic exposure did not play a significant role in the model for otherside arguments. Furlong (1993) also did not find a relationship between topic exposure and the presence of counterarguments (equivalent to otherside arguments) in an interview task. A relationship between topic knowledge and myside (but not otherside) arguments is not inconsistent with the observation by Perkins and colleagues (1991) that more intelligent undergraduates tend to generate a greater number of myside arguments, but not otherside arguments, compared to their relatively less intelligent peers. To the degree that more topic knowledge makes one "more intelligent" about that topic, perhaps a similar mechanism is at play.

Agreeableness was the only significant person-level predictor in any of the exploratory HLM models for Study 2, accounting for an additional 4% of the variance in myside arguments, beyond the item-level predictors. The NEO-FFI (with which Agreeableness was measured) was included for exploratory purposes and was not associated with *a priori* hypotheses. Agreeableness has not been associated with informal reasoning performance in the past, so this result should be interpreted with caution. However, a reasonable post hoc explanation is that more agreeable people may be reluctant to appear one-sided; perhaps this is more easily accomplished by listing fewer myside arguments rather than listing more otherside arguments (recall that Agreeableness was not retained as a significant predictor in the model for otherside arguments). Further research would be needed in order to firmly establish a link between Agreeableness and informal reasoning.[19]

I had predicted that cognitive ability would be related to the number of otherside arguments generated in the directive instruction conditions, and that personality factors would be related to performance in the non-directive instruction conditions. These predictions were based on the typical/maximal performance framework (Cronbach, 1990; Stanovich & West, 2008). Under maximal performance conditions, the goals of the

[19] A plausible explanation for the relationship between Agreeableness and myside arguments is that Agreeableness is related to prior opinion. For example, more agreeable people may hold less extreme opinions, or may be more likely to simply agree with the prompts. However, this does not appear to be the case: there were no significant correlations between Agreeableness and either opinion strength or the raw opinion scale (i.e., prior to folding it in half to obtain the strength indicator).

activity are clear and performance depends primarily on cognitive ability, assuming that individuals put forth maximal effort and perform as well as they are able. Performance under typical conditions is more strongly determined by individual differences in goals and motivation, sources of variability which are reduced in maximal performance conditions.

The extent to which typical/maximal conditions are present in any experimental situation can be difficult to assess. On the "maximal" side, it is questionable whether most (ethical) research designs are capable of eliciting maximal effort/performance to the same degree as real-life assessment situations, such as high-stakes standardized tests. On the "typical" side, participants in the current experiment may have felt obligated to search the Internet in response to the experimental manipulation, which may not reflect their behavior in unproctored settings. Thus, the current study does not exactly match the typical/maximal assessment settings as described by Cronbach (1990). However, the fact remains that the instructional manipulation clarified the goal of the activity, thus (presumably) reducing variance in performance resulting from differences in understanding of the goal of the task. In this way, the typical/maximal framework described here is more in line with Stanovich and West's (2008) specific instantiation of the idea in the context of informal reasoning, than with Cronbach's (1990) original description related to psychometric testing more generally. True assessment of "typical" informal reasoning likely would require naturalistic study designs in which participants are minimally aware of the goals of the research project, or even of their participation in it. This would be difficult from both logistical and ethical perspectives.

I had expected that separate considerations of performance in the experimental conditions would reveal relationships between person-level predictors of reasoning that had been masked by prior researchers' choice to collapse results across conditions (e.g., Furlong, 1993; Macpherson & Stanovich, 2007). However, this was not the case in either the planned tests of the hypotheses, or in the exploratory HLM analyses. Not only was there virtually no support for the specific hypotheses that the correlations between reasoning and ability/non-ability traits would be significantly different across instructional conditions, but there were few significant correlations at all. Of the ones that were significant, they were contrasted with correlations that barely missed the threshold for significance. Thus, in addition to failing to support the typical/maximal distinction hypotheses, these results also suggest that previous failures to link cognitive ability traits with reasoning performance are not due to aggregation across experimental conditions. Instead, the results of this study lend additional support to the notion that informal reasoning is not well predicted by traditional ability and personality factors.

Turning to the other exploratory analyses, the results related to the "bias" of neutral participants may challenge the traditional interpretation of the relative number of myside and otherside arguments. Although the issue typically has not been addressed directly, the theoretical approach of prior researchers (Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003) implies that a completely neutral reasoner should have minimal, if any, bias. In the present study, the participants who gave neutral opinion ratings provided an opportunity to examine this prediction in a preliminary manner.

The neutral participants were not neutral in their argument generation. Instead, on average, neutral participants exhibit positive bias (i.e., generate more pro arguments than con arguments). The difference between the number of pro and con arguments among the neutral participants was in the same range as the non-neutral participants' actual bias in the directive instruction conditions, and also consistent with myside bias reported by other investigators (Toplak & Stanovich, 2003). The pro/con "bias" represents bias assuming that all participants were supportive of the prompt; if the neutral option had not been available and participants had been forced to choose a rating on one side or the other, the degree of bias may have been somewhat higher than the pro/con bias. Although one could claim that these participants are not truly neutral on these topics, it is reasonable to assume that they have among the weakest opinions in the sample for the items in question. The fact that the majority of these participants essentially "fail" the argument generation task by not listing the same number of arguments for both sides suggests that prior opinion cannot be the only driver of myside bias, and raises the question of whether zero myside bias is a reasonable expectation for participants. However, it must be reiterated that there were not enough neutral cases for statistical tests, so it is possible that these results are an anomaly.

This study took the novel step of scoring arguments based on their quality. However, aggregation schemes for the quality ratings are associated with the same predictors as the raw count of otherside arguments, and nearly the same for myside arguments. The quality of arguments certainly varied substantially, ranging from compelling to trivial. One possible reason for the failure of the quality-based schemes to be associated with different predictors is that the task instructions asked participants to list arguments, not to list good ones. This was for consistency with previous uses of the argument generation task, because the main purpose of this study was to investigate the

Internet manipulation and the relationship between reasoning and individual differences under various conditions. Asking participants to list only good arguments would have changed the task considerably, as they would have had to set a standard for what constituted a good argument and then compare each argument with that standard. As Stanovich and West (2008) pointed out, cognitive ability is expected to be more strongly related to performance under maximal performance conditions, which by definition include a clear goal for the task. It is possible that cognitive ability is related to argument quality when the instructions are to generate high-quality arguments. The impact of task instructions on the relationship between cognitive ability and various task performance metrics has been demonstrated for simpler tasks, such as speed and accuracy in perceptual speed and psychomotor tasks (Ackerman & Ellingsen, 2016). It is plausible that instructions regarding argument quality would affect the relationship between argument generation and cognitive ability. However, the present project cannot address this question.

An alternative explanation for the lack of results obtained from the quality-based scoring methods is the possibility that, although two scorers were used, the scoring scheme was faulty. Although the rubric was based on prior work related to scoring participant arguments, it is possible that there is a disconnect between researchers' ideas of good arguments and laypeople's ideas of good arguments. This possibility could be addressed using a follow-up study in which a new sample of participants is asked to rate arguments that were generated by participants in this study, in order to create a scoring scheme that more accurately reflects laypeople's ideas of argument quality. Such a study could make an interesting contribution to the informal reasoning literature, but is well beyond the scope of the current project.

On the whole, there are four main takeaways from Study 2. First, use of the Internet increased the number of otherside arguments generated, but did not affect the number of myside arguments. This provides preliminary evidence that, at least for undergraduates doing an argument generation task as part of a research study, people do not use the Internet to simply search for evidence that confirms their prior beliefs. Results may be different in other populations and/or with more complex tasks, such essay writing or decision-making, and additional research would be needed in order to investigate this effect further. Second, person-level predictors did not exhibit consistent patterns of correlations with myside or otherside arguments generated in the four conditions, nor did they play a role in the exploratory HLM models for myside and otherside arguments (aside from Agreeableness in the myside model). This suggests that if there are person-level correlates of informal reasoning, they are not among the ones assessed in this study. However, some individual characteristics (particularly TIE and anti-intellectualism) were related to the amount of time spent on an item, which was a significant term in both the myside and otherside argument models. Future studies may be designed to investigate the relationship between individual predictors and the decision to stop searching for additional arguments, and the relationship between the stop decision and the number of myside and otherside arguments generated.

Third, the “bias” among the neutral participants raises questions regarding the reasonableness of holding participants to an ideal of listing exactly the same number of arguments for both sides of an issue, and considering any deviation to be indicative of faulty reasoning. While it is true that the argument generation task is a relatively simplistic measure of informal reasoning, this observation points to larger questions regarding how best to quantify high-quality informal reasoning, which must be addressed with the use of any informal reasoning task. Finally, it does not appear that considering the quality of the generated arguments meaningfully impacts conclusions about individual-level predictors or clarifies the relationship between reasoning and ability or personality factors.

CHAPTER 4
GENERAL DISCUSSION

This set of studies is the first to use HLM to investigate the effects of item-level and person-level predictors of reasoning performance in an argument generation task. Whereas prior researchers have relied on ANOVAs and zero-order correlations between predictors and reasoning indicators, HLM allows multiple levels of predictors to be considered simultaneously. This provides a more robust analysis and a clearer picture of the influence of the various possible predictors. The two studies in this project also addressed methodological questions relevant to the argument generation task and to informal reasoning research more generally.

Predictors of Reasoning Performance

Results from this set of studies were consistent with previous research (Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003; Wolfe & Britt, 2008) in finding that the number of otherside arguments was substantially smaller than the number of myside arguments. The present studies extend our understanding of the predictors of myside and otherside arguments by providing HLM analyses for both outcomes. The exploratory HLM analyses for both studies revealed that item-level variables were the primary predictors of both myside and otherside arguments. Topic exposure was a significant predictor of myside arguments in both studies, and opinion strength was a significant (negative) predictor of otherside arguments in both studies, as well as myside arguments in Study 1. Time per item predicted both reasoning outcomes in both studies; time, in turn, was associated with TIE and anti-intellectualism in Study 2. In Study 1, the only correlate of time per item was the vocabulary test. It is possible that due to the nature of the MTurk sample, with its unproctored setting and financial incentive, other factors (such as overall speed) influenced time per item, compared to the Georgia Tech sample used in Study 2.

An unexpected finding was that in the planned hypothesis testing for Study 2, Internet access was associated with an increase in otherside arguments, and did not affect myside arguments. This is counter to expectations but promising from a reasoning perspective, because it indicates that participants used the Internet to identify additional arguments that run counter to their own views. Whether participants sought out these otherside arguments or just managed to stumble across them is a question for another study. Regardless of their intent, participants did add (slightly) more otherside arguments to their lists when they used the Internet, which indicates that they did not simply pass over or avoid arguments that contradicted their opinions. Instead, it appears that participants who used the Internet allowed themselves to be exposed to a wider range of arguments, rather than focusing on finding arguments that supported their own position.

The argument generation task cannot assess whether participants took the additional otherside arguments seriously or whether they would have integrated these arguments into their views on the topics in any meaningful way. However, it is encouraging that in a setting with open access to the wealth of information on the Internet, participants appear to consider a wider range of arguments than when they do not have such access. This is an important addition to previous research, which has found that participants consider more otherside arguments when they are provided with a limited set of sources to peruse (e.g., Wolfe & Britt, 2008). Participants in an experimental setting may be more likely to access all available information when they have enough time to do so. It is therefore important to find that participants locate and list otherside arguments when they are able to access the nearly endless amount of information that is available on the Internet.

Methodological Considerations

From a methodological perspective, this set of studies demonstrated two things that should be considered in the design of future experiments. First, Study 1 showed that, although the generation of arguments is positively correlated across prompts, the rank ordering of participants' performance across prompts is not so consistent that prompts can be considered interchangeable. This appears to be due to the fact that item-level variables, such as prior opinion and topic exposure, were the primary predictors of both myside and otherside arguments in both studies. The positive zero-order correlations between performance on individual items in both studies, as well as the significant intercept variance in the HLM models, demonstrated that performance is not *entirely* due to item-level predictors. However, if the goal is to identify individual-level characteristics that predict reasoning performance, it is advisable to use multiple prompts, probably more than typically have been used in this field. Another point of note, also relevant to item consistency, is that results of Study 1 demonstrated substantial differences in prior opinion and in the number and pro/con distribution of arguments generated across prompts. Unless there is a specific reason for a researcher to desire a prompt with an uneven or non-centered distribution of prior opinion, it is important to pilot item prompts in a sample drawn from the population of interest in order to ensure that the distribution is acceptable for purposes of the research.

The second methodological point is that the value of considering argument quality in the scoring scheme for the argument generation task appears to be minimal, at least with the version of the argument generation task used here. Assessing argument quality was not the main goal of the present study, and in order to maintain consistency with previous studies, the argument generation task instructions did not give any guidance about the quality of arguments to be listed. It is possible that a version of the task with specific instructions to include only arguments of a certain quality would yield different results. Whether individual traits predict the ability to generate high-quality arguments is an interesting question in its own right, but cannot be adequately addressed in the present study. Based on the present results, use of quality-based scoring with the traditional argument generation task is not advisable.

Limitations

The results of these two studies must be interpreted in light of several limitations. First, the argument generation task is a relatively simplistic task, and cannot address questions such as whether or how participants would integrate arguments to form their opinions or to build a case. Inclusion of otherside arguments in a list does not mean that participants take them seriously. However, being aware of such arguments is the first step in integrating them into one's opinion on a topic, and this set of studies was designed as a first step toward understanding the impact of open access to outside information on the types of arguments that people consider.

A second limitation is the samples used in this study. The MTurk sample used in Study 1 had the advantage of being a non-student sample, but the disadvantages of being unproctored and, perhaps, in a hurry to finish in order to obtain compensation. The

Georgia Tech student sample used in Study 2 likely suffered from range restriction on the ability tests and/or the reasoning task, which works against efforts to relate ability to reasoning performance. A third limitation is the somewhat structured nature of the Internet search task in Study 2. Participants may have felt obligated to search the Internet when it was allowed (although many declined this opportunity at least once), but may not bother to seek out information when facing real-life reasoning situations. Also, adding to the artificiality of the task, efforts were made to present the same results to everyone by requiring participants to use DuckDuckGo.com as their search engine, and by clearing the browser history after each participant. Commonly used search engines such as Google "learn" user preferences and adapt search results accordingly (Pariser, 2011). Over time, users of such tools may be less likely to encounter otherside information without intentionally seeking it out. This effect may be more pronounced for individuals with stronger views on various topics, who are already more inclined to list fewer otherside arguments. The present study cannot address this interesting and timely question.

Conclusion

The argument generation task and other similar tasks typically paint an unfavorable picture of informal reasoning. Previous researchers have reported substantial myside bias that is mitigated, but not eliminated, by explicit instructions to consider both sides of an issue (Furlong, 1993; Macpherson & Stanovich, 2007; Toplak & Stanovich, 2003; Wolfe, 2012). Although many researchers have attempted to find individual-level predictors of informal reasoning performance, results have been inconsistent. This reality has been disappointing, given that the kinds of problems used in informal reasoning tasks may better reflect real-world reasoning scenarios, compared to the more artificial formal reasoning tasks that typically exhibit significant relationships with other cognitive ability tests (Galotti, 1989).

The present set of studies helps to clarify the factors that are associated with informal reasoning on an argument generation task. Using HLM to consider item-level and person-level predictors simultaneously, I found that item-level predictors emerged as the main drivers of variability in both myside and otherside arguments. Opinion strength predicted both myside (positively) and otherside arguments (negatively); it also was associated with the likelihood that participants would use the Internet when it was available. That is, opinion strength is associated not only with the distribution of arguments that a person generates, but also with the likelihood that he or she will forego the opportunity to use outside sources when creating the list of arguments. Interestingly, topic exposure—which was measured using questions that queried general engagement with the topic, such as following news about it—did not exhibit the same pattern of results. Instead, topic exposure was only related to an increased number of myside arguments. Further research on the relationship between opinion strength, topic exposure, and informal reasoning is needed to clarify the ways in the three factors relate to each other, including how their relationships may change over time as opinions develop and solidify.

Conclusions about person-level predictors are more difficult to draw. Aside from low but significant correlations with the vocabulary test in Study 1, ability measures did not correlate with reasoning performance. It is possible that item-level predictors simply are more important than person-level ones, especially for otherside arguments. However, the moderate positive correlations between performance on individual items, as well as the acceptable Cronbach's α levels observed for the myside and otherside "scales," suggest that item-level variables are not the only factors involved. The restriction of range in ability in the Study 2 sample likely hindered efforts to identify person-level variables related to the argument generation task. Future researchers seeking to establish a relationship between traditional ability tests and informal reasoning likely would benefit from using samples not drawn from selective universities. It may also be useful to administer several different informal reasoning tasks (as opposed to multiple items of the same task, as was done in this study), in order to investigate the nomological network of these tasks and perhaps clarify possible relationships between them and more traditional ability tests.

Results of this project also raise questions about the appropriateness of using myside bias as a reasoning outcome, given that most participants who rate themselves as neutral do not generate an unbiased list of arguments. Future researchers may wish to reconsider the best way to score this task, or select a task that provides a more complete picture of informal reasoning ability than can be provided by having participants develop a list of bullet points. Taking argument quality into account may be one approach, provided that participants are told to only list high-quality arguments. It may also be the case that more complex tasks, which can provide more nuanced and multifaceted snapshots of reasoning performance, would be preferable to the relatively simplistic argument generation task. As many previous researchers have discovered, however, scoring such tasks is often difficult (e.g., Sadler & Zeidler, 2005; Wolfe & Britt, 2008). On this front, at least, the argument generation task may provide a reasonable starting point for trying out new directions of research in informal reasoning.

This pair of studies represents an attempt to consider a variety of possible predictors of informal reasoning as measured by the argument generation task. Although there was not clear evidence for individual-level predictors of myside and, especially, otherside arguments, these two studies do demonstrate the importance of item-level predictors for informal reasoning performance. Results also suggest that personality factors such as TIE and anti-intellectualism are related to the amount of time that a person spends searching for arguments, whether online or in one’s head. Future research on these and other correlates of reasoning performance may aid in understanding how topic-related opinions and knowledge develop and how they influence people’s ability to reason about complex issues.

APPENDIX A

Argument Generation Task Instructions (Study 1)

The following instruction screen appeared before participants continued to the first reasoning prompt:

> “On each of the next several screens, a question is listed at the top of the screen, followed by several text boxes. Please read each question and take your time to think about it. Your task is to generate as many arguments as you can about each issue. Type each argument into a separate text box below each question.
>
> “You may take up to 10 minutes to think about and respond to each question. *Prior research with questions like these shows that people often come up with more ideas after they’ve been thinking about a question for a few minutes, so please take your time to think.* You can move on to the next page when you are really out of ideas, but try to come up with as many arguments as you can. In addition, it is important that you have read these instructions so that you know what you are supposed to do. To make sure that you have read these instructions, in the text box at the bottom of this screen, you should ignore the question and instead type “I have read the instructions.” This is important because we need to know that you know what you are supposed to do.
>
> **“Please DO NOT look up any information about these issues, or ask for anyone else's input.** We want to know what YOU can come up with just by thinking about them. **If you consult outside sources, you will invalidate the results.”**

Each argument was presented on a separate screen, along with 25 open-response text boxes and the following text:

> “Type as many arguments you can think of for the question below. Enter each argument into a separate box.
>
> “It is okay if you do not have an argument for each box . We don’t expect people to generate 25 arguments. We just don't want anyone to run out of space.”

APPENDIX B

Argument Generation Task Instructions (Study 2)

Component	Condition	Text
Instructions	Non-directive	Please read the following question and take your time to think about it.
	Directive	Please read the following question and take your time to think about it from opposite sides.
Prompt	All	[See Appendix A for item texts.]
Instructions	Non-directive	Your task is to generate as many arguments as you can about this issue. Type each argument into a separate box below.
	Directive	Your task is to generate as many arguments as you can about this issue. Type each argument into a separate box below. We would like you to put aside your personal beliefs on this issue. Write as much as you can, and try to give reasons both for and against.
Internet	No Internet	You may not use the Internet or any outside sources when answering this question.
	Internet	For this question, you may use the Internet to look up any information you wish. Please use Firefox as your browser. If you use a search engine, please use DuckDuckGo.com. Do not use Google or other search engines (they have been blocked on this computer).
Instructions	All	It is okay if you do not have an argument for each box. You can move on to the next page when you are out of ideas, but try to come up with as many arguments as you can.

APPENDIX C

Order of Reasoning Conditions and Prompts (Study 2)

Order	Number	Instruction	Internet	Prompt
1	1	Non-Directive	No	LDS
	2	Non-Directive	Yes	TRI
	3	Directive	No	GMO
	4	Directive	Yes	TV
2	1	Non-Directive	No	GMO
	2	Non-Directive	Yes	LDS
	3	Directive	No	TV
	4	Directive	Yes	TRI
3	1	Non-Directive	No	TRI
	2	Non-Directive	Yes	TV
	3	Directive	Yes	LDS
	4	Directive	No	GMO
4	1	Non-Directive	No	TV
	2	Non-Directive	Yes	GMO
	3	Directive	Yes	TRI
	4	Directive	No	LDS
5	1	Non-Directive	Yes	LDS
	2	Non-Directive	No	TRI
	3	Directive	Yes	GMO
	4	Directive	No	TV
6	1	Non-Directive	Yes	GMO
	2	Non-Directive	No	LDS
	3	Directive	Yes	TV
	4	Directive	No	TRI
7	1	Non-Directive	Yes	TRI
	2	Non-Directive	No	TV
	3	Directive	No	LDS
	4	Directive	Yes	GMO
8	1	Non-Directive	Yes	TV
	2	Non-Directive	No	GMO
	3	Directive	No	TRI
	4	Directive	Yes	LDS

APPENDIX D

Quality Scoring (Study 2)

Arguments in Study 2 were scored for quality using the rubric below. To facilitate scoring, the rubric was arranged to mirror academic grades (A through D, corresponding to numerical scores of 60—99). To obtain a final score prior to analysis, 59 was subtracted from each score to create a 1—40 point scale (excluded arguments received a score of 0).

Arguments could be divided into two general types: “would” arguments, which focus on consequences of proposed actions, and “should” arguments, which focus on assertions regarding whether or not an action is right, without specific reference to consequences. For the prompt about restriction of violence on television, an example of a “would” argument is “It would reduce bullying in schools.” An example of a “should” argument is “We should not promote violence in society.” Although arguments were not explicitly coded as “would” or “should” arguments, these two types did require somewhat different guidance for quality scores. The general guidance for “would” arguments was based on two related questions that have been found to be factors in laypeople’s judgments of argument strength (Hahn & Oaksford, 2007): How likely is the stated outcome, and how important (favorable/aversive) is it? Arguments citing more likely outcomes are considered more compelling than arguments citing highly unlikely outcomes. Similarly, outcomes that would be very favorable or very aversive are more compelling than ones whose positive or negative impact would be minimal. Secondary considerations for quality ratings of “would” arguments were whether the stated outcome could be accomplished by other means, and whether a stated problem could be avoided by making a minor amendment to the proposal, both of which lowered the quality rating of the argument (for a scoring scheme that emphasizes these two considerations, see Kuhn et al., 1997).

There is less guidance in the literature for evaluating the strength of “should” arguments. Because the “should” arguments explicitly or implicitly appeal to individual or societal ideals, the criteria for these arguments relate to these ideals: How well-established is the underlying principle being invoked, and how clearly is it related to the prompt? Although the degree to which an ideal is well-established is open to a greater amount of interpretation than other components of the scoring scheme, in reality it was possible to rank-order the ideals invoked. For example, many of the TV violence arguments referred to the first amendment or free speech, which was considered to be a well-established ideal. In comparison, a vague argument such as “violence shouldn’t be shown on TV” is not closely linked with an identifiable, well-established ideal, and would receive a lower score.

There was not an explicit consideration of formal or informal logical fallacies (e.g., slippery slope arguments, appeal to authority, argument from silence), because some arguments that technically are logical fallacies can in fact carry varying levels of strength in real-world reasoning, depending on the context (Hahn & Oaksford, 2007). For an extended treatment of real-life argumentation schemes as an alternative to the traditional reasoning fallacies, see Walton et al. (2008).

Argument Quality Rubric

A (90-99 points)

- Highly compelling arguments
- Relatively difficult to refute directly and could not be easily dismissed by a reasonable opponent

- May cite specific evidence
- Clearly based on well-accepted definitions, rights, etc.
- Address outcomes that are highly likely and/or very favorable/aversive.
- Arguments in this class don't have to be perfect. However, they do have to be difficult to dismiss.

Examples:

- Violence in media is a form of free speech and does not present a clear and present danger.
- Viewing violence may make people more likely to commit violent acts in real life.

B (80-89 points)

- Moderately compelling arguments
- Moderately likely outcome and/or moderately favorable/aversive outcome
- Reasonably well-accepted claims, but may be only indirectly tied to principles that would qualify for an "A" rating

Examples:

- Violence shouldn't be restricted because the writers put that much in on purpose in order to engage the audience and put them more in the story.
- If a person is violent, then it doesn't matter if they watch violent TV, they will still be violent.

C (70-79 points)

- Somewhat compelling arguments
- Outcome is not very likely and/or not very favorable/aversive
- Not difficult to address with a reasonable accommodation or minor modification to the prompt
- Vague references to principles that may be invoked by more highly-rated arguments
- May focus on one relevant component of the problem but miss the "big picture"

Examples:

- Violence should not be restricted because that would mean restricting what's shown on the news, and people have a right to know what's going on in the world.
- Violence can teach children from rough neighborhoods how to fight.

D (60-69 points)

- Minimally compelling arguments
- Trivial outcomes and/or very unlikely outcomes

- Very vague
- May invoke simple personal opinion

Address a minor or non-essential component of the problem

Examples:

- That would be censorship, which I don't agree with.
- I don't like violent TV shows.
- Some people get emotional over what they see on TV.

REFERENCES

Ackerman, P. L. (2000). Domain-specific knowledge as the “dark matter” of adult intelligence: Gf/Gc, personality and interest correlates. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *55*(2), P69–P84. http://doi.org/10.1016/S0160-2896(96)90016-1

Ackerman, P. L., & Beier, M. E. (2007). Further explorations of perceptual speed abilities in the context of assessment methods, cognitive abilities, and individual differences during skill acquisition. *Journal of Experimental Psychology. Applied*, *13*(4), 249–272. http://doi.org/10.1037/1076-898X.13.4.249

Ackerman, P. L., & Ellingsen, V. J. (2016). Speed and accuracy indicators of test performance under different instructional conditions: Intelligence correlates. *Intelligence*, *56*, 1–9. http://doi.org/10.1016/j.intell.2016.02.004

Ackerman, P. L., & Kanfer, R. (1993). Integrating laboratory and field study for improving selection: Development of a battery for predicting air traffic controller success. *Journal of Applied Psychology, 78*(3), 413-432.

Altemeyer, B. (2002). Dogmatic behavior among students: testing a new measure of dogmatism. *The Journal of Social Psychology*, *142*(6), 713–721. http://doi.org/10.1080/00224540209603931

Baron, J. (1991). Beliefs about thinking. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 169–186). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking & Reasoning*, *1*(3), 221–235. http://doi.org/10.1080/13546789508256909

Beier, M. E., & Ackerman, P. L. (2005). Age, ability, and the role of prior knowledge on the acquisition of new domain knowledge: promising results in a real-world learning environment. *Psychology and Aging*, *20*(2), 341–55. http://doi.org/10.1037/0882-7974.20.2.341

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive Science: A Multidisciplinary Journal*, *24*(4), 573–604. http://doi.org/10.1016/s0364-0213(00)00033-1

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116–131.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307.

Chang Rundgren, S.-N., & Rundgren, C.-J. (2010). SEE-SEP: From a separate to a holistic view of socioscientific issues. *Asia-Pacific Forum on Science Learning and Teaching*, *11*(1), 1–24.

Christensen, C., & Elstein, A. S. (1991). Informal reasoning in the medical profession. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 17–36). Hillsdale, N.J.: Routledge.

Christenson, N., Chang Rundgren, S. N., & Höglund, H. O. (2012). Using the SEE-SEP model to analyze upper secondary students' use of supporting reasons in arguing socioscientific issues. *Journal of Science Education and Technology*, *21*(3), 342–352. http://doi.org/10.1007/s10956-011-9328-x

Christenson, N., Chang Rundgren, S. N., & Zeidler, D. L. (2014). The relationship of discipline background to upper secondary students' argumentation on socioscientific issues. *Research in Science Education*, *44*(4), 581–601. http://doi.org/10.1007/s11165-013-9394-6

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1957). The two disciplies of scientific psychology. *American Psychologist*, *12*(671–684).

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York, NY: HarperCollins.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change" -- Or should we? *Psychological Bulletin*, *74*(1), 68–80. http://doi.org/10.1037/h0029382

Crowson, H. M. (2009). Does the DOG scale measure dogmatism? Another look at construct validity. *The Journal of Social Psychology*, *149*(3), 265–283.

Desjarlais, M., & Willoughby, T. (2007). Supporting learners with low domain knowledge when using the internet. *Journal of Educational Computing Research*, *37*(1), 1–17. http://doi.org/10.2190/K788-MK86-2342-3600

Eigenberger, M. E., & Sealander, K. A. (2001). A scale for measuring students' anti-intellectualism. *Psychological Reports*, *89*(2), 387–402. http://doi.org/10.2466/pr0.2001.89.2.387

Ekstrom, R. B., French, J. W., Harman, H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests*. *Educational Testing Service*. Princeton, NJ. Retrieved from http://www.ets.org/Media/Research/pdf/Manual_for_Kit_of_Factor-Referenced_Cognitive_Tests.pdf

Evans, J. S. B. T., & Thompson, V. A. (2004). Informal reasoning: theory and method. *Canadian Journal of Experimental Psychology*, *58*(2), 69–74. http://doi.org/10.1037/h0085797

Everett, J. A. C. (2013). The 12 Item Social and Economic Conservatism Scale (SECS). *PLoS ONE*, *8*(12), 1-11. http://doi.org/10.1371/journal.pone.0082131

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–60. http://doi.org/10.3758/BRM.41.4.1149

Furlong, P. R. (1993). Personal factors influencing informal reasoning of economic issues and the effect of specific instructions. *Journal of Educational Psychology*, *85*(1), 171–181. http://doi.org/10.1037/0022-0663.85.1.171

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin*, *105*(3), 331–251.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. http://doi.org/10.1037/0033-295X.103.4.650

Goff, M., & Ackerman, P. L. (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology*, *84*(4), 537–552. http://doi.org/10.1037/0022-0663.84.4.537

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732. http://doi.org/10.1037/0033-295X.114.3.704

Heck, R. H., Thomas, S. L., & Tabata, L. N. (2014). *Multilevel and longitudinal modeling with IBM SPSS* (2nd ed.). New York, NY: Routledge.

Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: Thinking aloud during online searching. *Educational Psychologist*, *39*(1), 43–55. http://doi.org/10.1207/s15326985ep3901_5

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. http://doi.org/10.2307/4615733

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY:

Routledge.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). New York, NY: Guildford Press.

Jonassen, D. H. (1997). Instructional design models for well-structured and III-structured problem-solving learning outcomes. *Educational Technology Research and Development*, *45*(1), 65–94. http://doi.org/10.1007/BF02299613

Kardash, C. M., & Scholes, R. J. (1996). Effects of preexisiting beliefs, epistemological beliefs, and need for cognition on interpretation of controversial issues. *Journal of Educational Psychology*, *88*(2), 260–271. http://doi.org/10.1037/0022-0663.88.2.260

Kerstetter, D., & Cho, M.-H. (2004). Prior knowledge, credibility and information search. *Annals of Tourism Research*, *31*(4), 961–985. http://doi.org/10.1016/j.annals.2004.04.002

King, P. M., & Kitchener, K. S. (2002). The Reflective Judgment Model: Twenty years of research on epistemic cognition. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 37–62). Mahwah, NJ.

Klaczynski, P. A., & Gordon, D. H. (1996). Everyday statistical reasoning during adolescence and young adulthood: motivational, general ability, and developmental influences. *Child Development*, *67*(6), 2873–2891.

Klaczynski, P. A., & Robinson, B. (2000). Personal theories, intellectual ability, and epistemological beliefs: Adult age differences in everyday reasoning biases. *Psychology and Aging*, *15*(3), 400–416. http://doi.org/10.1037/0882-7974.15.3.400

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, *56*, 28–55.

Kortland, K. (1996). An STS case study about students' decision making on the waste issue. *Science Education*, *80*, 673–689.

Kruglanski, A. W. (1990). Motivations for judging and knowing: Implications for causal attribution. In E. T. Higgins & R. M. Sorrentino (Eds.), *The handbook of motivation and covnition: Foundation of social behavior (Vol. 2)* (pp. 333–368). New York, NY: Guilford Press.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing." *Psychological Review*, *103*(2), 263–283. http://doi.org/10.1037/0033-295X.103.2.263

Kuhn, D. (1991). *The skills of argument*. New York, NY: Cambridge University Press.

Kuhn, D., Shaw, V., & Felton, M. (1997). Effects of dyadic interaction on argumentative reasoning. *Cognition and Instruction*, *15*(3), 287–315.

Lawrence, J. A. (1991). Informal reasoning in the judicial system. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 59–82). Hillsdale, N.J.: Routledge.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison, WI: The University of Wisconsin Press.

Macpherson, R., & Stanovich, K. E. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences*, *17*, 115–127. http://doi.org/10.1016/j.lindif.2007.05.003

Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction*, *14*(2), 139–178. http://doi.org/10.1207/s1532690xci1402_1

Nussbaum, E. M., & Kardash, C. M. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology*, *97*(2), 157–169. http://doi.org/10.1037/0022-0663.97.2.157

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we thing*. New York, NY: Penguin Press.

Perkins, D. N. (1985a). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology*, *77*(5), 562–571. http://doi.org/10.1037/0022-0663.77.5.562

Perkins, D. N. (1985b). Reasoning as imagination. *Interchange*, *16*(1), 14–26. http://doi.org/10.1007/BF01187588

Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 83–105). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher*, *18*(1), 16–25.

Ricco, R. B. (2007). Individual differences in the analysis of informal reasoning fallacies. *Contemporary Educational Psychology*, *32*(3), 459–484.

http://doi.org/10.1016/j.cedpsych.2007.01.001

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences*, *50*(1), 90–94. http://doi.org/10.1016/j.paid.2010.09.004

Rolfhus, E. L., & Ackerman, P. L. (1996). Self-report knowledge: At the crossroads of ability, interest, and personality. *Journal of Educational Psychology*, *88*(1), 174–188. http://doi.org/10.1037//0022-0663.88.1.174

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*, *41*(5), 513–536. http://doi.org/10.1002/tea.20009

Sadler, T. D., Chambers, F. W., & Zeidler, D. L. (2004). Student conceptualizations of the nature of science in response to a socioscientific issue. *International Journal of Science Education*, *26*(4), 387–409. http://doi.org/10.1080/0950069032000119456

Sadler, T. D., & Donnelly, L. A. (2006). Socioscientific argumentation: The effects of content knowledge and morality. *International Journal of Science Education*, *28*(12), 1463–1488.

Sadler, T. D., & Zeidler, D. L. (2005). The significance of content knowledge for informal reasoning regarding socioscientific issues: Applying genetics knowledge to genetic engineering issues. *Science Education*, *89*(1), 71–93. http://doi.org/10.1002/sce.20023

Shin, N., Jonassen, D. H., & McGee, S. (2003). Predictors of well-structured and ill-structured problem solving in an astronomy simulation. *Journal of Research in Science Teaching*, *40*(1), 6–33. http://doi.org/10.1002/tea.10058

Slosson, R. L. (1981). *Slosson Intelligence Test*. New York, NY: Slosson Educational Publications.

Snijers, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis* (1st ed.). Thousand Oaks, CA: SAGE Publications.

Stanovich, K. E. (1986). Matthew effects in reading: some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. http://doi.org/10.1598/RRQ.21.4.1

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*(2), 342–357. http://doi.org/10.1037/0022-0663.89.2.342

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought.

Journal of Experimental Psychology: General, *127*(2), 161–188. http://doi.org/10.1037//0096-3445.127.2.161

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*(3), 225–247. http://doi.org/10.1080/13546780600780796

Stanovich, K. E., & West, R. F. (2008). On the failure of cognitive ability to predict myside and one-sided thinking biases. *Thinking & Reasoning*, *14*(2), 129–167. http://doi.org/10.1080/13546780701679764

Sternberg, R. J. (1984). Toward a triarchic theory of human intelligence. *The Behavioral and Brain Sciences*, *7*, 269–315.

Symons, S., & Pressley, M. (1993). Prior knowledge affects text search success and extraction of information. *Reading Research Quarterly, 28*(3), 250–261.

Toplak, M. E., & Stanovich, K. E. (2003). Associations between myside bias on an informal reasoning task and amount of post-secondary education. *Applied Cognitive Psychology*, *17*(7), 851–860. http://doi.org/10.1002/acp.915

Voss, J. F. (1991). Informal reasoning and international relations. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 37–58). Hillsdale, N.J.: Routledge.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. New York, NY: Cambridge University Press.

Wawro, A. (2013). Get your privacy ducks in a row with DuckDuckGo.*PCWorld.* Retrieved from http://www.pcworld.com/article/2035703/get-your-privacy-ducks-in-a-row-with-duckduckgo.html

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence*. Toronto, ON: The Psychological Corporation, Harcourt Brace Jovanovich, Inc.

Weinstock, M. P., Neuman, Y., & Glassner, A. (2006). Identification of informal reasoning fallacies as a function of epistemological level, grade level, and cognitive ability. *Journal of Educational Psychology*, *98*(2), 327–341. http://doi.org/10.1037/0022-0663.89.2.327

Whitmire, E. (2004). The relationship between undergraduates' epistemological beliefs, reflective judgment, and their information-seeking behavior. *Information Processing and Management*, *40*(1), 97–111. http://doi.org/10.1016/S0306-4573(02)00099-7

Willoughby, T., Anderson, S. A., Wood, E., Mueller, J., & Ross, C. (2009). Fast searching for information on the Internet to use in a learning context: The impact of domain knowledge. *Computers & Education*, *52*(3), 640–648. http://doi.org/10.1016/j.compedu.2008.11.009

Wittmann, W. W., & Suss, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik symmetry. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 77–108). Washington, DC: American Psychological Association.

Wolfe, C. R. (2012). Individual differences in the "myside bias" in reasoning and written argumentation. *Written Communication*, *29*(4), 477–501. http://doi.org/10.1177/0741088312457909

Wolfe, C. R., & Britt, M. A. (2008). The locus of the myside bias in written argumentation. *Thinking & Reasoning*, *14*(1), 1–27. http://doi.org/10.1080/13546780701527674

Woo, S. E., Harms, P. D., & Kuncel, N. R. (2007). Integrating personality and intelligence: Typical intellectual engagement and need for cognition. *Personality and Individual Differences*, *43*(6), 1635–1639.

Zeidler, D. L., & Schafer, L. E. (1984). Identifying mediating factors of moral reasoning in science education. *Journal of Research in Science Teaching*, *21*(1), 1–15. http://doi.org/10.1002/tea.3660210102