The Use Of Alternative Reasons In Probabilistic Judgment

Burcu Gurcay-Morris

University of Pennsylvania, bgurcaymorris@gmail.com

Follow this and additional works at: https://repository.upenn.edu/edissertations

Part of the Psychology Commons

Recommended Citation


This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/2321
For more information, please contact repository@pobox.upenn.edu.
The Use Of Alternative Reasons In Probabilistic Judgment

Abstract
This dissertation investigates people's ability to search for and use alternative reasons while making probabilistic judgments, with the goals of devising and testing (1) a new actively open-minded thinking (AOT) measure that assesses thinking behavior by looking at people's ability to generate alternative reasons/contradicting evidence, and (2) a short online module to train people in actively open-minded thinking. In Chapter 1 we assessed individual differences in actively open-minded thinking on probabilistic judgment tasks by using both belief and behavioral measures. In the first three studies subjects made probabilistic judgments in three-choice almanac questions, while in the fourth study subjects made point and confidence interval estimates for numerical almanac questions. Compared to the low scoring subjects, subjects who score high on the new behavioral AOT measure were more likely to have more accurate probability judgments when they did not know the correct answer to the question. Higher scores on the behavioral actively open-minded thinking measure were also associated with lower overconfidence. In Chapter 2 we tested the effectiveness of making subjects consider alternative reasons and a one-hour long online training module in AOT. Studies 4 and 5 tested whether making subjects consider alternative reasons would improve their accuracy and decrease their overconfidence. In Study 4 we observed that this intervention was successful in increasing the number of alternative reasons and subjects' accuracy when subjects did not know the correct answer. There was also a slight decrease in overconfidence as a result of this intervention. Study 5, which used point estimate questions, did not show any benefits of the intervention in accuracy or overconfidence. Studies 6 and 7 tested the effectiveness of new online AOT training modules we designed for adults. The training module in Study 6 did not show any improvement in subjects' accuracy, but the training condition in Study 7 increased subjects' accuracy scores when subjects did not know the correct answer. We observed some effect of the training on subjects' overconfidence such that going through the training decreased subjects' unwarranted confidence. Chapter 3 discusses the relation between the behavioral and belief measures of AOT, their effects on accuracy and overconfidence. We specifically argue that while the belief measure of AOT assesses general AOT tendencies, the behavioral measure of AOT assesses task specific AOT behavior. The results show that considering alternative reasons increases subjects' accuracy by lowering their overconfidence. We finally discuss our results from training adult population in AOT and suggest potential testing scenarios for our training module.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Psychology

First Advisor
Jonathan Baron

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/2321
Keywords
actively open-minded thinking, individual differences, myside bias

Subject Categories
Psychology

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/2321
THE USE OF ALTERNATIVE REasons IN Probabilistic jUDGMENT

Burcu Gürçay-Morris

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

________________________

Jonathan Baron, PhD, Professor of Psychology

Graduate Group Chairperson

________________________

Sara Jaffee, PhD, Professor of Psychology

Dissertation Committee

Barbara Mellers, PhD, I. George Heyman University Professor

Uri Simonsohn, PhD, Professor of Operations, Information and Decisions
THE USE OF ALTERNATIVE REASONS IN PROBABILISTIC JUDGMENT

COPYRIGHT

2016

Burcu Gürçay-Morris

This work is licensed under the
Creative Commons Attribution-
NonCommercial-ShareAlike 3.0
License

To view a copy of this license, visit

https://creativecommons.org/licenses/by-nc-sa/3.0/us/
ACKNOWLEDGMENT

This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.
This dissertation investigates people’s ability to search for and use alternative reasons while making probabilistic judgments, with the goals of devising and testing (1) a new actively open-minded thinking (AOT) measure that assesses thinking behavior by looking at people’s ability to generate alternative reasons/contradicting evidence, and (2) a short online module to train people in actively open-minded thinking. In Chapter 1 we assessed individual differences in actively open-minded thinking on probabilistic judgment tasks by using both belief and behavioral measures. In the first three studies subjects made probabilistic judgments in three-choice almanac questions, while in the fourth study subjects made point and confidence interval estimates for numerical almanac questions. Compared to the low scoring subjects, subjects who score high on the new behavioral AOT measure were more likely to have more accurate probability judgments when they did not know the correct answer to the question. Higher scores on the behavioral actively open-minded thinking measure were also associated with lower overconfidence. In Chapter 2 we tested the effectiveness of making subjects consider alternative reasons and a one-hour long online training module in AOT. Studies 4 and 5 tested whether making subjects consider alternative reasons would improve their accuracy and decrease their overconfidence. In Study 4 we observed that this intervention was successful in
increasing the number of alternative reasons and subjects’ accuracy when subjects did not know the correct answer. There was also a slight decrease in overconfidence as a result of this intervention. Study 5, which used point estimate questions, did not show any benefits of the intervention in accuracy or overconfidence. Studies 6 and 7 tested the effectiveness of new online AOT training modules we designed for adults. The training module in Study 6 did not show any improvement in subjects’ accuracy, but the training condition in Study 7 increased subjects’ accuracy scores when subjects did not know the correct answer. We observed some effect of the training on subjects’ overconfidence such that going through the training decreased subjects’ unwarranted confidence. Chapter 3 discusses the relation between the behavioral and belief measures of AOT, their effects on accuracy and overconfidence. We specifically argue that while the belief measure of AOT assesses general AOT tendencies, the behavioral measure of AOT assesses task specific AOT behavior. The results show that considering alternative reasons increases subjects’ accuracy by lowering their overconfidence. We finally discuss our results from training adult population in AOT and suggest potential testing scenarios for our training module.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ACKNOWLEDGMENT</th>
<th>iii</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF ILLUSTRATIONS</td>
<td>ix</td>
</tr>
<tr>
<td>MEASUREMENT OF INDIVIDUAL DIFFERENCES IN ACTIVELY OPEN-MINDED THINKING</td>
<td>1</td>
</tr>
<tr>
<td>Actively Open-Minded Thinking and Intelligence</td>
<td>2</td>
</tr>
<tr>
<td>Actively Open-Minded Thinking and Other Individual Differences</td>
<td>5</td>
</tr>
<tr>
<td>Measures of Actively Open-Minded Thinking</td>
<td>6</td>
</tr>
<tr>
<td>Self-Report Measures</td>
<td>7</td>
</tr>
<tr>
<td>Behavioral Measures</td>
<td>9</td>
</tr>
<tr>
<td>Overview of the Present Chapter</td>
<td>12</td>
</tr>
<tr>
<td>Studies 1A and 1B</td>
<td>13</td>
</tr>
<tr>
<td>Method</td>
<td>13</td>
</tr>
<tr>
<td>Summary and Discussion of Results</td>
<td>17</td>
</tr>
<tr>
<td>Results</td>
<td>18</td>
</tr>
<tr>
<td>Study 2</td>
<td>27</td>
</tr>
<tr>
<td>Method</td>
<td>27</td>
</tr>
<tr>
<td>Summary and Discussion of Results</td>
<td>28</td>
</tr>
<tr>
<td>Results</td>
<td>30</td>
</tr>
<tr>
<td>Study 3</td>
<td>35</td>
</tr>
<tr>
<td>Method</td>
<td>35</td>
</tr>
<tr>
<td>Summary and Discussion of Results</td>
<td>37</td>
</tr>
<tr>
<td>Results</td>
<td>37</td>
</tr>
<tr>
<td>INTERVENTIONS FOR MYSIDE BIAS AND TEACHING ACTIVELY OPEN-MINDED THINKING</td>
<td>44</td>
</tr>
<tr>
<td>Interventions for Myside Bias</td>
<td>44</td>
</tr>
<tr>
<td>Teaching Actively Open-Minded Thinking</td>
<td>45</td>
</tr>
<tr>
<td>Overview of the Chapter</td>
<td>48</td>
</tr>
<tr>
<td>Study 4</td>
<td>48</td>
</tr>
</tbody>
</table>


# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Actively open-minded thinking scale used in studies 1A and 1B.</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>Actively open-minded thinking scale used in studies 2 through 7.</td>
<td>41-42</td>
</tr>
<tr>
<td>3</td>
<td>Summary table of correlation test results in studies 1A through 7.</td>
<td>108</td>
</tr>
<tr>
<td>4</td>
<td>Summary table of linear mixed-effect model results in studies 1A through 7.</td>
<td>109-111</td>
</tr>
<tr>
<td>5</td>
<td>Summary table of t-test results in studies 4, 6, and 7.</td>
<td>112</td>
</tr>
<tr>
<td>Figure</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>Scatterplots of question difficulty versus correlation coefficients between AOT-Reasons and Brier scores in Studies 1A, 1B, 2, 4, 6, and 7.</td>
<td>43</td>
</tr>
<tr>
<td>2</td>
<td>Subjects’ AOT-Reasons scores in the first and second rounds of Study 7.</td>
<td>96</td>
</tr>
</tbody>
</table>
MEASUREMENT OF INDIVIDUAL DIFFERENCES IN ACTIVELY OPEN-MINDED THINKING

In the case of any person whose judgment is really deserving of confidence, how has it become so? Because he has kept his mind open to criticism of his opinions and conduct. Because it has been his practice to listen to all that could be said against him; to profit by as much of it as was just, and expound to himself, and upon occasion to others, the fallacy of what was fallacious. Because he has felt, that the only way in which a human being can make some approach to knowing the whole of a subject, is by hearing what can be said about it by persons of every variety of opinion, and studying all modes in which it can be looked at by every character of mind. No wise man ever acquired his wisdom in any mode but this; nor is it in the nature of human intellect to become wise in any other manner.


Actively open-minded thinking, a prominent theory of good thinking in judgment and decision making literature, describes a good thinker as someone who is not only fair to new information regardless of her preferred beliefs, but also as someone who seeks out new information to challenge her pet conclusions (Baron, 2008). Baron (1988) proposed this general framework to discuss thinking in terms of the thinker’s existing beliefs and goals, and how these goals and beliefs might affect one’s search of evidence and interpretation of the found evidence. This framework has its roots in the ideal thinking behavior described by John Stuart Mill (1859/1863) in *On Liberty*, but also drew on the work of Irvis L. Janis and colleagues on groupthink behavior (Janis, 1982; Herek, Janis, & Huth, 1987). Nickerson (1988) also talks about a similar approach to good thinking and calls it “fair mindedness.”
Actively Open-Minded Thinking and Intelligence

Actively open-minded thinking (AOT) is a dispositional theory of good thinking. Dispositions are more malleable (see Baron, 1985 and Perkins, Jay, & Tishman, 1993 for an in-depth discussion of dispositional theories of thinking), therefore posing a contrast with cognitive ability, which includes capacities that affect the performance in traditional psychometric intelligence tests.

Given the contrast drawn between dispositions and cognitive abilities, does it mean that AOT is different from intelligence? The answer to this question depends on how one defines intelligence. Many theorists find the common definition of intelligence restrictive (Baron, 1985; Perkins et al., 1993; Stanovich, 2011; Stanovich et al., 2013). Baron (1985) defined intelligence as general abilities that help people succeed in achieving their goals in various situations, which includes not only fixed cognitive abilities, but also thinking dispositions like AOT. Stanovich (2011), on the other hand, called for a separation of the narrower concept of intelligence from thinking dispositions. Both authors agree that AOT is a component of human rationality that is not captured well by traditional psychometric intelligence tests.

The difference of AOT from cognitive ability is suggested by several results. Stanovich and West (1997) showed that AOT predicted variance in an argument evaluation task about real life situations even after cognitive ability measures such as SAT scores and vocabulary test scores were partialled out. Similarly, Kokis et al. (2002) showed that even though cognitive ability was associated with analytic thought on
different tasks, AOT still explained the variance in analytic responding after variance shared with cognitive ability had been controlled. Sá et al. (2005) observed that while AOT predicted variance in subjects’ reiteration or elaboration of the original theory when the variance predicted by cognitive ability was controlled, cognitive ability failed to uniquely predict variance in reiteration of elaboration of the original theory when AOT was partialled out.\(^1\) Argument types offered by subjects with high and low intelligence was highly similar. Klaczynski (1997) looked at adolescents’ everyday reasoning and its relationship with cognitive ability and personal theories, and found similar results others observed with adults. While verbal ability was the best predictor of everyday reasoning, neither verbal ability nor other cognitive abilities predicted biases adolescents showed in everyday reasoning. Adolescents’ personal theories were the strongest predictors of biases in everyday reasoning.

Despite the evidence that AOT can account for different reasoning abilities that traditional intelligence tests cannot account for, there is a large literature that shows that intelligence is related to some belief biases in syllogistic reasoning tasks (e.g., Evans, 2002). Additionally, Stanovich and West (1997) and Sá and Stanovich (2001) both found that the ability to avoid belief bias is predicted by cognitive ability. These results contrast with those observed by Klaczynski and colleagues (e.g., Klaczynski, 1997; Klaczynski & Gordon, 1996) where they found that reasoning was independent of cognitive ability.

\(^1\) However, analyses that partial out cognitive ability do not necessarily remove this variable, because the measures of cognitive ability are not perfect, so we do not remove much of the variance by including cognitive ability measures in the models. Kahneman (1965) brought this problem to the attention of psychologists, but his argument was limited to problems of reliability. Westfall and Yarkoni (2016) also address the issue and propose a solution.
The difference between these two sets of results seem to be due to the lack or presence of decontextualization requirements in the tasks (Sá et al., 2005; Stanovich, West, & Toplak, 2013). While subjects are directly instructed to ignore their favored beliefs in tasks such as syllogisms (Stanovich & West, 2008), in the informal reasoning tasks used by Klacynzki and colleagues, there is no direct instruction to decouple favored beliefs and biases from the reasoning process. When subjects are instructed to keep their personal opinions out of the reasoning process, this signals to them that decoupling is a necessary requirement of the reasoning task. This leads to the correlation observed between bias avoidance and cognitive ability, because different levels of cognitive abilities determine how successful subjects will be in decoupling their personal beliefs. Therefore, decoupling favored beliefs from reasoning results in correlations between cognitive ability and reasoning, while this relation becomes absent when subjects are not specifically instructed to do so.

AOT seems to be a related but different dimension of human rationality. Studies have shown that AOT predicts performance in informal reasoning tasks even after cognitive abilities are partialled out. Baron, Gürçay, and Metz (2016) argues that even if AOT is not properly part of intelligence, it affects measures of intelligence; and it is very likely that AOT itself is affected by cognitive abilities. They further argue that partialling out intelligence or IQ does not make much sense, as any higher correlation observed

---

2 See Stanovich et al. 2013 for a discussion of task differences and the correlation between cognitive ability and reasoning.
between AOT and a reasoning task than one between a reasoning task and IQ, just means that the reasoning task measures AOT better than IQ.³

**Actively Open-Minded Thinking and Other Individual Differences**

Measures of AOT that currently exist in the literature are descendants of the dogmatism scale of Rokeach (1960), the need for cognition scale (NFC) of Cacioppo and Petty (1982), and the Openness-Ideas and Values facets of the Revised NEO Personality Inventory of Costa and McCrae (1992). Baron, Gürçay, and Metz (2016) argues that the current measures of AOT and related personality measures such as Reflectivity/Impulsivity (R/I) are not defined in terms of cognitive processes, but instead are defined in terms of where they fall in a multi-dimensional semantic space. It is possible to observe overlaps between AOT and other scales such as need for cognition and cognitive reflection.

For example, Baron et al. (2015) observed that the belief measure scale for AOT was correlated with performance on an extended version of the Cognitive Reflection Task (CRT) (Frederick, 2005) and R/I measures. They further argued that AOT is consistent with a disposition to question initial answers and to search thoroughly before giving a response, but it is the latter disposition that might be determining the relation between measures of CRT and AOT. Similarly, Campitelli and Labollita (2010) also showed that CRT includes more characteristics than Frederick (2005) proposed, and suggested that CRT is related to AOT. Kardas and Scholes (1996) observed that subjects who scored

---

³ Baron, Gürçay, and Metz (2016) also discusses the problem of statistical control in studying correlations between different variables.
highly on the NFC Scale were more likely to draw conclusions that reflected the inconclusive nature of the mixed evidence subjects read during the study. This study did not look at how AOT performed in the same task, but the fairness of evaluation high NFC subjects showed is also a disposition high AOT subjects have.

Despite the similarities between AOT and other individual differences, the overlap is not perfect, and there is evidence in literature that shows that certain judgment tasks are better at measuring AOT than others. Toplak, West, and Stanovich (2011) showed that CRT measure was the strongest predictor of subjects’ performance on heuristics-and-biases tasks than measures of thinking dispositions, which included AOT. However, Haran, Ritov, and Mellers (2013) observed that AOT was the only reliable predictor of performance in an estimation task, even though AOT was positively correlated with NFC, CRT, and Openness to Experience from the Big Five personality dimensions (Gosling, Rentfrow, & Swann, 2003). AOT seems to be correlated with other individual differences such as NFC, CRT, and R/I, but the predictiveness of these individual differences are dependent on the task characteristics.

**Measures of Actively Open-Minded Thinking**

Some measures of AOT are self-reports, where subjects are asked to report levels of agreement or disagreement with statements about how people should think (e.g., Stanovich & West, 1997). Other measures are behavioral, where researchers ask subjects to express their opinions on certain issues and assess myside bias, favoritism towards preexisting beliefs even in the face of conflicting new evidence (e.g., Perkins et al., 1986;
Baron, 1995); or other relevant biases to AOT such as belief overkill, the irrational tendency to interpret new evidence as supporting a favored opinion (Baron, 2009).

**Self-Report Measures**

Baron (1991) argued that beliefs affect how people behave, and showed that assessment of beliefs can be a good measure of how people think (Baron, 1995). Stanovich and West (1997, 1998) devised an AOT subscale as part of a thinking dispositions questionnaire. Their design was influenced by the critical thinking literature (e.g., Nickerson, 1988; Perkins et al., 1993) but also by the work of Baron (1985, 1988).

Their AOT scale is a composite of several scales. These were (1) Flexible Thinking scale, which taps into reflectivity, willingness to consider contradictory beliefs, willingness to consider alternative opinions and explanations, a tolerance for ambiguity, and willingness to postpone closure; (2) Openness-Ideas and (3) Openness-Values subscales from the Revised NEO Personality Inventory (Costa & McCrae, 1992); (4) Absolutism subscale, which was based on Perry’s (1970) stages of epistemological development in young adulthood; (5) Dogmatism subscale based on Rokeach’s (1960) Dogmatism scale; and (6) Categorical Thinking subscale, which borrowed three items from Epstein and Meier’s (1989) Constructive Thinking Inventory to assess black-and-white kind of thinking in subjects. To construct the AOT scale, the authors summed the scores on the Flexible Thinking, Openness-Ideas, and Openness-Values scales and subtracted the sum of the Absolutism, Dogmatism, and Categorical Thinking scales. Higher scores on this scale indicated flexibility in thinking and belief change, while lower scores indicated rigidity in
thinking and resistance to belief change. We should note that this scale was constructed to assess AOT specifically in college students.

Baron selected items from the AOT Scale of Stanovich and West and added other items to devise a short scale, which was appropriate for the general population to assess beliefs about thinking. This 7-item scale consists of statements that subjects indicate agreement or disagreement with on a 5-point scale. Some items on the scale are: “Allowing oneself to be convinced by an opposing argument is a sign of good character” and “It is important to persevere in your beliefs even when evidence is brought to bear against them.” (The first 7 items on the AOT Scale in Table 1 are the original items from Baron’s AOT Scale.)

Both the long and short versions of the AOT Scale have been successful in predicting performance in other tasks as we mentioned in the previous section. However, these scales only assess beliefs about good thinking but not how people actually think. Additionally, Baron’s AOT Scale is weighted towards myside bias, irrational belief persistence, and inference. Baron et al. (2015) used an AOT Scale with an eighth item that also measured search. In the studies we report, we have added three more items to assess dispositions such as tendency to “keep an open mind” before prematurely settling on a conclusion and conduct proper amount of search that are also part of AOT.
**Behavioral Measures**

Most behavioral measures of AOT are designed to assess myside bias. These myside bias tasks ask subjects to list arguments on a given issue, and then experimenters score subjects’ arguments.

Stanovich and West (1997, 1998) developed a behavioral measure to evaluate myside bias called Argument Evaluation Test (AET) to evaluate such arguments. This test had 23 items. Subjects were tasked with evaluating a fictitious individual’s arguments. Each item began with a statement from the fictitious individual regarding a social issue such as, “The welfare system should be drastically cut back in size.” Subjects indicated how much they agree or disagree with this statement. The fictitious individual then offered a justification for his opinion such as, “The welfare system should be drastically reduced in size because welfare recipients take advantage of the system and buy expensive foods with their food stamps.” A critic then presented an argument against this justification such as, “Ninety-five percent of welfare recipients use their food stamps to obtain the bare essentials for their families.” Subjects were told to make the assumption that this counterargument was factually correct. Finally, the fictitious individual offered a rebuttal of the counterargument such as, “Many people who are on welfare are lazy and don’t want to work for a living.” Subjects were again told to assume that the rebuttal was factually correct, and then asked to evaluate the strength of the rebuttal offered by the fictitious character independent of their own opinions or beliefs. Subjects evaluations were then compared to a summary measure of eight expert judges. The expert judges
consisted of philosophy professors and the authors themselves. The authors estimated myside bias by trying to predict subjects’ ratings from both the expert judges’ ratings and the subjects’ preexisting opinions about the issue. Subjects who showed myside bias were those whose ratings of the rebuttal’s strength deviated from expert judges’ ratings and instead were in the direction of their preexisting opinions.

Similar methods were also used by Baron (1995) and Baron (2009). Baron (1995) assessed myside bias in thinking about abortion. Subjects were asked to prepare for a hypothetical class discussion on the topic “Are abortions carried out in the first day of pregnancy (e.g., by the “morning after” pill) morally wrong?” by generating a list of arguments. In a different study, the same subjects were also asked to evaluate arguments regarding abortion from fictitious students by grading them. The author classified the arguments generated by subjects as good or bad according to the justifications offered, and observed that many arguments had questionable warrants, showing that subjects made very little effort to look for evidence against their arguments. Baron (2009) developed a similar method to AET to examine belief overkill in political judgments. Subjects were asked to evaluate hypothetical candidates with conflicted positions on two different topics. For example, subjects saw the following on a page:

Candidate 1 favors a constitutional amendment defining marriage as between a man and a woman and favors increased income taxes on those with high incomes, to reduce the deficit.
Candidate 2 opposes a constitutional amendment defining marriage as between a man and a woman and opposes increased income taxes on those with high incomes, to reduce the deficit.

What is your position on the two candidates (assuming they both have equally acceptable positions on everything else you care about)?

Strongly favor Candidate 1 — Favor Candidate 1 — Neutral — Favor Candidate 2 — Strongly favor Candidate 2

What is the effect of the candidates’ position on a constitutional amendment defining marriage as between a man and a woman on your evaluation of the two candidates?

Strongly favors Candidate 1 . . .

[The same question was asked about the other issue.]

The issues presented on each page were a combination of moral and non-moral issues. There were 6 moral and 6 non-moral issues, and subjects saw all combinations of these issues. Baron observed that the non-moral issue (income taxes in the example above) was affected by the subject’s stance on the moral issue (marriage equality in the example above). Additional behavioral measures have also been developed to assess myside bias and AOT in children (Baron et al. 1993) and adolescents (Kokis et al., 2002; Metz, 2016).

While these behavioral measures are more direct than self-assessments, they require a lot of items to derive a reliable AOT measure, and are usually scored by experts.
and/or experimenters themselves. Additionally, the issues used in such argument tasks need to appeal to a wider population (and be age appropriate in the case of younger populations) such that people are familiar enough with these issues to have preexisting beliefs or opinions about them.

**Overview of the Present Chapter**

In this chapter we report results from studies where we tested individual differences in AOT on probabilistic judgment tasks. We used a behavioral measure similar to the one used by Koriat, Lichtenstein, and Fischhoff (1980) and Hoch (1985) in debiasing procedures, where subjects were asked to list reasons that was either for or against subjects’ pet beliefs. In the first three studies subjects made categorical predictions in almanac questions, while in the fourth study subjects made point and confidence interval estimates for numerical estimation questions. In all these studies subjects additionally wrote and classified the reasons they wrote as for their preferred answers or other options. This method allowed for an easier and simpler scoring of subjects’ reasoning on these tasks as compared to some of the other behavioral measures of AOT such as AET. We investigated whether this behavioral AOT measure would correlate with Baron’s AOT Belief Scale, subjects’ accuracy, defined as Brier scores, and confidence. We hypothesized that the behavioral AOT measure would correlate positively with AOT Belief Scale, and negatively with Brier scores, and expected subjects who scored high on the new behavioral measure to be less overconfident in their predictions than those who scored lower on this measure.
**Studies 1A and 1B**

The main purpose of this study was to investigate individual differences in AOT, with a focus on people’s ability to be able to come up with arguments against their favored judgments. Additionally, we tried to find questions from domains that would be familiar to our subjects, as two pilot studies revealed that some questions such as matrices and logic questions were too difficult for subjects to answer correctly (see Appendix A). Therefore, we picked two new domains: U.S. metropolitan area population and food content. We thought that these would be two domains where people would have a greater knowledge base compared to other questions asked in the pilot studies. This, in return, would give us a better idea regarding the individual differences in AOT and subjects’ ability to come up with arguments against their preferred judgments. Due to the number of questions we collected, we decided to split the questions into two sets, and test them on two different subject groups. Therefore, the method and results sections of these studies are reported together.

**Method**

**Subjects in Study 1A.** Seventy-four subjects participated in the study. The subjects’ age ranged from 22 to 75 (Median = 50); 63.5% were female. The subjects were from a panel of about 1200 people who volunteered to do studies for pay on the Internet over the last 15 years, through advertising, links from various web sites, and word of mouth. They were mostly Americans, varying considerably in age, income, and education level, but with women over-represented. Subjects who did not take previous studies
seriously had been removed over the years. The panel was divided into three groups in order to use different samples for closely related studies. Subjects were paid $6 for participating in this study (through PayPal). A panel of 200 subjects were notified via email when the study was ready, and the study was removed when there were about 70 responses, aiming for 80 subjects.

**Subjects in Study 1B.** Ninety-three subjects participated in the study. The subjects’ age ranged from 21 to 72 (Median = 49); 66.7% were female. The make-up of the subjects in this study were the same as those in Study 1A. Subjects were paid $6 for their participation (through PayPal). A panel of about 200 subjects who did not receive an invitation to do Study 1A was notified by email when the study was ready, and the study was removed when there were about 90 subjects, aiming for 100.

**Questions in Study 1A.** Subjects answered 20 questions with three choice options A, B, and C in the same order. We used a fixed order in this study to reduce extraneous variance. The questions asked subjects which of the U.S. metropolitan areas was the largest one. The cities picked for the twenty questions came from a list on the Web (“Cities and metropolitan areas,” n. d.). Appendix B-1 lists the questions and answers presented to the subjects.

**Questions in Study 1B.** Subjects answered 20 questions with three choice options A, B, and C in the same order. The first quarter of the questions asked subjects to pick the food with the highest calorie content per serving; the second quarter of the questions asked subjects to pick the food with the most about of fiber content per 100 grams; the
third quarter of the questions asked subjects to pick the food with the highest fat content per 100 grams; and the last quarter of questions asked subjects to pick the food with the highest protein content per serving (see Appendix B-1 for a full list of questions).

**Procedure.** The procedures for both studies were similar with one exception, which we describe below. Subjects did the task on their own computers over the Internet. They first read the instructions regarding the task (see Appendix C-1). The instructions and the full studies can be found on [http://finzi.psych.upenn.edu/~baron/ex/bg/args/args2c.html](http://finzi.psych.upenn.edu/~baron/ex/bg/args/args2c.html) (Study 1A) and [http://finzi.psych.upenn.edu/~baron/ex/bg/args/args2f.html](http://finzi.psych.upenn.edu/~baron/ex/bg/args/args2f.html) (Study 1B).

Subjects were told that they would be asked 20 questions, each with three possible answers A, B, and C. They were informed that we were interested in how they thought about answers. Once they read the instructions, they could enter their age, gender, and e-mail address, and proceed with the study. Subjects answered one question on each page. On each page, they saw a short note defining metropolitan area and what the question was asking (see Appendix C-1). They were also given some example reasons. Below the short information the subjects saw the question with three options, and underneath that they were given six note spaces to list their reasons for and against their preferred answer. They were not required to use all six text input spaces but they had to write at least two reasons. In Study 1A, after listing reasons, they were asked to indicate their preferred answer, and then they were asked to state the probability that each of the three answers could be the correct answer. In Study 1B, subjects picked a preferred answer both before...
and after they listed their reasons. The subjects were also given choices for probabilities to indicate their confidence in any answer option being correct, and were instructed to choose the one that was closest to what they think. In Study 1A, subjects assigned their probability judgments after they indicated their preferred answer, while in Study 1B, subjects assigned the probabilities after they were given a chance to change their preferred answer. The options for probability judgments were 0, 1, 5, 10, 20, 30, 40, 50, 60, 70, 90, 95, 99, and 100 (in percent). The subjects knew that 100% would mean they were completely certain that an answer was correct, but they would have a probability of 33% being correct if they were guessing. The subjects were also informed on the instruction page that their probability judgments for the three answers would need to add up to about 100% (no less than 90% and not greater than 110% to allow for the limited set of options). After completing the page, they could move onto the the next page by clicking a button where they could classify their reasons as being for or against the choice options. Once they completed the classification task, they moved onto the next page, again by clicking a button, where they were presented with a new question and they repeated the same procedure for the remaining questions. The classifications of reasons were later checked by one of the authors (BG) to ensure that subjects’ reasons were classified correctly.

After answering all questions, the subjects were presented with an AOT scale (see Table 1). In this survey, subjects read eight statements regarding beliefs about thinking, and were asked to indicate their agreement with each statement on a 5-point scale (1 =
completely agree, 5 = completely disagree). Once they completed the survey, they clicked a button to submit their responses.

Summary and Discussion of Results

In Studies 1A and 1B we looked at individual differences in actively open-minded thinking as well as what kinds of questions might be suitable for studies to test the validity of a new actively open-minded measure based on subjects’ reasoning. We did not observe any correlations between the two AOT measures or between AOT-Beliefs and Brier scores. We observed a correlation between AOT-Reasons and Brier scores.

One surprising result we observed in Study 1B was that the correlation between AOT-Reasons and Brier scores strengthened as the question difficulty increased. We additionally found partial support for the effect of AOT-Reasons on subjects’ Brier scores, and ran additional linear mixed-effect models to understand our results further. Linear mixed-effect models showed an interaction between AOT-Reasons and subjects’ answering questions correctly, but AOT-Reasons had no effect on subjects’ answering questions correctly. Additionally, when subjects were incorrect, higher scores on AOT-Reasons were associated with lower Brier scores (hence better), but when subjects were correct, higher AOT-Reasons were associated with higher Brier scores. Our analyses failed to show any effect of AOT-Beliefs on subjects’ Brier scores.

We assessed the relations between overconfidence and AOT measures by two different overconfidence measures. The first measure compared subjects’ probability judgments for their preferred answers to the percentage of questions they answered
correctly. The second overconfidence measure used Brier scores but was also loosely analogous to the first overconfidence measure. The correlational tests using these different overconfidence measures yielded similar results. We observed no statistically significant correlations between AOT-Reasons and the overconfidence measures in Study 1A, while we did in Study 1B. AOT-Beliefs scores showed no statistically significant correlations with either overconfidence measure. The second overconfidence measure allowed us to run linear mixed-effect models. We found supporting evidence for our hypothesis that higher scores on AOT-Reasons are associated with lower overconfidence.

Even though these studies showed promise in terms of measuring AOT behaviorally, the two different domains the questions came from showed differences in degrees of difficulty, and some of our results appeared to be inconsistent between these two datasets. Therefore, going forward, we decided to systematically select the questions to ask subjects rather than assessing the difficulty of questions in hindsight. Next we discuss the results summarized here in more detail.

Results

Since the experimental designs for studies 1A and 1B were identical except for the topics of the questions, we present the results of these studies together.

Individual Differences Measures for Actively Open Minded Thinking

We looked at two measures to assess individual differences in AOT. The first measure was constructed by looking at the responses subjects gave on the AOT scale, which we call AOT-Beliefs. Subjects’ responses were on a 1 to 5 scale, where higher
numbers meant more actively open-minded thinking except for the reverse-scored items, 4 through 7. First we subtracted subjects’ responses from 3 so that their responses now varied between -2 and 2. Then we took the arithmetic mean of the 8 items for each subject to come up with an AOT scale measure for each subject. The subjects’ AOT-Beliefs scores ranged from -0.750 to 1.75 in Study 1A ($M_{1A} = 0.672$, $SD_{1A} = 0.630$), and from -1.454 to 1.727 in Study 1B ($M_{1B} = 0.748$, $SD_{1B} = 0.545$).

Next we calculated a new measure by looking at the reasons subjects listed for each question, which we call AOT-Reasons. According to this measure, subjects gained 1 point for giving a reason that is for an option other than their preferred option or a reason that is against their preferred option. They did not get any points for giving a reason that supports their preferred option. The subjects’ mean AOT-Reasons scores ranged from 0.00 to 1.90 in Study 1A ($M_{1A} = 0.276$, $SD_{1A} = 0.415$), while they ranged from 0.00 to 1.45 in Study 1B ($M_{1B} = 0.332$, $SD_{1B} = 0.355$). A Pearson’s product-moment correlation testing whether the two individual differences measures were positively correlated showed a negative and statistically non-significant correlation between these two measures in both studies ($r = -0.18$, $t(72) = -1.52$, $p = 0.13$, two-tailed, in Study 1A; $r = -0.09$, $t(91) = -0.83$, $p = 0.41$, two-tailed, in Study 1B).

We did not observe a correlation between subjects’ mean Brier scores and AOT-Beliefs in either study, $r = -0.02$, $t(67) = -0.2$, $p = 0.84$, two-tailed, in Study 1A; $r = 0.05$, $t(90) = 0.47$, $p = 0.64$, two-tailed, in Study 1B. We next hypothesized that the greater the mean AOT-Reasons scores are, the lower the mean Brier scores would be, and found
some support for it, \( r = -0.18, t(67) = -1.53, p = 0.07 \), one-tailed, in Study 1A; \( r = -0.27, t(90) = -2.66, p = 0.01 \), one-tailed, in Study 1B. A recent study found AOT to be associated with lower overconfidence (Haran, Ritov, & Mellers, 2013). Given that Brier scores uses subjects’ probability judgment accuracy in its calculation, it is possible that the way AOT benefits subjects is by lowering their extreme confidence judgments. Then, this mechanism should be most beneficial in cases where subjects pick wrong answers, namely difficult questions. Therefore, we looked at how the correlation between Brier scores and mean AOT-Reasons behaved as the question difficulty varied. While we did not observe a statistically significant correlation in Study 1A, the correlation was in the expected direction, \( r = -0.28, t(18) = -1.22, p = 0.24 \), two-tailed. In Study 1B, we observed that as the questions got more difficult, the correlation between Brier scores and AOT reasoning got stronger, \( r = -0.56, t(18) = -2.89, p = 0.01 \), two-tailed.

**Linear Mixed-Effect Models.** We used the lmer() function in the lme4 package in R (Bates et al., 2014) to look at the effects of AOT-Reasons and AOT-Beliefs on subjects’ Brier scores. We first regressed the Brier scores of subjects on AOT-Reasons with subjects and items as crossed random effects. We observed no effect of the measure on subjects’ Brier scores in Study 1A, but the sign of the coefficient was in the right direction (-0.009, 95% c.i. -0.043 to 0.025, Wald method). However, in Study 1B, the same analysis showed an effect of AOT-Reasons on subjects’ Brier scores (-0.030, 95% c.i. -0.058 to -0.002, Wald method). Thus, we found partial support for our hypothesis
that people who write more reasons that go against their preferred options have lower Brier scores.

One reason why we did not observe any effect of AOT-Reasons in the first round could be that the usefulness of considering alternative reasons might be dependent on whether the correct answer is known. If a subject knows the correct answer, they might not feel the need to list alternative reasons. To test this hypothesis we first regressed the Brier scores of subjects on AOT-Reasons and Correct (whether subjects guessed the answer correctly where 1 = correct and 0 = incorrect with subjects and items as crossed random effects). AOT-Reasons on its own had no effect on subjects’ Brier scores in either study (-0.012, 95% c.i. -0.038 to 0.014, Wald method; -0.001, 95% c.i. -0.023 to 0.019, Wald method), and as it would be expected Correct did have an effect on the Brier scores in both studies (-0.368, 95% c.i. -0.383 to -0.353, Wald method; -0.439, 95% c.i. -0.454 to -0.424, Wald method). We also observed an interaction effect of AOT-Reasons and Correct on the Brier scores in both studies (0.099, 95% c.i. 0.081 to 0.118, Wald method; 0.201, 95% c.i. 0.148 to 0.255, Wald method), showing that the effect of AOT-Reasons on Brier scores is dependent on whether subjects answered the questions correctly or not.

To get a better understanding of what is going on, we split our data into two based on whether subjects guessed a specific question correctly or incorrectly. For each dataset, we regressed subjects’ Brier scores on AOT-Reasons with subjects and items as crossed random effects. The results showed a good effect when subjects guessed the question incorrectly such that higher AOT-Reasons scores were associated with lower Brier scores
in both studies (-0.086, 95% c.i. -0.134 to -0.038, Wald method; -0.099, 95% c.i. -0.153 to -0.044, Wald method). However, when subjects answered the question correctly, we observed a bad effect of AOT-Reasons on subjects’ Brier scores such that higher AOT-Reasons scores were associated with higher Brier scores. While this effect was not statistically significant in Study 1A (0.029, 95% c.i. -0.026 to 0.085, Wald method), it was statistically significant in Study 1B (0.056, 95% c.i. 0.023 to 0.089, Wald method). Finally, we ran a generalized linear mixed-effect model to determine any effect of AOT-Reasons on Correct but we did not observe any effect of AOT-Reasons on Correct in either study (-0.071, 95% c.i. -0.399 to 0.258, Wald method; 0.020, 95% c.i. -0.280 to 0.321, Wald method). It seems that subjects’ being correct on a given question is not dependent upon subjects’ AOT-Reasons scores. These analyses explain the strengthening of the correlation between Brier scores and AOT-Reasons scores as the difficulty of questions increases. Since subjects are more likely to answer difficult questions wrong, writing more alternative reasons as to why their preferred option might be wrong lowers their Brier scores.

---

4 The lmer function used in R was lmer(BS ~ Ar+(1+Ar|S)+(1+Ar|C),d1[d1$Correct<0,],) where BS is Brier scores, Ar is AOT-Reasons, S is subjects, C is items, and d1[d1$Correct<0,] is the data frame that only contained data for incorrectly answered questions.

5 The lmer function used in R was lmer(BS ~ Ar+(1+Ar|S)+(1+Ar|C),d1[d1$Correct<0,],) where BS is Brier scores, Ar is AOT-Reasons, S is subjects, C is items, and d1[d1$Correct<0,] is the data frame that only contained data for correctly answered questions.

6 The generalized lmer model in R was glmer((Correct+1)/2 ~ Ar+(1+Ar|S)+(1+Ar|C), d1 , family = “binomial”), where Ar is AOT-Reasons, S is subjects, C is items, and d1 is the data frame used.
An additional linear mixed-effect model analysis showed that AOT-Beliefs did not have an effect on the Brier scores (-0.006, 95% c.i. -0.069 to 0.056, Wald method, in Study 1A; 0.017, 95% c.i. -0.053 to 0.087, Wald method, in Study 1B).

**A New Confidence Measure for Three-Choice Questions**

In binary judgment tasks such as yes/no forecasting questions, subject assigns two probabilities, one for success and one for failure. If success is 1 and failure is 0, then subject’s expected score is her judged probability of success, or her confidence. The way we usually assess subject’s confidence is by comparing her confidence or probability of success to her actual score, which is the percentage of questions she answered correctly.

In our case, the probability judgment tasks require subjects to assign three probabilities, one of which is implied by the other two. Thus, subjects can be overconfident in giving a probability that is too high to the wrong answer, or one that is too low for the correct answer. Given this, we wanted to calculate subjects’ confidence such that the measure would consider all three probabilities assigned by the subjects. Going along with the idea that confidence is the expected probability of success, we thought we could calculate confidence by what we call the Expected Brier Score (EBS). We now explain how to calculate this score:

Let \( x, y, \) and \( z \) be the probabilities the subject assigns to each option on a three-choice question. Then,

\[
EBS = x(1-x)^2 + (1-x)y^2 + y(1-y)^2 + (1-y)z^2 + z(1-z)^2 + (1-z)x^2
\]
Thus, $x(1-x)^2$ is the component for $x$ if $x$ is true, which has a probability of $x$; and $(1-x)x^2$ is the component for $x$ if $x$ is false, which has a probability of $(1-x)$, and so on.

To assess subject’s confidence, we compare EBS to her actual Brier score. If the subject is appropriately confident, then the mean of this confidence measure will equal the subject’s mean Brier Score. If the mean EBS score is lower than the mean of the subject’s Brier score, then we say that the subject is overconfident. Thus, we calculate the overconfidence a subject shows by subtracting the mean EBS from the mean Brier Score. Therefore, this new overconfidence measure, which we will refer to as Over2, is loosely analogous to the more widely-used overconfidence measure we described above.\(^7\) The advantage of this measure is that it provides a measure for each response, and uses all three probability judgments the subject makes.

In the correlations we ran, we used both the widely-used overconfidence measure (Over1 from now on) and our new measure, Over2. Both of these measures showed that

---

\(^7\) The new measure of overconfidence is sensible but not equivalent to the more commonly used overconfidence measures. We can make such a comparison in the case where there are only two outcomes (e.g., yes/no) for a given question. Let’s assume that $P$ is the subject’s probability judgment of an answer being correct. By using the formula for EBS given above, the EBS for a judgment of $P$ is $2P^2(1-P)$. That is, a proportion $P$ of the outcomes are true, and if this happens, the Brier Score is $2(1-P)^2$, and a proportion of $1-P$ are false, and if that happens, the Brier Score is $2P^2$. Thus, the EBS is $2P^2(1-P)^2 + 2(1-P)^2$, and this simplifies to $2P^2(1-P)$. We can ignore the 2 coefficient as it is a result of the symmetry, and assume that our formula is $P^2(1-P)$. Next, let’s assume that $Q$ is the actual proportion of cases like this that are correct. Then, the actual Brier Score is $2Q^2(1-P)^2 + (1-Q)^2$. We can again drop the 2 coefficient, and simplify the formula to $Q-2QP + P^2$. Then the overconfidence measure, actual Brier Score minus EBS, is $(1-2P)(Q-P)$. The corresponding ordinary overconfidence measure is $Q-P$. The $(1-2P)$ could be written as $2(0.5-P)$, and it comes from the fact that probabilities below 0.5 should never happen while predicting the probability of something being true, because they should be inverted. But this means that the function is relatively flat near $P = 0.5$, compared to the usual overconfidence measure of $P-Q$. For example, if a person is overconfident half the time by 10% (by the standard measure), and underconfident half the time by 10%, the mean will be greater than 0. However, this is still acceptable for the purpose of asking whether AOT reduces overconfidence, as the same measure is used everywhere.
on average subjects are overconfident in their probability judgments (Study 1A: $M_{Over1} = 0.18; M_{Over2} = 0.43$; Study 1B: $M_{Over1} = 0.30; M_{Over2} = 0.55$).

We ran a correlation to see whether lower overconfidence was associated with higher numbers of reasons that went against subjects’ preferred option. We observed a statistically significant correlation between Over1 and AOT-Reasons only in Study 1B. The correlation coefficient with Over1 as the overconfidence measure was $-0.06$, $t(67) = -0.51$, $p = 0.31$, one-tailed, in Study 1A; while in Study 1B, the correlation coefficient was $-0.24$, $t(90) = -2.35$, $p = 0.01$, one-tailed. Using Over2 as the overconfidence measure, we observed a correlation between Over2 and AOT-Reasons in both studies, but only in Study 1B this relation was statistically significant. The correlation coefficient for this test in Study 1A was $-0.20$, $t(67) = -1.66$, $p = 0.051$, one-tailed; while in Study 1B the correlation coefficient was $-0.30$, $t(90) = -3.02$, $p = 0.002$, one-tailed.

We additionally looked at the correlation between AOT-Beliefs and the overconfidence measures. In Study 1A the correlation coefficient between AOT-Beliefs and Over1 was $0.09$, $t(67) = 0.73$, $p = 0.77$, one-tailed; and in Study 1B the correlation coefficient between these two variables was $0.10$, $t(90) = 0.94$, $p = 0.82$, one-tailed. There was no correlation between Over1 and AOT-Beliefs in either study. When we ran the same correlations with Over2, we again observed no correlation between AOT-Beliefs and overconfidence in either study ($r = 0.02$, $t(67) = 0.15$, $p = 0.88$, two-tailed, for Study 1A; $r = 0.06$, $t(90) = 0.57$, $p = 0.57$, two-tailed, for Study 1B).
**Linear Mixed-Effect Models.** We ran linear mixed-effect models to see whether number of against reasons had an effect on subjects’ confidence ratings. Our new overconfidence measure, Over2, can be computed on a single trial, thus we regressed Over2 on AOT-Reasons with subjects and items as crossed random effects. In both sets of data Over2 was lower when AOT-Reasons was higher (-0.034, 95% c.i. -0.067 to -0.001, Wald method, for Study 1A; -0.045, 95% c.i. -0.074 to -0.017, Wald method, for Study 1B), meaning that if subjects wrote more reasons that go against their preferred option, they were less overconfident. However, we did not observe an effect of AOT-Beliefs on Over2 in Study 1A (0.007, 95% c.i. -0.079 to 0.093, Wald method) or in Study 1B (0.029, 95% c.i. -0.072 to 0.130).

We also regressed subjects’ probability judgments for their preferred answer (Prob from now on) on AOT-Reasons with items and subjects as crossed random effects. We observed a statistically significant effect of AOT-Reasons on Prob in both studies such that higher AOT-Reasons scores were associated with lower probability judgments (-0.044, 95% c.i. -0.069 to -0.019, Wald method; -0.048, 95% c.i. -0.069 to -0.028, Wald method). A linear mixed-effect model regressing Prob on AOT-Reasons and Correct with items and subjects as crossed random effects showed an effect of AOT-Reasons on Prob in both studies (-0.044, 95% c.i. -0.069 to -0.019, Wald method; -0.048, 95% c.i. -0.069 to -0.028, Wald method) and no effect of Correct on Prob in either study (0.002, 95% c.i.

---

8 The `lmer` function used in R was `lmer(Prob ~ Ar + (1+Ar|C) + (1+Ar|S),d1)`, where Ar is AOT-Reasons, C is items, S is subjects, and d1 is the data frame used.
-0.003 to 0.008, Wald method; -0.001, 95% c.i. -0.007 to 0.005, Wald method).\(^9\) While scoring high on AOT-Reasons was associated with lower probability judgments for preferred answers, no associations were observed between answering questions correctly and subjects’ probability judgments for preferred answers.

**Study 2**

In this study we aimed to study individual differences in actively open-minded thinking, and more specifically, investigate people’s ability to list reasons against their preferred answer.

**Method**

**Subjects.** Eighty-six subjects participated in the study. The subjects’ age ranged from 19 to 75 (Median = 46.5); 67.4% were female. The make-up of the subjects in this study were the same as those in previous studies. Subjects were paid $6 for their participation (through PayPal). A panel of about 200 subjects was notified by email when the study was ready, and the study was removed when there were about 85 subjects, aiming for 90.

**Questions.** Subjects answered 20 questions with three answer options A, B, and C in the same randomized order. The questions came from the pool of questions used in Study Pilot Study 2 (see Appendix A), and were picked as described in the results section of that study. A full list of questions and answers along with their difficulty and discrimination scores can be found in Appendix B-1.

---

\(^9\) The lmer function in R for this model was lmer(Prob ~ Ar + Correct + (1+Ar|C) + (1+Ar|S),d1), where C is items, S is subjects, and d1 is the data frame used.
**Procedure.** Subjects did the task on their own computers over the Internet. They first read the instructions regarding the task (see Appendix C-1). The instructions and the full study can be found on http://finzi.psych.upenn.edu/~baron/ex/bg/args/args4.html.

The procedure for this study was similar to that of Study 1B with a small change to the AOT scale that was used. At the end of the task, subjects were given an AOT scale but this time the scale had three new additional items, making it an 11-item scale (α = 0.82, see Table 2). These items were added because the 8-item AOT scale was weighted towards inference but did not have enough items regarding information search and keeping an open-mind. As in the previous studies, the classifications of reasons were checked by one of the authors (BG) after the data collection was complete to ensure that subjects’ reasons were classified correctly.

**Summary and Discussion of Results**

In this study we attempted to measure individual differences based on subjects’ reasoning, and testing the effect of actively open-minded thinking on subjects’ Brier scores and overconfidence. We did not observe any significant relations between different measures of AOT or between AOT measures and overconfidence. There was also no correlation between Brier scores and AOT-Reasons scores, but we did observe a statistically significant correlation between AOT-Beliefs and Brier scores. We were also surprised that the correlation between question difficulty and the coefficients for the correlation between mean Brier scores and AOT-Reasons showed a small-to-medium but statistically non-significant correlation. The relation we observed between question
difficulty and the correlation between mean Brier scores and AOT-Reasons did not exist in this study.

Linear mixed-effect models we ran to understand the relations between AOT-Reasons, Brier scores, and subjects’ correctness on questions showed that an interaction between AOT-Reasons and subjects’ answering questions correctly predicted subjects’ Brier scores. Additionally, we observed a marginally significant good effect of AOT-Reasons on Brier scores when subjects answered questions incorrectly, but a bad effect of AOT-Reasons on Brier scores when subjects were correct. Specifically, when subjects were incorrect, higher scores on AOT-Reasons were associated with lower Brier scores, but when subjects were correct, higher AOT-Reasons were associated with higher Brier scores. Our analyses failed to show any effect of AOT-Beliefs on subjects’ Brier scores.

Looking at subjects’ reasons showed us that many subjects wrote either irrelevant comments such as “Salad is tasty” while answering calorie/nutrition questions or said “I don’t know.” We excluded these comments while calculating the AOT-Reasons\textsuperscript{10}, but maybe some of these comments should have been included. Comments like “I don’t know” might indicate useful information such as confidence in an answer, and even irrelevant comments might tell us something in terms of subjects’ thinking process such as searching for evidence. AOT-Reasons as it stands now has this potential weakness. However, Figure 1 shows that the relation we observed in the previous studies still exists. When the questions are more difficult, the correlation between the Brier scores and AOT-Reasons was statistically significant.\textsuperscript{10}

\textsuperscript{10} When we ran the same statistical tests with the original coding of the subjects, some results were statistically significant.
Reasons scores are negative for these questions, which is what we observed before.

Given this and the additional analyses we ran, it seems that this result seems believable: people who think in a more open-minded manner have lower Brier scores on really difficult items.

We additionally ran linear mixed-effect models to see whether AOT-Reasons predicted subjects’ overconfidence or the probabilities they assigned to their preferred answers. We observed no effect of AOT-Reasons on subjects’ overconfidence, but we observed a statistically significant effect of AOT-Reasons on subjects’ probability judgment for their preferred answers. Specifically, higher number of reasons against preferred answers predicted lower probability judgments for preferred answers. Writing reasons against subjects’ preferred reasons seem to decrease subjects’ confidence in their answers. Similar analyses with AOT-Beliefs as the independent variable did not show a statistically significant effect of this variable on subjects’ overconfidence. Next, we describe the statistical analyses and results in more detail.

Results

Since this study was identical in design to Studies 1A and 1B, we ran the same tests to analyze the data.

Individual Differences Measures for Actively Open Minded Thinking

We used two AOT measures, AOT-Beliefs and AOT-Reasons, to look at individual differences in AOT as we did in Studies 1A and 1B. We calculated AOT-Beliefs as we did in the previous studies, but this time we had three additional items.
Subjects’ responses were on a 1 to 5 scale, where higher numbers meant more actively open-minded thinking except for reverse-scored items. First we subtracted subjects’ responses from 3 so that their responses now varied between -2 and 2. Items 4 through 7 and 9 on the scale were reverse-scored, and then we took the arithmetic mean of the 11 items for each subject to come up with an AOT scale measure for each subject. The subjects’ AOT-Beliefs scores ranged from -0.909 to 1.727 ($M = 0.697$, $SD = 0.597$). We calculated AOT-Reasons the same way we did in previous studies. The subjects’ mean AOT-Reasons scores ranged from 0.00 to 1.30 ($M = 0.247$, $SD = 0.278$). The correlation between AOT-Beliefs and mean AOT-Reasons was 0.08, and was not statistically significant $t(84) = 0.76$, $p = 0.45$, two-tailed.

There was a small correlation between mean Brier scores and AOT-Beliefs, showing that people who had lower Brier scores measured higher on AOT-Beliefs, $r = -0.19$, $t(84) = -1.75$, $p = 0.04$, one-tailed. We expected that higher mean AOT-Reasons scores would be correlated with lower mean Brier scores, but this hypothesis was not supported, $r = -0.04$, $t(84) = -0.4$, $p = 0.35$, one-tailed. Additionally, we looked at how the correlation between Brier scores and mean AOT-Reasons behaved as the question difficulty varied. We observed a small-to-medium but not statistically significant correlation in the expected direction, $r = -0.24$, $t(18) = -1.04$, $p = 0.31$, two-tailed. Our hypothesis that as the questions get more difficult, the correlation between Brier scores and AOT reasoning get stronger, was not supported.
**Linear Mixed-Effect Models.** We ran linear mixed-effect models to assess the effect of actively open-minded measures on subjects’ Brier scores. We first regressed Brier scores on AOT-Beliefs scores and did not observe any effect of AOT-Beliefs on Brier scores even though the sign of the estimate coefficient was in the correct direction (-0.065, 95% c.i. -0.139 to 0.008, Wald method). Next we regressed the Brier scores of subjects on AOT-Reasons with subjects and items as crossed random effects. We observed an effect of the measure on subjects’ Brier scores (0.044, 95% c.i. 0.012 to 0.077, Wald method); however, this effect was in the opposite direction of the hypothesis. According to this analysis, higher numbers of reasons that go against subjects’ preferred answers were associated with higher Brier scores. This reverse association probably resulted from reverse causality, when one thinks of more reasons against their hypothesis because it is more likely to be wrong. Good thinking here was probably not much help, because subjects were still highly limited by their knowledge base.

We additionally regressed the Brier scores of subjects on AOT-Reasons and Correct (whether subjects guessed the answer correctly where 1 = correct and 0 = incorrect with subjects and items as crossed random effects). AOT-Reasons had an effect on subjects’ Brier scores but it was in the opposite direction from what we would have expected (0.025, 95% c.i. 0.003 to 0.047, Wald method), and as it would be expected Correct did have an effect on the Brier scores (-0.449, 95% c.i. -0.463 to -0.435, Wald method). We also observed an interaction effect of AOT-Reasons and Correct on the Brier scores (0.058, 95% c.i. 0.038 to 0.079, Wald method).
To get a better understanding of what is going on, we split our data into two based
on whether subjects guessed a specific question correctly or incorrectly. For each dataset,
we regressed subjects’ Brier scores on AOT-Reasons with subjects and items as crossed
random effects. The results showed a good effect when subjects guessed the question
incorrectly such that higher AOT-Reasons scores were associated with lower Brier scores,
but it was not statistically significant (-0.052, 95% c.i. -0.119 to 0.015, Wald method).\textsuperscript{11}
However, when subjects answered the question correctly, we observed a bad effect of
AOT-Reasons on subjects’ Brier scores such that higher AOT-Reasons scores were
associated with higher Brier scores (0.052, 95% c.i. 0.013 to 0.091, Wald method).\textsuperscript{12}
Finally, we ran a generalized linear mixed-effect model to determine any effect of AOT-
Reasons on Correct but we did not observe any effect of AOT-Reasons on Correct
(-0.220, 95% c.i. -0.632 to 0.193, Wald method). It seems that subjects’ being correct on a
given question is not dependent upon subjects’ AOT-Reasons scores. These analyses help
us understand why we observe the strengthening of the correlation between Brier scores
and AOT-Reasons scores as the difficulty of questions increases. Since subjects are more
likely to answer difficult questions wrong, writing more alternative reasons as to why
their preferred option might be wrong lowers their Brier scores.

\textbf{Overconfidence and Actively Open-Minded Thinking}

\textsuperscript{11} The lmer function used in R was \texttt{lmer(BS ~ Ar+(1+Ar|S)+(1+Ar|C),d1[d1$Correct<0,])}, where
BS is Brier scores, Ar is AOT-Reasons, S is subjects, C is items, and d1[d1$Correct<0,] is the
data frame that only contained data for incorrectly answered questions.

\textsuperscript{12} The lmer function used in R was \texttt{lmer(BS ~ Ar+(1+Ar|S)+(1+Ar|C),d1[d1$Correct>0,])}, where
BS is Brier scores, Ar is AOT-Reasons, S is subjects, C is items, and d1[d1$Correct>0,] is the
data frame that only contained data for correctly answered questions.
We assessed the relation between overconfidence and AOT measures by using two overconfidence measures as we did in Studies 1A and 1B. For both measures, the mean overconfidence scores were positive, meaning that subjects were overconfident in their probability judgments ($M_{\text{Over1}} = 0.15; M_{\text{Over2}} = 0.39$).

We hypothesized that higher scores on AOT-Reasons would be associated with lower overconfidence. The correlation between AOT-Reasons and the widely-used overconfidence measure, Over1, was not statistically significant, $r = -0.06$, $t(84) = -0.58$, $p = 0.28$, one-tailed. The correlation between AOT-Reasons and Over2 was not statistically significant either, $r = -0.10$, $t(84) = -0.88$, $p = 0.19$, one-tailed. We additionally looked at the correlations between AOT-Beliefs and overconfidence measures, but observed no correlations between AOT-Beliefs and either overconfidence measure, $r = -0.08$, $t(84) = -0.77$, $p = 0.22$, one-tailed (with Over1 as the overconfidence measure); $r = -0.08$, $t(84) = -0.73$, $p = 0.23$, one-tailed (with Over2 as the overconfidence measure).

**Linear Mixed-Effect Models.** We regressed Over2 on AOT-Reasons with subjects and items as crossed random effects. Similar to our results in Pearson’s product-moment correlations, we did not observe a significant effect of AOT-Reasons on Over2 ($0.021$, 95% c.i. -0.009 to 0.052, Wald method). Similarly, we regressed Over2 on AOT-Beliefs, and did not observe a significant effect of AOT-Beliefs on Over2 even though the effect was in the expected direction (-0.032, 95% c.i. -0.117 to 0.053). In conclusion, there was no effect of the AOT measures on subjects’ overconfidence.
We also regressed subjects’ probability judgments for their preferred answer (Prob from now on) on AOT-Reasons with items and subjects as crossed random effects. We observed a statistically significant effect of AOT-Reasons on Prob such that higher AOT-Reasons scores were associated with lower probability judgments (-0.050, 95% c.i. -0.072 to -0.027, Wald method). A linear mixed-effect model regressing Prob on AOT-Reasons and Correct with items and subjects as crossed random effects showed an effect of AOT-Reasons (-0.048, 95% c.i. -0.070 to -0.026, Wald method) and an effect of Correct on Prob (0.027, 95% c.i. 0.020 to 0.033, Wald method). Scoring high on AOT-Reasons was associated with lower probability judgments for preferred answers, and answering questions correctly was associated with higher probability judgments for preferred answers.

Study 3

The aim of this study was to study individual differences in actively open-minded thinking by using numerical questions.

Method

Subjects. Eighty-two subjects participated in the study. The subjects’ age ranged from 19 to 75 (Median = 50); 67% were female. The make-up of the subjects in this study were the same as those in previous studies. Subjects were paid $8 for their participation.

---

13 The lmer function used in R was lmer(Prob ~ Ar + (1+Ar|C) + (1+Ar|S),d1), where Ar is AOT-Reasons, C is items, S is subjects, and d1 is the data frame used.

14 The lmer function in R for this model was lmer(Prob ~ Ar + Correct + (1+Ar|C) + (1+Ar|S),d1), where C is items, S is subjects, and d1 is the data frame used.
(through PayPal). A panel of about 200 subjects was notified by email when the study was ready, and the study was removed when we had about 100 subjects.

Questions. Subjects answered 25 questions in randomized order. The questions came from the pool of questions used in Pilot Study 3, and were picked as described in the results section of that study. A full list of questions and answers along with their difficulty and discrimination scores can be found in Appendix B-1.

Procedure. Subjects did the task on their own computers over the Internet. They first read the instructions regarding the task (see Appendix C-1). The instructions and the full study can be found on http://finzi.psych.upenn.edu/~baron/ex/bg/args/args7.html.

Each subject answered 25 questions in an order randomized for each subject. Subjects saw one question on each page and were asked to submit their best estimate, a lower and an upper number to give them a range so that they would be correct 90% of the time. After subjects put in the three values, they were asked to list at least two reasons that affected their answer and classify each one. They could say their argument was about “why my answer makes sense”, “why it could be too high”, “why it could be too low”, “why it could be too low or too high”, or “none of the above.” Upon completion of this stage, they were given the opportunity to adjust their estimates if they chose to do so. The subjects repeated this procedure for all the questions, and then took the AOT scale (see Table 2). As in the previous studies, the classifications of reasons were checked by one of the authors (BG) after the data collection was complete to ensure that subjects’ reasons were classified correctly.
Summary and Discussion of Results

This study tried to replicate the results of the previous studies with numerical questions. We observed some discrepancy between our results in this study and previous ones. Earlier studies showed no correlation between the two AOT measures, but in this study we observed a significant correlation in the hypothesized direction. Additionally, higher AOT-Beliefs and AOT-Reasons scores were associated with lower error scores if the error scores were based on confidence interval judgments, but this correlation disappeared if the error scores were calculated by using subjects’ best estimates. Finally, linear mixed-effect analyses did not show any effect of the AOT measures on either accuracy score.

Results

In Pilot Study 3, we tested two potential accuracy scores called Overconf and CIscore (see Appendix A, pages 117-120 for an in-depth discussion of these scores). We decided that Overconf was a more suitable measure for calculating accuracy than CIscore since we are primarily interested in overconfidence, so for the analyses where we wanted to look at the effects of AOT on accuracy scores from subjects’ confidence interval judgments, we used Overconf rather than CIscore. We additionally used another accuracy score called Besterror, which considered subjects’ best point estimate for a given question while calculating error scores.\textsuperscript{15}

\textsuperscript{15} See Pilot Study 3 for an in-depth discussion of Besterror accuracy score.
Individual Differences Measures for Actively Open Minded Thinking

The subjects’ AOT-Beliefs scores ranged from -0.909 to 1.727 (M = 0.542, SD = 0.540), while their mean AOT-Reasons scores ranged from 0.00 to 0.98 (M = 0.043, SD = 0.267). We expected that the mean of AOT-Beliefs and AOT-Reasons would be positively correlated, and our hypothesis was supported, r = 0.27, t(80) = 2.53, p = 0.01, two-tailed.

We observed no correlation between AOT-Beliefs scores and mean Besterror scores, r = -0.14, t(80) = -1.28, p = 0.10, one-tailed. Next we looked at the relation between AOT-Beliefs and Overconf, expecting that higher AOT-Beliefs scores would be associated with lower Overconf scores. This hypothesis was not supported. We observed no correlation between these two variables, r = -0.01, t(80) = -0.05, p = 0.48, one-tailed.

We next wanted to see whether higher mean AOT-Reasons scores would be associated with lower mean Besterror scores, but found no support for this association, r = -0.11, t(80) = -1.02, p = 0.15, one-tailed. However, we observed a statistically significant negative correlation between mean Overconf scores and mean AOT-Reasons scores such that higher number of against reasons were associated with better accuracy, r = -0.25, t(80) = -2.29, p = 0.01, one-tailed. In previous experiments we looked at how the correlation between Brier scores and mean AOT-Reasons behaved as the question difficulty varied. We could potentially run a similar analysis using Besterror or Overconf to replace Brier scores, but we think that it is not very appropriate to do that in this task. Question difficulty is reflected in the subjects’ intervals in this study, but in the previous studies, which used categorical answers rather than numerical estimates, many subject
errors were the result of guesses based on irrelevant information due to lack of
knowledge in question domains. Therefore, these errors sometimes led to below-chance
judgments. Given this, we were not able to run a similar test.

**Linear Mixed-Effect Models.** We wanted to look at the effect of AOT-Reasons
on the subjects’ accuracy scores based on their best estimate (Besterror) and confidence
interval judgments (Overconf). We first regressed Besterror on AOT-Reasons with
subjects and items as crossed random effects. We observed no effect of the measure on
subjects’ Besterror scores (-0.002, 95% c.i. -0.038 to 0.043, Wald method). Next, we
regressed Overconf on AOT-Reasons with subjects and items as crossed random effects
again. Once again, we did not observe any effect of AOT-Reasons on the accuracy score
(-0.008, 95% c.i. -0.027 to 0.011, Wald method).

Additionally, we looked at the effect of AOT-Beliefs on the subjects’ accuracy
scores based on their Besterror and Overconf scores. We crossed each accuracy score on
subjects’ AOT-Beliefs scores with subjects and items as crossed random effects, and did
not find any effect of the mean of AOT-Beliefs on either accuracy score (For Besterror as
the dependent variable the estimate was -0.066, 95% c.i. -0.159 to 0.028, Wald method;
for Overconf as the dependent variable the estimate was -0.002, 95% c.i. -0.071 to 0.067,
Wald method).
Table 1. The following table shows the Actively Open-Minded Scale presented to subjects for Studies 1A and 1B. Questions 4, 5, 6, and 7 are reverse coded.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td><strong>Allowing oneself to be convinced by an opposing argument is a sign of good character.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completely agree</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td><strong>People should take into consideration evidence that goes against their beliefs.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completely agree</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td><strong>People should revise their beliefs in response to new information or evidence.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completely agree</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td><strong>Changing your mind is a sign of weakness.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completely agree</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td><strong>Intuition is the best guide in making decisions.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completely agree</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>6</strong></td>
<td><strong>It is important to persevere in your beliefs even when evidence brought to bear against them.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completely agree</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td><strong>One should disregard evidence that conflicts with one's established beliefs.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completely agree</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>8</strong></td>
<td><strong>People should search actively for reasons why their beliefs might be wrong.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completely agree</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 2. The following table shows the Actively Open-Minded Scale presented to subjects starting in Study 2. Questions 4, 5, 6, 7, and 9 are reverse coded.

|   | 
|---|---
| 1 | **Allowing oneself to be convinced by an opposing argument is a sign of good character.** |
|   | Completely agree | 1 | 2 | 3 | 4 | 5 | Completely disagree |
| 2 | **People should take into consideration evidence that goes against their beliefs.** |
|   | Completely agree | 1 | 2 | 3 | 4 | 5 | Completely disagree |
| 3 | **People should revise their beliefs in response to new information or evidence.** |
|   | Completely agree | 1 | 2 | 3 | 4 | 5 | Completely disagree |
| 4 | **Changing your mind is a sign of weakness.** |
|   | Completely agree | 1 | 2 | 3 | 4 | 5 | Completely disagree |
| 5 | **Intuition is the best guide in making decisions.** |
|   | Completely agree | 1 | 2 | 3 | 4 | 5 | Completely disagree |
| 6 | **It is important to persevere in your beliefs even when evidence brought to bear against them.** |
|   | Completely agree | 1 | 2 | 3 | 4 | 5 | Completely disagree |
| 7 | **One should disregard evidence that conflicts with one's established beliefs.** |
|   | Completely agree | 1 | 2 | 3 | 4 | 5 | Completely disagree |
| 8 | **People should search actively for reasons why their beliefs might be wrong.** |
|   | Completely agree | 1 | 2 | 3 | 4 | 5 | Completely disagree |
9 When we are faced with a new question, the first answer that occurs to us is usually best.

<table>
<thead>
<tr>
<th>Completely agree</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Completely disagree</th>
</tr>
</thead>
</table>

10 When faced with a new question, we should consider more than one possible answer before reaching a conclusion.

<table>
<thead>
<tr>
<th>Completely agree</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Completely disagree</th>
</tr>
</thead>
</table>

11 When faced with a new question, we should look for reasons why our first answer might be wrong, before deciding on an answer.

<table>
<thead>
<tr>
<th>Completely agree</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Completely disagree</th>
</tr>
</thead>
</table>
Figure 1. The figure shows the scatterplots of question difficulty versus correlation coefficients between AOT-Reasons and Brier scores in Studies 1A, 1B, 2, 4, 6, and both rounds of Study 7. The expected sign for the correlation between AOT-Reasons and Brier scores is negative. The hypothesis is that this correlation coefficient becomes more negative (i.e., gets stronger) as the question difficulty increases. The red line in each scatterplot is the regression line.
INTERVENTIONS FOR MYSIDE BIAS AND TEACHING ACTIVELY OPEN-MINDED THINKING

Baron (1988) argued that when people deviate from the normative model of good thinking, they do so by showing favoritism towards evidence that would support their preferred beliefs and conclusions and ignorance or under-weighing of evidence that would contradict these beliefs and conclusions. Perkins, Bushey, and Faraday (1986) calls this set of failures in thinking “myside bias.”

Baron (2008) argued that people’s tendency to seek evidence and make inferences favoring their beliefs is the most serious problem in thinking. He further argued that myside bias may lead to irrational belief persistence. Favoring one-side in search for evidence might make false beliefs to last longer, and could even strengthen the existing false beliefs when they should get weaker.

Interventions for Myside Bias

As we discussed in the previous section, people’s tendency to conduct insufficient search for evidence and possibilities and ignore or dismiss new evidence affect people’s reasoning negatively. In addition to unsubstantiated strengthening of arguments on one side and interpreting evidence in a self-serving way, research has also shown that people who do not follow the principles of AOT are also overconfident and less accurate than those who are good thinkers (Koriat et al., 1980; Hoch, 1985).

If overconfidence and inaccuracy are problems resulting from ignorance of alternative evidence or possibilities, then an intervention that would make subjects
consider alternatives should decrease, if not eliminate, unwarranted overconfidence, and improve accuracy. What makes people consider alternatives? Studies found that making the alternatives salient was successful in getting subjects to consider alternatives, decreasing unwarranted overconfidence (e.g., Weinstein, 1980), and increasing subjects’ accuracy (e.g., Hoch, 1985). Researchers employed different methods such as presenting likelihoods of data under both the focal and alternative hypotheses (Weinstein, 1980; Trope & Bassok, 1982 and 1983; Bassok & Trope, 1984; Trope & Mackie, 1987), asking questions that mention both hypotheses as opposed to just one (Baron, Beattie, & Hershey, 1988), asking participants to generate their own hypotheses (Koehler, 1994), making subjects generate reasons why an alternative might be true (Koriat et al., 1980; Slovic & Fischhoff, 1977), and telling subjects explicitly to consider the alternative (Lord, Lepper, & Preston, 1984) to increase the saliency of alternatives.

**Teaching Actively Open-Minded Thinking**

Researchers have argued for teaching good thinking through instruction (Nickerson, 1988; Baron, 1993) and it has been shown that AOT can be taught this way. Selz (1935) (as cited in Baron et al., 2016) is one of the earliest training research studies conducted, where students, aged 11 to 13, were given an intelligence test with various types of problems. The training group was given training on only one type of problem, where students were instructed to consider the requirements of the task, and test whether each solution met each task requirement. These students were also taught to explain why certain answers did not meet the requirements and to justify their answers when they
seemed to meet the requirements. Compared to the no-training group, the training group showed more improvement in their scores on the intelligence test in all problem types. Additionally, the training group was more likely to change their answers to correct it in the post-training administration of the intelligence test than they were in the pre-training administration of the test. As a result of the training, students were more likely to look for evidence against their initial answers and change them.

Perkins et al. (1986) also designed experimental reasoning courses for high school students based on previous research in informal reasoning. These courses were taught over four consecutive weeks for a total of 16 lessons. The courses emphasized generating reasons by using existing knowledge, and avoiding myside bias. Additionally, students were trained to be true, relevant, and complete while reasoning. The classes mostly involved students’ critiquing their own and peers’ reasoning performances with these reasoning standards in mind. Students’ reasoning about some controversial issues at the end of the course was compared to their reasoning on these issues before the course. The authors observed that the number of arguments students listed on the other side from their own and the quality of these arguments increased as a result of the course. Moreover, when the authors examined the effects of law school classes, high school debate classes, a first-year college class on critical thinking, and a graduate course on thinking, they observed no effect of these courses on students’ tendency to write other side reasons or the quality of these reasons, but they did observe an increase in the number of myside reasons and their quality as a result of taking these classes. The authors concluded that
even though actively open-minded thinking can be taught, the instructions should be
more specific to teach this way of thinking, and not just require students to think.

Similarly, Baron, Badgio, and Gaskins (1985) trained students with reading
disability for 8-months. These students exhibited certain cognitive styles that prevented
them from being academically successful. More specifically, these students were
impulsive, rigid, and non-persistent in their thinking such that they did not spend enough
time to think about problems, failed to consider alternatives to initial possibilities, and
were unable to complete longer tasks. Thus, the training focused on changing these
cognitive styles. Students were taught to take time to think, consider alternatives, and
keep trying by examples, practice exercises, and feedback. Compared to the control
group, students in the training condition took more time to think in various laboratory
tasks, and those students who were particularly impulsive improved their overall
accuracy in these tasks. Finally, Graumlich and Baron (1991) taught two decision making
courses to sixth graders and high school students, which included discussion of AOT. The
authors observed informally that students who used to be very closed-minded before the
course, became more open-minded and considered multiple options and evaluated these
options appropriately.

The studies listed above tried to train mostly students in high school and younger.
To the best of our knowledge, there are no studies that were conducted with adults. In this
chapter we discuss results from two training studies, where we trained adult subjects in
AOT on the World Wide Web and tested the effects of this training on a probabilistic judgment task.

Overview of the Chapter

In this chapter we report results from studies where we either instructed subjects to consider alternative reasons or trained subjects in AOT. In studies 4 and 5 we tested whether telling subjects to consider alternative reasons would improve subjects’ Brier scores and reduce overconfidence. These were replication studies of Koriat et al. (1980) and Hoch (1985) with different questions. We hypothesized that subjects’ Brier scores in the intervention condition should be lower when compared to the Brier scores in the control condition. Additionally, we expected the subjects in the intervention condition to have lower overconfidence than those in the control condition. In studies 6 and 7, we designed a one-hour online AOT training for adults. We hypothesized that subjects should have lower Brier scores and lower overconfidence in the training condition than in the no-training condition. We additionally expected that compared to the no-training condition, subjects would write more reasons that went against their preferred answers in the training condition.

Study 4

The main purpose of this study was to investigate whether forcing subjects to give a reason that goes against their preferred choice or a reason that supports a non-preferred answer would improve their Brier scores.
Method

Subjects. Sixty-three subjects participated in the study. The subjects’ age ranged from 21 to 71 (Median = 46); 58.7% were female. The make-up of the subjects in this study were the same as those in previous studies. Subjects were paid $8 for their participation (through PayPal). A panel of about 200 subjects was notified by email when the study was ready, and the study was removed after 72 hours.

Questions. Subjects answered 20 questions with three answer options A, B, and C in randomized order. The questions were the same as those used in Study 2 with one exception. We replaced an arithmetic question with a population question as it would be very difficult for subjects to come up with an against reason if they knew the correct answer. A full list of questions and answers along with their difficulty and discrimination scores can be found in Appendix B-1.

Procedure. Subjects did the task on their own computers over the Internet. They first read the instructions regarding the task (see Appendix C-1). The instructions and the full study can be found on http://finzi.psych.upenn.edu/~baron/ex/bg/args/args5.html.

The procedure for this study was similar to that of Study 2 with the following modifications. There were two conditions: (1) a control condition where subjects would answer questions like the subjects in Study 2, and list at least two reasons but they were not required to give reasons that were against their preferred answer or for a non-preferred answer; (2) an experimental condition where subjects had to again list at least two reasons but this time, one of the reasons had to be either against their preferred
answer or for a non-preferred answer (Twoside condition from now on). All subjects did the task in both conditions, and answered half the questions in the control condition, and the other half in the experimental condition. Half of the subjects did the task where the sequence of control and experimental conditions were completely randomized. The other half of the subjects did the task where the first five items were in the control condition, while for items 6 through 20, they did ten of the questions in the experimental condition, and five of them in the control condition. Subjects were assigned to these two different sequences randomly. The reason for the different sequences was to test for possible carry-over effects due to the within-subject design of the study. Subjects in both conditions then made probability judgments about the correctness of each option after listing their reasons.

Another way this study differed from the previous studies was that for reason classifications; we added a seventh option labeled “None in particular” for reasons subjects might come up with, that are more like thoughts that do not correspond to being for or against any of the answer options.

After the data collection was completed, the classifications of reasons were checked by one of the authors (BG) to ensure that subjects’ reasons were classified correctly.

Summary and Discussion of Results

In this study we tested whether simply requiring subjects to list at least one reason against their preferred answers would change subjects’ Brier scores and confidence.
We observed small but statistically non-significant correlations between subjects’ Brier scores and AOT measures. Additionally, we observed that the relation between AOT-Reasons and Brier scores strengthens as the question difficulty increases. As Figure 1 shows, the correlation between AOT-Reasons and Brier scores is negative for the most difficult questions. This effect we observed is quite large, and it seems that listing more reasons that go against one’s preferred answers lowers their Brier scores by lowering confidence.

Linear mixed-effect models we ran to understand the relations between AOT-Reasons, Brier scores, and subjects’ correctness on questions showed an interaction between AOT-Reasons and subjects’ answering questions correctly, as well as an effect of AOT-Reasons on subjects’ answering questions correctly such that higher AOT-Reasons scores were associated with subjects’ answering questions correctly. Additionally, we observed a good effect of AOT-Reasons on Brier scores when subjects answered questions correctly, but a bad effect of AOT-Reasons on Brier scores when subjects were incorrect. Specifically, when subjects were incorrect, higher scores on AOT-Reasons were associated with lower Brier scores, but when subjects were correct, higher AOT-Reasons were associated with higher Brier scores. We did not observe any effect of AOT-Beliefs on subjects’ Brier scores.

Using subjects as the unit of analysis we did not observe any statistically significant correlations between either measures of AOT and overconfidence. However, AOT-Reasons had an effect on subjects’ overconfidence and their probability judgments
for their favored answers in the linear mixed-effect models we ran such that subjects who had higher AOT-Reasons scores tended to have lower overconfidence.

Comparisons between the control and experimental conditions showed that subjects scored higher on AOT-Reasons, and wrote more reasons in the experimental condition, but even though the Brier scores and overconfidence were lower in the experimental condition, this difference between the two conditions did not reach statistical significance. However, we observed lower Brier scores in the Twoside condition than in the control condition, when subjects did not know the correct answers to the questions. The linear mixed-effect models we ran to discern the effects of our experimental manipulation on subjects’ confidence showed an effect of Twoside instructions on subjects’ confidence, but this effect disappeared or diminished when other variables such as the number of reasons against preferred answers or the number of reasons for preferred answers were included in the model. A mediation analysis showed that the effect of Twoside on subjects’ confidence is mediated by its effect on the aforementioned variables. We next report our results in more detail.

Results

Individual Differences Measures for Actively Open Minded Thinking

We calculated the individual differences measures as described in Study 2. The subjects’ AOT-Beliefs scores ranged from -1.626 to 1.364 (M = 0.789, SD = 0.636), while their mean AOT-Reasons scores ranged from 0.50 to 1.40 (M = 0.82, SD = 0.188). A
correlation between AOT-Beliefs and mean AOT-Reasons scores was not statistically significant, \( r = -0.07, t(61) = -0.55, p = 0.59, \) two-tailed.

Next we ran a correlation between subjects’ mean Brier scores and AOT-Beliefs. There was a small and statistically non-significant correlation between these two variables in the expected direction, \( r = -0.12, t(60) = -0.90, p = 0.19, \) one-tailed. We expected a negative correlation between mean AOT-Reasons and mean Brier scores, showing that lower Brier scores would be associated with higher AOT-Reasons scores, but the correlation we observed between these two variables was small and not statistically significant, \( r = -0.14, t(60) = -1.08, p = 0.14, \) one-tailed.

Finally, we looked at how the correlation between Brier scores and mean AOT-Reasons behaved as the question difficulty varied. We observed a large correlation in the expected direction, \( r = -0.52, t(18) = -2.62, p = 0.01, \) two-tailed. Our hypothesis that as the questions get more difficult, the correlation between Brier scores and AOT reasoning get stronger, was supported.

**Linear Mixed-Effect Models.** Next we ran linear mixed-effect models to assess the effect of actively open-minded measures on subjects’ Brier scores. We first regressed the Brier scores of subjects on AOT-Reasons with subjects and items as crossed random effects. We did not observe a statistically significant effect of the measure on subjects’ Brier scores (-0.035, 95% c.i. -0.086 to 0.016, Wald method).

We additionally regressed the Brier scores of subjects over AOT-Reasons and Correct (whether subjects guessed the answer correctly where 1 = correct and 0 =
incorrect with subjects and items as crossed random effects). AOT-Reasons had an effect on subjects’ Brier scores (-0.063, 95% c.i. -0.098 to -0.029, Wald method), and as it would be expected Correct did have an effect on the Brier scores (-0.486, 95% c.i. -0.521 to -0.452, Wald method). We also observed an interaction effect of AOT-Reasons and Correct on the Brier scores (0.110, 95% c.i. 0.076 to 0.144, Wald method).

Next we split our data into two based on whether subjects guessed a specific question correctly or incorrectly. For each dataset, we regressed subjects’ Brier scores on AOT-Reasons with subjects and items as crossed random effects. The results showed a good effect when subjects guessed the question incorrectly such that higher AOT-Reasons scores were associated with lower Brier scores (-0.130, 95% c.i. -0.200 to -0.059, Wald method). However, when subjects answered the question correctly, we observed a bad effect of AOT-Reasons on subjects’ Brier scores such that higher AOT-Reasons scores were associated with higher Brier scores (0.066, 95% c.i. 0.030 to 0.102, Wald method). Finally, we ran a generalized linear mixed-effect model to determine any effect of AOT-Reasons on Correct but we did not observe any effect of AOT-Reasons on Correct (-0.111, 95% c.i. -0.391 to 0.169, Wald method). These results are similar to the results we observed in previous studies; answering a question correctly is independent of having

---

16 The lmer function used in R was lmer(BS ~ Ar+(1+Ar|S)+(1+Ar|C),d1[d1$Correct<0,]), where BS is Brier scores, Ar is AOT-Reasons, S is subjects, C is items, and d1[d1$Correct<0,] is the data frame that only contained data for incorrectly answered questions.

17 The lmer function used in R was lmer(BS ~ Ar+(1+Ar|S)+(1+Ar|C),d1[d1$Correct>0,]), where BS is Brier scores, Ar is AOT-Reasons, S is subjects, C is items, and d1[d1$Correct>0,] is the data frame that only contained data for correctly answered questions.
high or low AOT-Reasons scores, and they shed some light onto the stronger relation between Brier scores and AOT-Reasons when questions are more difficult.

An additional linear mixed-effect model analysis did not show a statistically significant effect of AOT-Beliefs on subjects’ Brier scores either (-0.031, 95% c.i. -0.099 to 0.037, Wald method).

**Overconfidence and Actively Open-Minded Thinking**

We used two overconfidence measures in assessing the relations between AOT and overconfidence. The first measure, Over1, compared subjects’ probability judgments for their preferred answers to the percentage of correct answers. The second measure, Over2, compared subjects’ expected Brier scores (EBS, see the results section of Studies 1A and 1B for an explanation of this measure) to their actual Brier scores. For both measures, the mean overconfidence scores were positive, meaning that subjects were overconfident in their probability judgments ($M_{Over1} = 0.09$; $M_{Over2} = 0.35$).

We hypothesized that subjects who had higher AOT-Reasons scores would have lower overconfidence. We observed no statistically significant correlations between AOT-Reasons and each overconfidence measure. The correlation coefficient between AOT-Reasons and Over1 as the overconfidence measure was $-0.02$, $t(60) = -0.16$, $p = 0.44$, one-tailed. The correlation between Over2 and AOT-Reasons was small and not statistically significant, $r = -0.18$, $t(60) = -1.41$, $p = 0.08$, one-tailed. Similarly, a correlation between either overconfidence measure and AOT-Beliefs scores was not
statistically significant either (with Over1 as the overconfidence measure, $r = -0.18$, $t(60) = -1.44$, $p = 0.08$, one-tailed; with Over2, $r = -0.12$, $t(60) = -0.94$, $p = 0.17$, one-tailed).

**Linear Mixed-Effect Models.** We regressed the new overconfidence measure (Over2) on AOT-Reasons with subjects and items as crossed random effects. Our analysis showed a significant effect of AOT-Reasons on Over2 (-0.073, 95% c.i. -0.121 to -0.027, Wald method). Subjects who wrote more reasons that went against their preferred answers were less overconfident about their answers. However, we did not observe a statistically significant effect of AOT-Beliefs on Over2 (-0.039, 95% c.i. -0.118 to 0.041) even though the sign of the regression coefficient was in the expected direction.

We also regressed subjects’ probability judgments for their preferred answer (Prob from now on) on AOT-Reasons with items and subjects as crossed random effects.18 We observed a statistically significant effect of AOT-Reasons on Prob such that higher AOT-Reasons scores were associated with lower probability judgments (-0.067, 95% c.i. -0.091 to -0.042, Wald method). A linear mixed-effect model regressing Prob on AOT-Reasons and Correct with items and subjects as crossed random effects showed an effect of AOT-Reasons (-0.065, 95% c.i. -0.090 to -0.040, Wald method) and an effect of Correct on Prob (0.024, 95% c.i. 0.014 to 0.034, Wald method).19 As also observed in Study 2, scoring high on AOT-Reasons was associated with lower probability judgments

---

18 The lmer function used in R was `lmer(Prob ~ Ar + (1+Ar|C) + (1+Ar|S),d1)`, where Ar is AOT-Reasons, C is items, S is subjects, and d1 is the data frame used.

19 The lmer function in R for this model was `lmer(Prob ~ Ar + Correct + (1+Ar|C) + (1+Ar|S),d1)`, where C is items, S is subjects, and d1 is the data frame used.
for preferred answers, and answering questions correctly was associated with higher probability judgments for preferred answers.

**Effect of Twoside Instructions on Subjects’ Reasons**

We first wanted to check whether our subjects followed the instructions and wrote at least one argument that was against their preferred answer when they were asked to do so. Based on the previous studies, we knew that most subjects do not list reasons that go against their preferred answers, so we thought that running a two sample t-test to see whether their AOT-Reasons score were higher in Two-sided condition were higher than their AOT-Reasons score in the control condition would be an appropriate way of doing this. We calculated average AOT-Reasons scores for subjects in the control and Two-sided conditions, and ran a one-tailed two sample t-test. We expected the mean AOT-Reasons scores in the Two-sided condition to be higher than those in the control condition if subjects followed the instructions in the study. This turned out to be the case, $t = 12.93, p < 0.001, N = 126$, one-tailed ($M_{control} = 0.53, M_{twoside} = 1.10$).

We additionally looked at whether instructing subjects to list at least one reason that went against their preferred answer would affect the number of reasons they list. In this study, subjects were told to give at least two reasons for each question, and sometimes they were required to list one reason that was against their answer. Therefore, there was nothing in the instructions that specifically asked subjects to list more than two reasons in any case. A question we had was whether prompting a change in the kind of reasons they write would also lead them to make a more thorough search for evidence
about their answer, and thus writing more reasons overall. We compared the average number of reasons written by each subject in both conditions. The result of a two sample t-test showed that our hypothesis was supported, $t = 2.01$, $p = 0.02$, $N = 126$, one-tailed ($M_{\text{Control}} = 2.38$, $M_{\text{Twoside}} = 2.60$). We are not sure exactly how the Two-sided condition increases the number of reasons submitted. One reason might be that because people were forced to think against their answer, they had to first go through the reasons that supported their answer until they could come up with a reason that met the expectations. Another way the instructions might have led to this increase in number of reasons would be that the challenge of having to come up with a reason that went against their preferred answer made the subjects more likely to search for additional evidence.

**Effect of Twoside Instructions on Brier Scores**

Next we wanted to see whether instructing people to write reasons that went against their answers would lower their Brier scores. To test this hypothesis, we calculated the mean average Brier scores for each condition and ran a two sample t-test. The mean Brier score for the Two-sided condition was lower but not significantly so, $t = -1.01$, $p = 0.16$, $N = 124$, one-tailed ($M_{\text{Control}} = 0.66$, $M_{\text{Twoside}} = 0.63$).

**Linear Mixed-Effect Models.** We regressed subjects’ Brier scores on Twoside with subjects and items as crossed effects, and observed that Twoside instructions were associated with lower Brier scores, but this effect was not statistically significant (-0.031, 95% c.i. -0.085 to 0.023, Wald method). Regressing Brier scores on Correct and Twoside also showed an effect of Correct (-0.422, 95% c.i. -0.451 to -0.395, Wald method),
Twoside (-0.042, 95% c. i. -0.080 to -0.005, Wald method), as well as an interaction effect of Correct and Twoside on Brier scores (0.057, 95% c. i. 0.019 to 0.095, Wald method). Lower Brier scores were associated with Twoside instructions and correct answers. The interaction effect showed that the effect of Twoside on Brier scores varies with whether subjects knew the correct answer or not. To investigate further, we divided our data into two so that we could test the effect of Twoside instructions on Brier scores when subjects were correct and incorrect. When subjects were incorrect, Twoside had a big effect on Brier scores (-0.083, 95% c. i. -0.139 to -0.028, Wald method). When subjects were correct, Twoside again had an effect on Brier scores, but this effect was smaller, not statistically significant and in the opposite direction (0.027, 95% c. i. -0.002 to 0.055, Wald method). It seems that Twoside instructions are most beneficial when subjects do not know the correct answers, and it might be even hurtful if subjects know the correct answers.

**Effect of Twoside Instructions on Overconfidence**

Finally, we looked at whether the manipulation made subjects less overconfident.

We first compared Over1 scores for each subject between the two conditions. Even though Over1 scores were lower in the Twoside condition than those in the control condition, this difference between the conditions was not statistically significant, $t = -0.83$, $p = 0.20$, $N = 124$, one-tailed ($M_{Control} = 0.11$, $M_{Twoside} = 0.08$). We next compared the mean Over2 scores for each subject between the two conditions. A two-sample t-test showed that the overconfidence was slightly lower in the Twoside condition, but this
result was again not statistically significant, \( t = -1.41, p = 0.08, N = 124 \), one-tailed \((M_{\text{Control}} = 0.38, M_{\text{Twoside}} = 0.32)\).

**Linear Mixed-Effect Models.** We ran additional regression models to understand the relationship between the Twoside instructions and subjects’ confidence. First, we regressed \( \text{Prob} \) (subject’s probability judgment for their preferred option) on Twoside (whether the question was answered in the experimental (=1) or control (=0) conditions), and observed a statistically significant effect of Twoside on \( \text{Prob} \) such that subjects assigned lower probabilities to their preferred options if they answered the question in Twoside condition (-0.039, 95% c.i. -0.060 to -0.018, Wald method).\(^{20}\)

Next, we regressed \( \text{Prob} \) on \( \text{Ar} \) (the number of reasons against the subject’s preferred option), \( \text{Arb} \) (the number of reasons for the subject’s preferred option), \( \text{Correct} \) (whether the subject answered the question correctly (=1) or not (=0)), and Twoside with items and subjects as crossed random effects.\(^{21}\) We observed significant effects of \( \text{Ar} \) (-0.048, 95% c.i. -0.76 to -0.020, Wald method), \( \text{Arb} \) (0.035, 95% c.i. 0.015 to 0.055, Wald method), and \( \text{Correct} \) (0.022, 95% c.i. 0.007 to 0.037, Wald method) but no significant effect of Twoside on \( \text{Prob} \) (-0.000, 95% c.i. -0.025 to 0.024, Wald method).

Higher number of reasons for subjects’ preferred options and answering the questions correctly were associated with higher confidence, while higher number of reasons against subject’s preferred options was associated with lower confidence. The Twoside

\[^{20}\] The \texttt{lmer} function used in R was \( \texttt{lmer} (\text{Prob} \sim \text{Twoside} + (1 + \text{Twoside}|C) + (1 + \text{Twoside}|S), \texttt{d1}) \), where \( C \) is items, \( S \) is subjects, and \( \texttt{d1} \) is the data frame.

\[^{21}\] The \texttt{lmer} function used in R was \( \texttt{lmer} (\text{Prob} \sim \text{Ar} + \text{Correct} + \text{Arb} + \text{Twoside} + (1|C) + (0 + \text{Ar} + \text{Correct} + \text{Arb}|C) + (1|S) + (0 + \text{Ar} + \text{Correct} + \text{Arb}|S), \texttt{d1}) \), where \( C \) is items, \( S \) is subjects, and \( \texttt{d1} \) is the data frame.
manipulation did not have an effect on subjects’ probability judgment for their preferred answer when additional variables were added to the model.

Additionally, we observed that Twoside did not interact with anything or did not have any effect on Prob on itself as long as Ar and Arb are included in the model. Higher number of reasons for subjects’ preferred options were associated with higher confidence, while higher number of reasons against subjects’ preferred options were associated with lower confidence. We regressed Prob on Ar, Arb, Twoside, the interaction between Twoside and Ar, and the interaction between Twoside and Arb with items and subjects as crossed random effects.\textsuperscript{22} Once again we observed significant effects of Ar (-0.052, 95\% c.i. -0.088 to -0.017, Wald method) and Arb (0.042, 95\% c.i. 0.017 to 0.067, Wald method), but no significant effect of Twoside on Prob (-0.012, 95\% c.i. -0.098 to 0.074, Wald method). There was also no interaction of Twoside with Ar (0.026, 95\% c.i. -0.037 to 0.089, Wald method) or with Arb (-0.010, 95\% c.i. -0.036 to 0.017, Wald method).

While Twoside instructions affect the number of against reasons people write for their preferred answer, it does not seem to change any of the relationships that also exist.

**Mediation Analysis.** While we observed an effect of Twoside on Prob when Twoside is the only independent variable in the linear-mixed effect model, this effect disappeared when we added other independent variables to the model. A possible explanation is that the effect of Twoside is mediated by its effect on Ar and Arb variables. Therefore, we decided to run a simple mediation model as recommended by Baron and

\textsuperscript{22} The lmer function used in R was lmer(Prob ~ (Ar + Arb) * Twoside + (1|C) + (0 + Ar + Arb|C) + (0 + Ar * Twoside|C) + (0 + Arb * Twoside|C) + (1|S) + (0 + Ar + Arb|S) + (0 + Ar * Twoside|S) + (0 + Arb*Twoside|S), d1), where C is items, S is subjects, and d1 is the data frame.
Kenny (1986). We already established above that Twoside has a statistically significant effect on Prob. Next, we regressed Ar on Twoside. We observed an effect of Twoside on Ar such that when subjects answered the questions in Twoside condition, they wrote more reasons that were against their preferred answers (0.574, 95% c.i. 0.482 to 0.665, Wald method). We also regressed Arb on Twoside, and observed a significant effect of Twoside on Arb such that when subjects answered the questions in Twoside condition, they wrote fewer reasons that were for their preferred answers (-0.257, 95% c.i. -0.370 to -0.143, Wald method). Finally, we regressed Prob on Ar and Twoside, and observed a statistically significant effect of Ar on Prob (-0.069, 95% c.i. -0.101 to -0.037, Wald method), but no effect of Twoside on Prob (0.003, 95% c.i. -0.027 to 0.032, Wald method). When we regressed Prob on Arb and Twoside, we observed a statistically significant effect of Arb on Prob (0.048, 95% c.i. 0.028 to 0.068, Wald method) and a statistically significant effect of Twoside on Prob (-0.026, 95% c.i. -0.046 to -0.005, Wald method). Twoside’s effect on Prob disappeared when Ar was included in the model, and the beta coefficient for Twoside has become smaller when Arb was included in the model even though it still had a significant effect on Prob. Thus, it seems that Ar and Arb mediate the effect of Twoside on Prob.

23 The lmer function used in R was lmer(Ar ~ Twoside + (1 + Twoside|C) + (1 + Twoside|S), d1), where C is items, S is subjects, and d1 is the data frame.
24 The lmer function used in R was lmer(Arb ~ Twoside + (1 + Twoside|C) + (1 + Twoside|S), d1), where C is items, S is subjects, and d1 is the data frame.
25 The lmer function used in R was lmer(Prob ~ Ar + Twoside + (1 + Ar + Twoside|C) + (1 + Ar + Twoside|S), d1), where C is items, S is subjects, and d1 is the data frame.
26 The lmer function used in R was lmer(Prob ~ Arb + Twoside + (1 + Arb + Twoside|C) + (1 + Arb + Twoside|S), d1), where C is items, S is subjects, and d1 is the data frame.
**Carryover Effects.** Given the design of our study, it is possible that there might have been some carryover effects as subjects switched between control and training conditions throughout the task. To test for any potential carryover effects, we regressed $Ar0$ (the number of against reasons subjects wrote in control condition) and $Ar1$ (the number of against reasons subjects wrote in Twoside condition) on the position of each item in the presentation order (Item), and the number of Twoside items through the current item (Twoside.cumsum) with subjects as crossed random effects. We did not observe any carryover effects in either regressions. There was no effect of Item (0.009, 95% c.i. -0.023 to 0.041, Wald method) or Twoside.cumsum (-0.011, 95% c.i. -0.071 to 0.049, Wald method) on $Ar0$. Similarly, there was no effect of Item (0.008, 95% c.i. -0.010 to 0.026, Wald method) or Twoside.cumsum (-0.013, 95% c.i. -0.048 to 0.022, Wald method) on $Ar1$.

**Study 5**

The aim of this study was to investigate whether forcing subjects to write reasons that considered why their estimates might be too high or too low would make their estimates more accurate. This study was similar to Study 4 with the exception that we used numerical judgment questions rather than multiple choice questions.

**Method**

**Subjects.** Seventy-six subjects participated in the study. The subjects’ age ranged from 21 to 71 (Median = 46); 57.6% were female. The make-up of the subjects in this study were the same as those in previous studies. Subjects were paid $9 for their
participation (through PayPal). A panel of about 200 subjects was notified by email when the study was ready, and the study was removed when we had about 75 subjects.

**Questions.** Subjects answered 20 questions in an order randomized for each subject. The questions came from the pool of questions used in Pilot Study 3, and were picked as described in the results section of that study. A full list of questions and answers along with their difficulty and discrimination scores can be found in Appendix B-1.

**Procedure.** Subjects did the task on their own computers over the Internet. They first read the instructions regarding the task (see Appendix C-1). The instructions and the full study can be found on http://finzi.psych.upenn.edu/~baron/ex/bg/args/args8.html.

The procedure for this study was similar to that of Study 4 with the exception that we used numerical questions rather than a probability task. There were two conditions: (1) a control condition where subjects would answer questions and list arguments as in Study 4, and list at least two reasons but they were not required to give reasons why the correct answer could be outside or inside of their interval; (2) an experimental condition where subjects had to again list two reasons but this time, one of the reasons had to be about why the correct answer could be outside their interval, and one reason why the correct answer could be inside their interval. All subjects did the task in both conditions, and answered half the questions in the control condition, and the other half in the experimental condition. Half of the subjects did the task where the sequence of control and experimental conditions were completely randomized. The other half of the subjects did the task where the first five items were in the control condition, while for items 6
through 20, they did ten of the questions were in the experimental condition, and five of them in the control condition. Subjects were assigned to these two different sequences randomly. The reason for the different sequences was to test for possible carry-over effects due to the within-subject design of the study.

Upon answering all the questions, the subjects were again presented with the AOT scale (see Table 2). After the data collection was completed, the classifications of reasons were checked by one of the authors (BG) to ensure that subjects’ reasons were classified correctly.

**Results and Discussion**

This study tried to force people to write more reasons against why their estimates could be inaccurate. After looking at the data, we saw that the study did not work. The flaw of the experimental design was that subjects’ AOT-Reasons scores were pretty high even without the experimental manipulation, which asked people to think of one reason why the correct estimate could be within their given confidence interval, and one reason why the correct estimate would be outside of their given confidence interval. Many of the subjects’ reasons were about why their interval was too small before the manipulation. Exposing them to the experimental manipulation made them give reasons why their interval was not too small. Thus, the instructions increased the number of reasons given by subjects, but there was no increase in subjects’ AOT-Reasons scores. This also made AOT-Reasons not correlate with any of the other variables.
Study 6

The aim of this study was to see whether a training module could teach subjects to be more actively open-minded thinkers, and make them more accurate on a probability task.

Method

Subjects. One hundred and sixty-eight subjects participated in the study. The subjects’ age ranged from 20 to 79 (Median = 50); 71.4% were female. The make-up of the subjects in this study were the same as those in previous ones. Subjects were paid $16 for their participation (through PayPal). A panel of about 400 subjects were notified by email when the study was ready, and the study was removed after 4 days have passed.

Training. The training module consisted of three parts: (1) Introduction to the concept of actively open-minded thinking and relevant vocabulary; (2) Introduction to the concept of myside bias and ways of avoiding this bias; (3) An exercise to help subjects test their thinking before attempting the upcoming task. The full training can be found in Appendix D-1.

In the first part of the training subjects saw short paragraphs explaining what good thinking is, and were taught relevant vocabulary such as “possibility”, “evidence”, and “conclusion” within the context of judgments. After the textbook introduction subjects did an exercise that asked them to categorize a hypothetical person’s thought process. Subjects had to decide whether the person was stating a possibility or evidence, drawing
a conclusion, or searching for possibility or evidence. After completing this exercise, they were shown their selected responses and given feedback.

In the second part of the training subjects read short paragraphs and examples, and learned about myside bias and ways people commit this bias.

In the third part of the training subjects had the opportunity to test their own thinking by answering at least 3 of the 6 three-choice choice questions presented to them. The questions in this section included a variety of topics such as country population and movie release dates. For each question subjects chose to answer, they had to first write down their thought process, then pick an answer, and finally give probability judgments for each option. After seeing the first three questions, subjects were presented with a reminder page where we modeled to them how someone who was using the principles of AOT and someone who was committing myside bias could think about the first question. After this reminder subjects continued with the third part of the training. At the end of the exercise questions, subjects were shown the reasons they listed, their preferred answer, and the correct answer, and were asked to self-evaluate their responses to see how successful they were applying the principles of actively open-minded thinking. Upon completion of this part, subjects were given a seven-item survey asking them about their experience of learning about AOT, and the subjects agreed or disagreed on a 5-point scale (1 = Strongly Disagree, 3 = Neither Agree Nor Disagree, 5 = Strongly Agree). After completing the short survey subjects were given the chance to submit any additional comments, and were forwarded to the next part of the study.
Questions. The questions were the same as those used in Study 2. A full list of questions and answers along with their difficulty and discrimination scores can be found in Appendix B-1.

Procedure. Subjects did the task on their own computers over the Internet. They first read the instructions regarding the task (see Appendix C-1). The instructions and the full study can be found on http://finzi.psych.upenn.edu/~baron/ex/bg/args/args9.html.

Subjects were randomly assigned either to a training condition or to a control condition ($N_{control} = N_{training} = 84$). Subjects who were in the training condition first completed the training described above, and moved onto the second part of the study, where they answered 20 questions. Those who were in the control condition did the third part of the training without self-evaluation (see Appendix E), and had to answer at least 4 questions. After completing this part, these subjects also moved onto the second part of the study where the same 20 questions. Subjects in both conditions were given 36 hours to complete both parts of the study.

In the second part of the study subjects answered one question on each page. Below the short information the subjects saw the question with three options, and underneath that they were given six note spaces to list their reasons for and against their preferred answer. They were not required to use all six text input spaces but they had to write at least two reasons. Subjects picked a preferred answer after they listed their reasons. They were also given choices for probabilities to indicate their confidence in any answer option being correct, and were instructed to choose the one that was closest to
what they think. The options for probability judgments were 0, 1, 5, 10, 20, 30, 40, 50, 60, 70, 90, 95, 99, and 100 (in percent). Subjects knew that 100% would mean they were completely certain that an answer was correct, but they would have a probability of 33% being correct if they were guessing. They were also informed on the instruction page that their probability judgments for the three answers would need to add up to about 100% (no less than 90% and not greater than 110% to allow for the limited set of options). After completing the page, they could move onto the next page by clicking a button where they could classify their reasons as being for or against the choice options. Once they completed the classification task, they moved onto the next page, again by clicking a button, where they were presented with a new question and they repeated the same procedure for the remaining questions.

After answering all questions, the subjects were presented with an AOT scale (see Table 2). They read 11 statements regarding beliefs about thinking, and were asked to indicate their agreement with each statement on a 5-point scale (1 = completely agree, 5 = completely disagree). Once the subjects completed the survey, they clicked a button to submit their responses. As in the previous studies, after the data collection was completed, the classifications of reasons were checked by one of the authors (BG) to ensure that subjects’ reasons were classified correctly.

**Summary and Discussion of Results**

This study attempted to teach people principles of AOT and how to reduce myside bias. Our training module was partially successful. We found some evidence showing that
going through the training module made subjects write more reasons, and also more reasons that went against their preferred answers. Even though subjects in the training condition had lower overconfidence and lower Brier scores, these differences were not statistically different from those of the subjects’ in the control condition. Linear mixed-effect models showed an effect of AOT on overconfidence such that subjects who score high on the AOT measures had lower overconfidence.

A criticism of this study concerns the training module. Even though the module explained the advantages of using alternative reasons to subjects, there was neither much incentive for subjects to push themselves to list alternative reasons nor enough opportunities to practice the newly learned thinking skills during the training exercises. Due to these reasons we decided to modify our training module and replicate the study.

Additionally, AOT-Reasons was not as useful as a measure in this study as it had been previously. One reason might be a slight change we did to the way subjects submitted their responses and reasons. Previous studies asked subjects to indicate their preferred answer first, and then list their reasons before they were given a chance to revise their answer. In this study we asked subjects to read the question, list reasons for or against any answer options, and then ask them to pick an answer. This might have led people to behave differently and make them follow their reasons to pick an answer, which would decrease their AOT-Reasons scores.

However, the lack of any significant AOT-Reasons effects on Brier scores is not a new observation, as we have seen similar results in previous studies as well. Therefore,
we conducted additional analyses to discern the underlying mechanism for these results as we did in previous studies, and got comparable results. Once again further analyses showed that the effect of this measure on the Brier scores is dependent on whether subjects were correct or not, such that when subjects answered the questions incorrectly, scoring higher on AOT measure from reasons was associated with lower Brier scores, while in the cases where subjects answered the questions correctly, scoring higher on AOT measure from reasons was associated with higher Brier scores. Additionally, subjects’ answering a question correctly was independent of what their AOT-Reasons scores were. These analyses explain the correlation we have observed over multiple studies where the correlation between Brier scores and the AOT from measure scores got stronger as the question difficulty increased. In the section below we describe our results in further detail.

Results

We omitted six subjects from data analysis due to various reasons. One subject did not complete the first part of the studies. Three subjects went through the training module too fast, and two subjects gave inadequate reasons. We also excluded seven additional subjects from the AOT-Reasons calculations and any tests run involving this measure, because these subjects wrote the answer they picked rather than providing reasons for their choices.
Individual Differences Measures for Actively Open Minded Thinking

We calculated the individual differences measures as described in Studies 1A and 1B. The subjects’ AOT-Beliefs scores ranged from -1.182 to 1.909 (\(M = 0.711\), \(SD = 0.602\)), while their mean AOT-Reasons scores ranged from 0.00 to 1.89 (\(M = 0.41\), \(SD = 0.345\)). A correlation between AOT-Beliefs and mean AOT-Reasons scores was not statistically significant, \(r = 0.05\), \(t(158) = 0.61, p = 0.54\), two-tailed.

We observed a moderate correlation between subjects’ mean Brier scores and AOT-Beliefs, \(r = -0.30\), \(t(164) = -4.08, p < 0.001\), one-tailed. Subjects who scored higher on the AOT scale had lower Brier scores. We expected that scoring higher on AOT-Reasons would be correlated with lower Brier scores, but this hypothesis was not supported, \(r = 0.02\), \(t(156) = 0.48, p = 0.60\), one-tailed. Finally, we looked at how the correlation between Brier scores and mean AOT-Reasons scores behaved as the question difficulty varied. In the earlier studies, we observed that the relation between Brier scores and the mean AOT-Reasons scores strengthens as the question difficulty increases. However, we did not observe this relation in the current study, \(r = -0.06\), \(t(17) = -0.26, p = 0.79\), two-tailed.

Linear Mixed-Effect Models. We ran linear mixed-effect models to look at the effects of AOT measures on subjects’ Brier scores. We used subjects and items as crossed random effects. While we observed an effect of AOT-Beliefs on Brier scores (-0.110, 95% c.i. -0.163 to -0.057, Wald method), we did not observe this effect when we regressed Brier scores on AOT-Reasons (-0.000, 95% c.i. -0.027 to 0.027, Wald method).
One possible reason why AOT-Reasons did not have any effects on subjects’ Brier scores could be that the usefulness of considering alternative reasons might be whether a subject can guess the answer correctly. If a subject knows the correct answer, they might not feel the need to list alternative reasons. We had found some support for this hypothesis in previous studies, so we wanted to test it in this study as well given the similarity of other results regarding the effects of AOT on Brier scores. We regressed the Brier scores of subjects over AOT-Reasons and Correct (whether subjects guessed the answer correctly where 1 = correct and 0 = incorrect with subjects and items as crossed random effects). AOT-Reasons on its own had no effect on subjects’ Brier scores (-0.008, 95% c.i. -0.026 to 0.011, Wald method), and as it would be expected Correct did have an effect on the Brier scores (-0.385, 95% c.i. -0.399 to -0.371, Wald method). We also observed an interaction effect of AOT-Reasons and Correct on the Brier scores (0.059, 95% c.i. 0.041 to 0.092, Wald method).

Next we split our data into two based on whether subjects guessed a specific question correctly or incorrectly as we did in previous studies. For each dataset, we regressed subjects’ Brier scores on AOT-Reasons with subjects and items as crossed random effects. The results showed a good effect when subjects guessed the question incorrectly such that higher AOT-Reasons scores were associated with lower Brier scores (-0.092, 95% c.i. -0.130 to -0.053, Wald method). However, when subjects answered the question correctly, we observed a bad effect of AOT-Reasons on subjects’ Brier scores.

27The lmer function used in R was lmer(BS ~ Ar+(1|Ar[S])+(1|Ar[C]),d1[d1$Correct<0,]), where BS is Brier scores, Ar is AOT-Reasons, S is subjects, C is items, and d1[d1$Correct<0,] is the data frame that only contained data for incorrectly answered questions.
such that higher AOT-Reasons scores were associated with higher Brier scores (0.057, 95% c.i. 0.040 to 0.075, Wald method). Finally, we ran a generalized linear mixed-effect model to determine any effect of AOT-Reasons on Correct but we did not observe any effect of AOT-Reasons on Correct (-0.122, 95% c.i. -0.270 to 0.026, Wald method). It seems that subjects’ being correct on a given question is not dependent upon subjects’ AOT-Reasons scores. These analyses explain the strengthening of the correlation between Brier scores and AOT-Reasons scores as the difficulty of questions increases. Since subjects are more likely to answer difficult questions wrong, writing more alternative reasons as to why their preferred option might be wrong lowers their Brier scores.

**Overconfidence and Actively Open-Minded Thinking**

We again used two overconfidence measures in assessing the relations between AOT and overconfidence. The first measure, Over1, compared subjects’ probability judgments for their preferred answers to the percentage of correct answers. The second measure, Over2, compared subjects’ expected Brier scores (EBS, see the results section of Studies 1A and 1B for an explanation of this measure) to their actual Brier scores. For both measures, the mean overconfidence scores were positive, meaning that subjects were overconfident in their probability judgments ($M_{Over1} = 0.10; M_{Over2} = 0.18$).

We expected lower overconfidence to be associated with higher scores on AOT-Reasons, but we did not observe a correlation between AOT-Reasons and either overconfidence measure (with Over1 as the overconfidence measure: $r = -0.01$, $t(156) =$

---

28 The lmer function used in R was `lmer(BS ~ Ar+(1+Ar|S)+(1+Ar|C),d1[d1$Correct<0,])`, where BS is Brier scores, Ar is AOT-Reasons, S is subjects, C is items, and `d1[d1$Correct>0,]` is the data frame that only contained data for correctly answered questions.
-0.12, \( p = 0.45 \), one-tailed; with Over2 as the overconfidence measure: \( r = -0.05, t(156) = -0.60, p = 0.28 \), one-tailed). However, we did observe a negative correlation between each overconfidence measure and AOT-Beliefs scores, showing that higher scores on AOT-Beliefs was associated with lower overconfidence (with Over1 as the overconfidence measure: \( r = -0.18, t(164) = -2.40, p = 0.01 \), one-tailed; with Over2 as the overconfidence measure: \( r = -0.20, t(164) = -2.55, p = 0.01 \), one-tailed).

**Linear Mixed-Effect Models.** We looked at the effects of AOT on overconfidence. Both measures of AOT showed an effect on overconfidence such that higher scores on these measures were associated with lower overconfidence. We regressed the variable Over2 (the difference between mean actual Brier scores and mean expected Brier scores) on AOT-Reasons with subjects and items as crossed random effects. Our analysis showed a significant effect of AOT-Reasons on the overconfidence measure (-0.045, 95% c.i. -0.074 to -0.016, Wald method), meaning that subjects who wrote more reasons that went against their preferred answers were less overconfident about their answers. Similarly, regressing Over2 on AOT-Beliefs with subjects and items as crossed random effects also showed a significant effect of AOT-Beliefs on the overconfidence measure (-0.089, 95% c.i. -0.168 to -0.021, Wald method). Subjects who scored higher on the AOT scale showed lower overconfidence in their answers.

We also regressed subjects’ probability judgments for their preferred answer (Prob from now on) on AOT-Reasons with items and subjects as crossed random effects.\(^{29}\) We

\(^{29}\) The `lmer` function used in R was `lmer(Prob ~ Ar + (1+Ar|C) + (1+Ar|S),d1)`, where `Ar` is AOT-Reasons, `C` is items, `S` is subjects, and `d1` is the data frame used.
observed a statistically significant effect of AOT-Reasons on Prob such that higher AOT-Reasons scores were associated with lower probability judgments (-0.054, 95% c.i. -0.068 to -0.040, Wald method). A linear mixed-effect model regressing Prob on AOT-Reasons and Correct with items and subjects as crossed random effects showed an effect of AOT-Reasons (-0.053, 95% c.i. -0.067 to -0.040, Wald method) and an effect of Correct on Prob (0.019, 95% c.i. 0.013 to 0.025, Wald method). Scoring high on AOT-Reasons was associated with lower probability judgments for preferred answers, and answering questions correctly was associated with higher probability judgments for preferred answers. These results are similar to those we observed in previous studies.

**Correlations Between Accuracy Scores, Number of Reasons, and Actively Open-Minded Thinking**

We also looked at the relation between subjects’ mean number of reasons they wrote (Nreas.s from now on) and their Brier scores. We hypothesized that higher number of reasons written would be correlated with lower Brier scores. We observed a small but statistically significant correlation between these two variables, $r = -0.16, t(164) = -2.04, p = 0.02$, one-tailed. We also observed a small but statistically significant correlation between Nreas.s and AOT-Beliefs, $r = 0.13, t(166) = 1.74, p = 0.04$, one-tailed. Scoring high on the AOT scale was weakly associated with writing more reasons while answering the questions in the study.

---

30 The lmer function in R for this model was `lmer(Prob ~ Ar + Correct + (1+Ar|C) + (1+Ar|S), d1)`, where C is items, S is subjects, and d1 is the data frame used.
Comparisons Between the Control and Training Conditions

We finally investigated whether the AOT training had any effects on subjects’ Brier scores, overconfidence, the number of reasons subjects listed, AOT-Reasons, and AOT-Beliefs. We expected the subjects in the training condition to have lower Brier scores but did not observe any differences in Brier scores between the conditions, $t = -0.44, p = 0.33, N = 166$, one-tailed ($M_{Control} = 0.66, M_{Training} = 0.65$). We also expected the subjects in the training condition to have lower overconfidence, but this hypothesis was not supported either (using Over1 as the overconfidence measure: $t = -0.002, p = 0.50, N = 166$, one-tailed ($M_{Control} = 0.097, M_{Training} = 0.097$); using Over2 as the overconfidence measure: $t = -0.07, p = 0.47, N = 166$, one-tailed ($M_{Control} = 0.181, M_{Training} = 0.178$)). Since the training module emphasized searching for evidence before making decisions, we expected that subjects in this condition would end up writing more reasons for each answer option than subjects in the control condition. Our hypothesis was supported, $t = 3.11, p = 0.001, N = 168$, one-tailed ($M_{Control} = 2.31, M_{Training} = 2.68$). We hypothesized that subjects in the training condition would also list more reasons that were against their preferred options, and we found support for this hypothesis as well, $t = 1.89, p = 0.03, N = 160$, one-tailed ($M_{Control} = 0.36, M_{Training} = 0.45$). Finally, we wanted to see whether going through the training would increase subjects’ AOT-Beliefs scores, but this hypothesis was not supported, $t = -0.37, p = 0.36, N = 168$, one-tailed ($M_{Control} = 0.73, M_{Training} = 0.69$).
Study 7

The aim of this study was to see whether a training module could teach subjects to be more actively open-minded thinkers, and make them more accurate on a multiple-choice trivia task. Based on the results from the previous study, we changed the training so that it included questions that would encourage the subjects to list arguments that went against their pet beliefs. This study had a within-subject design, and therefore, it included a pre-test.

Method

Subjects. One hundred and thirty-one subjects participated in the study. The subjects’ age ranged from 18 to 76 (Median = 45); 66.4% were female. The subjects were from a panel of about 1200 people who volunteered to do studies for pay on the Internet over the last 15 years, through advertising, links from various web sites, and word of mouth. They were mostly Americans, varying considerably in age, income, and education level, but with women over-represented. Subjects who did not take previous studies seriously had been removed over the years. The panel was divided into three groups in order to use different samples for closely related studies. Subjects were paid $15 if they were in the control condition and $25 if they were in the training condition for their participation (through PayPal). A panel of about 400 subjects were notified by email when the study was ready, and the study was removed after 5 days have passed. None of the subjects who did the previous study participated in this one.
Training. The training module had four parts: (1) Introduction to the training; (2) Introduction to the concept of actively open-minded thinking and relevant vocabulary; (3) Introduction to the concept of myside bias and how to avoid it; (4) An exercise which gave subjects the opportunity to apply what they have learned during their training. The full training can be found in Appendix D-2.

In the first part of the training subjects were informed about the nature of the training, and were presented with two questions to answer. The first question was a percentage question where subjects were asked to make a point estimate by assigning a value between 0 and 100. The second question was a multiple choice question where they had to choose one option. Additionally, for the second question subjects were also asked to make probability judgments for each option’s correctness. For both questions subjects were asked to tell the experimenters how they came up with their answers, and they were allowed to as many reasons as they could. Upon completing these steps they were given a chance to change their responses and probability judgments.

In the second part of the training subjects read about what thinking is, and learned relevant vocabulary such as “possibilities,” “evidence,” “goals,” and “conclusion.” They also read short paragraphs about what actively open-minded thinking is, why AOT is good thinking, and how it is useful.

In the third part of the training subjects first learned about myside bias and how this bias operates. Then they read the responses of two hypothetical respondents who modeled either myside bias or AOT for the two questions subjects answered at the
beginning of the training. Subjects were explicitly told why and how each respondent was displaying myside bias or actively open-minded thinking. After looking at these modeled responses, subjects got to see their own responses to these two questions and the correct answers. They were also asked to self-evaluate their responses in terms of how much myside bias or principles of AOT they displayed. After completing this exercise, subjects took a three-item review test to evaluate how well they understood the concepts of good and bad thinking. Upon answering all three questions, subjects were given feedback regarding their answers to reiterate their learning.

In the fourth and final part of the training, subjects were given six problems to think about. The first three problems were policy problems and subjects had to pick two of these problems to write solutions for. The last three problems were multiple choice questions that were similar to the ones in pre- and post-training surveys, and subjects had to answer all three. For the policy questions subjects had to list at least three solutions (but could list up to 5) to the problem presented and then pick their favorite solution. Afterwards, they were required to list at least two and up to a maximum of eight arguments. Upon listing their arguments, subjects proceeded onto the next step where they had to classify their arguments as “for” or “against” their favored solution or neither. After the classification task, subjects were given the chance to change what their favored solution was before proceeding onto the next question.

After subjects answered the policy questions, they answered the multiple choice questions. For each question subjects had to pick a favored option and then make
probability judgments regarding the correctness of each option. After completing these steps they had to list at least two reasons ‘‘for’’ or ‘‘against’’ their favored answer or neither, and classify their reasons like they did for the policy questions.

Upon answering the five questions, subjects received feedback regarding their answers and were asked to self-evaluate their responses. For the policy questions subjects were shown their solutions, their favored solution, and their arguments. Then they were asked to self-evaluate their reasoning for the shown question in terms of how successfully they applied the principles of AOT. For the multiple choice questions subjects saw the same information and did the same self-evaluation, but they were also told what the correct answers to the questions were.

At the end of the training subjects took a 7-item survey like in the previous training to rate their learning experience on a 5-item scale (1 = Strongly Disagree, 3 = Neither Agree Nor Disagree, 5 = Strongly Agree) and were asked to submit any additional comments they had before being forwarded onto the post-training part of the study as in Study 6.

Questions. The questions were the same as those used in Study 6. A full list of questions and answers along with their difficulty and discrimination scores can be found in Appendix B-1.

Procedure. Subjects did the task on their own computers over the Internet. They first read the instructions regarding the task (see Appendix C-1). The instructions and the first part of the study can be found on http://finzi.psych.upenn.edu/~baron/ex/bg/args/
args10a.html for the control condition, and on http://finzi.psych.upenn.edu/~baron/ex/bg/args/args10b.html for the training condition. The instructions and the last part of the study can be found on http://finzi.psych.upenn.edu/~baron/ex/bg/args/args10post.html.

Subjects were randomly assigned either to a training condition or to a control condition ($N_{control} = 60; N_{training} = 51$ after attrition). Every subject first did the 20-question task and took the AOT scale (see Appendix C). This task was the same as the ones in Study 6. After completing the task, the subjects who were assigned to the training condition were forwarded to the link to go through their training. Upon completing the training, subjects were forwarded to the same 20-question task they did before. Those who were in the control condition were forwarded to a page where they were told to re-answer the questions they did in the first part and that doing this could improve their performance. At the end of the last part all subjects took the AOT scale again. As in the previous studies, after the data collection was completed, the classifications of reasons were checked by one of the authors (BG) to ensure that subjects’ reasons were classified correctly.

Summary and Discussion of Results

In this study we tried to improve on the AOT training module we devised in Study 6. Subjects were given more opportunity to practice AOT skills during the training. We were able to replicate our results from previous studies.

As in most studies, we did not observe a correlation between the AOT measures, and the correlation between the Brier scores and our new AOT measure was small and
not significant. A similar result was also observed in the linear mixed-effect models where we tested the effect of the new measure on the accuracy scores. However, further analyses showed that the effect of this measure on the Brier scores is dependent on whether subjects were correct or not such that when subjects answered the questions incorrectly, scoring higher on AOT measure from reasons was associated with lower Brier scores, while in the cases where subjects answered the questions correctly, scoring higher on AOT measure from reasons was associated with higher Brier scores. Additionally, this AOT measure had no effect on subjects’ being correct on a given question. These analyses explain the correlation we have observed over multiple studies where the correlation between Brier scores and the AOT from measure scores got stronger as the question difficulty increased. These results might also shed some light onto the lack of correlation between the two AOT measures. While AOT-Beliefs indicates one’s overall tendency for good thinking, AOT-Reasons seems to be more task-specific in indicating good thinking behavior.

We used two overconfidence measures to look at the effects of AOT measures on subjects’ overconfidence. Both overconfidence measures indicated that subjects were overconfident in their judgments. Correlations we ran between each overconfidence measure and AOT-Reasons did not yield statistically significant results in the first round data, while the results were statistically significant in the second round data. The linear mixed-effect models we ran showed that higher scores on AOT-Beliefs and AOT-Reasons were associated with lower overconfidence. These results are comparable to the results
we observed in previous studies. We also found some partial support for our hypothesis that going through AOT training reduced overconfidence from the first round to the second round in the study.

Further comparisons of training and control conditions showed that going through the training increased subjects’ scores on both AOT measures. Subjects in the training condition did not have lower Brier scores than those in the control condition, but the number of questions answered correctly were greater for the training condition subjects than the control condition subjects. These results are puzzling, but this might be explained by a combination of two factors. Writing more reasons against preferred answers reduces subjects’ confidence, and therefore, even when the subjects pick the correct answer they do not assign the proper confidence judgment, and/or subjects’ confidence judgments are not accurate enough such that they might show underconfidence when correct as well as overconfidence when incorrect, and since the calculation of Brier scores is a combination of their confidence and ability to answer the question correctly, we might not see an improvement in their accuracy scores based on Brier scores even if we observe improvement in their AOT skills and overconfidence.

In the previous studies we observed that the effect of AOT-Reasons was dependent on whether subjects answered the questions correctly or not. With this idea in mind, we compared the changes in subjects’ Brier scores between the first and second round responses between the two conditions when subjects answered the questions incorrectly, and we observed that when subjects did not know the correct answer to the
question, going through the AOT training helped with their Brier scores. This result was not observed when subjects knew the correct answers.

The t-tests we ran to see whether going through the AOT training would reduce subjects' overconfidence showed that subjects in the control condition showed a smaller change in their overconfidence scores than those subjects in the training condition, but these changes were statistically significant when we used the overconfidence measure that compared subjects’ mean probability judgments for their preferred options to the percentage of questions they answered correctly, but not when we used our new overconfidence measure that took the probability judgments for three options into consideration. Even though our results were not always statistically significant, we observed that going through the AOT training decreased subjects’ overconfidence.

In the section below we go into more detail about the analyses we conducted to reach these conclusions.

**Results**

Since this study had a within-subject design, we used both the first round and the second round data for the analysis of individual differences, and the effects of training were evaluated by comparing the differences between the rounds of the training condition to the differences between the rounds of the control condition. When reporting the results of individual differences analyses, the first set of statistics come from the first round data, and the second set of statistics come from the second round data unless indicated otherwise.
We only excluded one subject from the computation of AOT-Reasons, and any analysis that involves this measure, because this subject failed to provide any relevant reasons for their answers, but instead put in “think so” twice for each question and categorized this reason as support for why they picked that particular option.

**Individual Differences Measures for Actively Open Minded Thinking**

The subjects’ AOT-Beliefs scores ranged from -1.364 to 1.818 \((M_{R1} = 0.719, SD_{R1} = 0.599)\) in the first round, and from -1.273 to 1.909 \((M_{R2} = 0.785, SD_{R2} = 0.677)\) in the second round of the task. The AOT-Reasons scores ranged from 0.00 to 1.50 \((M_{R1} = 0.36, SD_{R1} = 0.348)\), and from 0.00 to 2.00 \((M_{R2} = 0.37, SD_{R2} = 0.394)\) in the second round of the task. A Pearson’s product moment correlation between AOT-Beliefs and mean AOT-Reasons was not statistically significant in either rounds of the data \((r_1 = 0.007, t(129) = 0.08, p = 0.94, \text{two-tailed}; r_2 = 0.01, t(108) = 0.11, p = 0.91, \text{two-tailed})\).

Next we tested whether higher AOT-Beliefs scores were associated with lower Brier scores. There was a moderate correlation between these two variables in the first round data, \(r_1 = -0.31, t(125) = -3.58, p < 0.001, \text{one-tailed}\), while the correlation was smaller in the second round data, \(r_2 = -0.21, t(104) = -2.16, p = 0.02, \text{one-tailed}\). Subjects who had higher scores on AOT-Beliefs had lower Brier scores than those who scored lower on AOT-Beliefs. We also wanted to know whether a similar relation to the one observed between Brier scores and AOT-Beliefs would also exist between Brier scores and mean AOT-Reasons scores. Specifically, we hypothesized that the greater mean AOT-Reasons score is, the lower the subject’s mean Brier score would be. The observed
correlations were too small and not statistically significant in either round, \( r_1 = -0.05, t(125) = -0.60, p = 0.28, \) one-tailed; \( r_2 = -0.12, t(104) = -1.22, p = 0.11, \) one-tailed. Finally, we looked at how the correlation between Brier scores and mean AOT-Reasons behaved as the question difficulty varied. We observed a moderate but statistically nonsignificant correlation in the first round data, \( r_1 = -0.31, t(18) = -1.37, p = 0.19, \) two-tailed, but the correlation was large in the second round data, \( r_2 = -0.58, t(18) = -3.04, p = 0.01, \) two-tailed. We found partial support for the hypothesis that as the question difficulty increases, the relation between Brier scores and AOT-Reasons strengthens.

**Linear Mixed-Effect Models.** We first regressed the Brier scores of subjects on AOT-Reasons with subjects and items as crossed random effects. We did not observe an effect of the measure on subjects’ Brier scores in either dataset (-0.002, 95% c.i. -0.033 to 0.030, Wald method; -0.014, 95% c.i. -0.050 to 0.022, Wald method). One reason why we did not observe any effect of AOT-Reasons could be that the usefulness of considering alternative reasons might be dependent on subjects’ knowing the correct answer to the question. If a subject knows the correct answer, they might not feel the need to list alternative reasons. To test this hypothesis we first regressed the Brier scores of subjects on AOT-Reasons and Correct (whether subjects guessed the answer correctly where 1 = correct and 0 = incorrect with subjects and items as crossed random effects). AOT-Reasons on its own had no effect on subjects’ Brier scores again in the first round data (-0.009, 95% c.i. -0.030 to 0.012, Wald method), but it did in the second round data (-0.028, 95% c.i. -0.051 to -0.005, Wald method), and as it would be expected Correct did
have an effect on the Brier scores in both rounds (-0.408, 95% c.i. -0.423 to -0.392, Wald method; -0.442, 95% c.i. -0.459 to -0.426, Wald method). We also observed an interaction effect of AOT-Reasons and Correct on the Brier scores in both rounds (0.060, 95% c.i. 0.041 to 0.079, Wald method; 0.103, 95% c.i. 0.082 to 0.124, Wald method).

To get a better understanding of what is going on, we split our data into two based on whether subjects guessed a specific question correctly or incorrectly. For each dataset, we regressed subjects’ Brier scores on AOT-Reasons with subjects and items as crossed random effects. The results showed a good effect when subjects guessed the question incorrectly such that higher AOT-Reasons scores were associated with lower Brier scores (-0.063, 95% c.i. -0.101 to -0.029, Wald method; -0.112, 95% c.i. -0.159 to -0.066, Wald method). However, when subjects answered the question correctly, we observed a bad effect of AOT-Reasons on subjects’ Brier scores such that higher AOT-Reasons scores were associated with higher Brier scores (0.043, 95% c.i. 0.023 to 0.065, Wald method; 0.054, 95% c.i. 0.032 to 0.076, Wald method). Finally, we ran a generalized linear mixed-effect model to determine any effect of AOT-Reasons on Correct but we did not observe any effect of AOT-Reasons on Correct (-0.079, 95% c.i. -0.245 to 0.086, Wald method; -0.057, 95% c.i. -0.271 to 0.157, Wald method). It seems that subjects’ being correct on a given question is not dependent upon subjects’ AOT-Reasons scores. These

31 The lmer function used in R was lmer(BS ~ Ar+(1+Ar|S)+(1+Ar|C),d1[d1$Correct<0,]), where BS is Brier scores, Ar is AOT-Reasons, S is subjects, C is items, and d1[d1$Correct<0,] is the data frame that only contained data for incorrectly answered questions.

32 The lmer function used in R was lmer(BS ~ Ar+(1+Ar|S)+(1+Ar|C),d1[d1$Correct>0,]), where BS is Brier scores, Ar is AOT-Reasons, S is subjects, C is items, and d1[d1$Correct>0,] is the data frame that only contained data for correctly answered questions.
analyses explain the strengthening of the correlation between Brier scores and AOT-Reasons scores as the difficulty of questions increases. Since subjects are more likely to answer difficult questions wrong, writing more alternative reasons as to why their preferred option might be wrong lowers their Brier scores.

Next we regressed the Brier scores of subjects on their AOT-Beliefs scores with subjects and items as crossed random effects. We observed an effect of AOT-Beliefs on the Brier scores in both rounds (-0.104, 95% c.i. -0.162 to -0.047, Wald method; -0.063, 95% c.i. -0.121 to -0.006, Wald method). Higher AOT-Beliefs scores were associated with lower Brier scores.

**Overconfidence and Actively Open-Minded Thinking**

We again used two overconfidence measures in assessing the relations between AOT and overconfidence. The first measure, Over1, compared subjects’ probability judgments for their preferred answers to the percentage of correct answers. The second measure, Over2, compared subjects’ expected Brier scores (EBS, see the results section of Studies 1A and 1B for an explanation of this measure) to their actual Brier scores. In the first round data we observed that, for both measures, the mean overconfidence scores were positive, meaning that subjects were overconfident in their probability judgments ($M_{Over1} = 0.11; M_{Over2} = 0.19$). In the second round data, we once again observed positive overconfidence scores, but these scores were slightly lower than the first round overconfidence scores ($M_{Over1} = 0.08; M_{Over2} = 0.17$).
The correlations between AOT-Reasons and each overconfidence measure were small and statistically non-significant in the first round data (with Over1 as the overconfidence measure: $r_{\text{Over1}, R1} = -0.10$, $t(125) = -1.16$, $p = 0.07$, one-tailed; with Over2 as the overconfidence measure: $r_{\text{Over2}, R1} = -0.13$, $t(125) = -1.47$, $p = 0.12$, one-tailed). In the second round data the correlation between each overconfidence measure and AOT-Reasons was statistically significant (with Over1 as the overconfidence measure: $r_{\text{Over1}, R2} = -0.18$, $t(104) = -1.88$, $p = 0.03$, one-tailed; with Over2 as the overconfidence measure: $r_{\text{Over2}, R2} = -0.21$, $t(104) = -2.22$, $p = 0.01$, one-tailed), showing that higher AOT-Reasons scores were associated with lower overconfidence.

Additionally, we looked at the correlations between AOT-Beliefs and each overconfidence measure. Using Over1 as the overconfidence measure, we observed a negative correlation between overconfidence and AOT-Beliefs scores ($r_1 = -0.28$, $t(125) = -3.30$, $p < 0.001$, one-tailed; $r_2 = -0.21$, $t(104) = -2.16$, $p = 0.02$, one-tailed). Similarly, using Over2 as the overconfidence measure, we observed a negative correlation between overconfidence and AOT-Beliefs scores in both rounds ($r_1 = -0.28$, $t(125) = -3.32$, $p < 0.001$, one-tailed; $r_2 = -0.22$, $t(104) = -2.28$, $p = 0.01$, one-tailed) again, showing that higher scores on AOT was associated with lower overconfidence.

**Linear Mixed-Effect Models.** We regressed the variable Over2 (the difference between mean actual Brier scores and mean expected Brier scores) on AOT-Reasons with subjects and items as crossed random effects. Our analysis showed a significant effect of AOT-Reasons on the overconfidence measure on both rounds of data (-0.039, 95% c.i. 90}
-0.072 to -0.006, Wald method; -0.062, 95% c.i. -0.099 to -0.025, Wald method). Subjects who wrote more reasons that went against their preferred answers were less overconfident about their answers. Similarly, regressing Over2 on subjects’ AOT-Beliefs scores with subjects and items as crossed random effects also showed a significant effect of AOT-Beliefs on the overconfidence measure on both rounds of data (-0.121, 95% c.i. -0.193 to -0.049, Wald method; -0.086, 95% c.i. -0.160 to -0.012, Wald method). Subjects who scored higher on AOT-Beliefs showed lower overconfidence in their answers.

We also regressed subjects’ probability judgments for their preferred answer (Prob from now on) on AOT-Reasons with items and subjects as crossed random effects.\textsuperscript{33} We observed a statistically significant effect of AOT-Reasons on Prob such that higher AOT-Reasons scores were associated with lower probability judgments in both rounds (-0.041, 95% c.i. -0.054 to -0.027, Wald method; -0.057, 95% c.i. -0.073 to -0.040, Wald method). A linear mixed-effect model regressing Prob on AOT-Reasons and Correct with items and subjects as crossed random effects showed an effect of AOT-Reasons (-0.039, 95% c.i. -0.053 to -0.026, Wald method; -0.057, 95% c.i. -0.073 to -0.041, Wald method) and an effect of Correct on Prob in both rounds (0.029, 95% c.i. 0.023 to 0.036, Wald method; 0.024, 95% c.i. 0.017 to 0.032, Wald method).\textsuperscript{34} Scoring high on AOT-Reasons was associated with lower probability judgments for preferred answers, and answering

\textsuperscript{33} The lmer function used in R was lmer(Prob ~ Ar + (1+Ar|C) + (1+Ar|S),d1), where Ar is AOT-Reasons, C is items, S is subjects, and d1 is the data frame used.

\textsuperscript{34} The lmer function in R for this model was lmer(Prob ~ Ar + Correct + (1+Ar|C) + (1+Ar| S),d1), where C is items, S is subjects, and d1 is the data frame used.
questions correctly was associated with higher probability judgments for preferred answers. We were able to replicate the results we observed in previous studies.

**Additional Correlation Tests**

We also looked at the relation between subjects’ mean number of reasons (Nreas.s from now on) they wrote and their Brier scores. We hypothesized that higher number of reasons written would be correlated with lower Brier scores. We did not observe a correlation between these two variables in the first round data \( r = -0.03, t(125) = -0.34, p = 0.37, \) one-tailed), and in the second round data we observed a small but statistically non-significant correlation \( r = -0.14, t(104) = -1.40, p = 0.08, \) one-tailed). We did not observe a correlation between Nreas.s and AOT-Beliefs in the first round data \( r = 0.09, t(129) = 1.08, p = 0.14, \) one-tailed), and in the second round data we observed a small but statistically non-significant correlation, \( r = 0.13, t(108) = 1.32, p = 0.09, \) one-tailed. We were not able to replicate the results we observed in Study 6.

**Comparisons Between the Control and Training Conditions**

We finally investigated whether the actively open-minded thinking training had any effects on subjects’ accuracy scores, overconfidence, and the individual differences in actively open-minded thinking. We expected the subjects in the training condition to have greater change in their Brier scores as compared to those in the control condition, but did not observe any differences between the conditions, \( t = -0.34, p = 0.37, N = 111, \) one-tailed \((M_{\text{control}} = 0.034, M_{\text{training}} = 0.041)\). However, a t-test looking at the differences between mean Brier scores between the two conditions, where we only included data
from questions where subjects were incorrect, showed that compared to the control condition, the training condition improved their Brier scores as a result of training, \( t = -2.03, p = 0.02, N = 111 \), one-tailed \( (M_{\text{control}} = -0.039, M_{\text{training}} = 0.036) \). When we conducted the same two-sample t-test with items where subjects guessed the answer correctly, we did not observe an effect of improvement in Brier scores as a result of training, \( t = 0.42, p = 0.34, N = 111 \), one-tailed \( (M_{\text{control}} = 0.007, M_{\text{training}} = -0.001) \). We additionally looked at the difference between subjects’ accuracy scores based on Correct (see above for how this score is calculated), and observed that subjects in the training condition improved their accuracy more than those in the control condition, \( t = -1.80, p = 0.04, N = 111 \), one-tailed \( (M_{\text{control}} = 0.028, M_{\text{training}} = 0.065) \).

To assess the relation between training and overconfidence we subtracted the second round mean overconfidence measure from the first round mean overconfidence measure, and ran a Welch Two Sample paired t-test to compare the differences between subjects in the control and training conditions. We calculated the changes in mean overconfidence scores in this way for both Over1 and Over2 measures. We expected that the range differences in the control condition should be smaller than those in the training condition. Using Over1 as the overconfidence measure, we observed a significant difference between control and training subjects’ overconfidence score changes such that control condition subjects showed less change in their overconfidence score between the two rounds than the training condition subjects did, \( t = -2.53, p = 0.01, N = 111 \), one-tailed \( (M_{\text{control}} = 0.007, M_{\text{training}} = 0.065) \). Using Over2 as the overconfidence measure, we
observed a smaller difference between control and training subjects’ overconfidence score changes such that control condition subjects showed less change in their overconfidence score between the two rounds than the training condition subjects did, and this difference between the two conditions was not statistically significant, $t = -1.33$, $p = 0.09$, $N = 111$, one-tailed ($M_{control} = 0.013$, $M_{training} = 0.040$). A third analysis used the difference between the maximum and minimum probabilities subjects assigned in each question for both rounds of the task. We took an arithmetic mean of these values, and subtracted the values of the first round from the second round to compare the differences in a Welch Two Sample paired t-test in the two conditions. We hypothesized that the range differences in the control condition should be smaller than those in the training condition, and this hypothesis was supported, $t = -1.68$, $p = 0.05$, $N = 111$, one-tailed ($M_{control} = -0.026$, $M_{training} = 0.009$).

We looked at the differences both in AOT-Beliefs and AOT-Reasons between both rounds of the study. We observed that people in the training condition showed a change in their AOT-Beliefs scores for the better, $t = 2.11$, $p = 0.02$, $N = 111$, one-tailed ($M_{control} = -0.030$, $M_{training} = 0.157$). Next we hypothesized that subjects in the control condition would show a smaller change in the number of reasons they listed that were against their preferred options than those in the training condition. We subtracted subjects’ mean AOT-Reasons scores in the first round from those in the second round, and ran a Welch Two Sample paired test to compare the subjects in the two conditions. We observed that the change in AOT-Reasons was smaller for the subjects in the control condition than those.
in the training condition as expected, $t = -5.06, p < 0.001, N = 111$, one-tailed ($M_{control} = -0.053, M_{training} = 0.204$). Figure 2 also shows that the training condition subjects increased the mean number of reasons they listed against their preferred options as compared to the control condition subjects.
Figure 2. The figure shows each subject’s AOT-Reasons scores in the first and second rounds of Study 7. Red diamonds are training condition subjects, while black diamonds are control condition subjects. The gray diagonal line indicates where the data points would be if there was no change in AOT-Reasons scores. When subjects’ AOT-Reasons scores showed an increase in the second round compared to the first round, they fall above the gray line. When subjects’ AOT-Reasons scores showed a decrease in the second round compared to the first round, they fall below the gray line.
GENERAL DISCUSSION

In this dissertation, we investigated individual differences in AOT by using belief and behavioral measures, and tested the effectiveness of a simple intervention and online AOT training modules in improving subjects’ thinking and accuracy. Table 3 shows the summary results from correlation tests, Table 4 shows the summary results from linear mixed-effect models, and Table 5 shows the summary results from the t-tests we ran.

Belief vs. Behavioral Measures of Actively Open-Minded Thinking

The behavioral measure we tested took subjects’ reasons into consideration since an important component of AOT behavior is one’s tendency to think of arguments or reasons that go against their initially favored beliefs or conclusions. This behavioral measure has an advantage over Baron’s short AOT scale, because it measures subjects’ actual thinking behavior rather than their beliefs about thinking. We expected these two measures to be correlated with one another, but over seven studies, only one study (Study 3) showed a statistically significant correlation between these two measures (see Table 3 for a summary of correlations run in all studies).

Accuracy Scores and Actively Open-Minded Thinking Measures

We used Brier scores to measure accuracy in subjects’ probabilistic judgments in most studies, but used two different measures based on subjects’ point and interval estimates in Studies 3 and 5. Correlations run between AOT measures and Brier scores did not yield consistent and statistically significant results. AOT-Beliefs correlated negatively with Brier scores in all studies but Study 1B, but these correlations were not
always statistically significant. The correlations between AOT-Reasons and Brier scores were not statistically significant with the exception of Study 1B. Study 3 showed significant correlations with both measures when the accuracy measure was based on subjects’ interval estimates but not when the accuracy was based on subjects’ best estimate error.

We were surprised by the lack of a statistically significant correlation between AOT-Reasons and Brier scores, especially because AOT-Reasons scores were almost consistently correlated with the overconfidence measure such that people who scored high on AOT-Reasons had lower overconfidence, while AOT-Reasons had no significant effect on the accuracy score based on subjects’ correct answers except in Study 1B (Table 4). A possible explanation we came up with for the lack of correlation between AOT-Reasons and Brier scores was that the relation between these two variables might be affected by question difficulty. We calculated question difficulty by looking at the average number of subjects who answered a particular question correctly. Correlating this difficulty measure with the correlation coefficients between AOT-Reasons and Brier scores confirmed our hypothesis such that more difficult questions were associated with higher correlations between AOT-Reasons and Brier scores. The linear mixed-effect models we ran provided further support to our hypothesis. When subjects did not answer the questions correctly, higher AOT-Reasons scores were associated with lower, and thus, better Brier scores (see Table 4 for a summary of all the linear mixed-effect models we ran over 7 studies). The beta coefficient for the AOT-Reasons variable was statistically
significant in all studies but the second study, but the coefficient was still in the expected
direction in this study. This relation we observed between AOT-Reasons and Brier scores
disappeared or reversed when subjects answered the questions correctly.

Given these results, it seems that AOT-Reasons and AOT-Beliefs are related
measures of actively open-minded thinking, yet they do not overlap perfectly. AOT-
Reasons measure is most useful when there is greater question difficulty or uncertainty of
outcomes, and can differentiate the individual differences in AOT when people do not or
cannot know the absolute correct answer as in the case of forecasting questions. AOT-
Beliefs, on the other hand, is a more general measure that tells us about how individuals
think they should think.

When we started pilot testing for this dissertation, the 7-item version of the AOT
scale did not include any questions regarding information search. Baron et al. (2015) used
a version of the AOT scale with an eighth item that included a “search for information”
item, which we also included in Studies 1A and 1B. Starting with Study 2, we added
three more “search for information items.” Even though we did not allow our subjects to
look up information regarding the questions, they still had to search their own knowledge
base to be able to come up with reasons. The original 7 items in the AOT scale did not
address the search part, so we felt it was necessary to add these search items in order to
improve task and measure correspondence. Thus, AOT-Beliefs measure, in its current
form, is useful in assessing individual differences in how people think about thinking in
general, regardless of whether they know the correct answer or the outcome. However, as
we mentioned before, it is still a self-report measure, and we cannot exclude social
desirability effects or the discrepancy between actual thinking behavior and reported
thinking tendencies.

**Overconfidence and Actively Open-Minded Thinking Measures**

We measured overconfidence in subjects’ judgments in two ways. The first
measure compared subjects’ probability judgments for their preferred answers to the
percentage of their correct answers. The second measure utilized subjects’ probability
judgments for all possible options, and compared them to subjects’ Brier scores. The
correlations we ran using either overconfidence measure had comparable results. Using
the first measure, we observed AOT-Reasons scores were negatively correlated with
overconfidence in all studies even though these correlations did not reach statistical
significance (Table 3). Similarly, using the second overconfidence measure, we observed
negative correlations in all studies but Study 6. Higher AOT-Reasons scores were
associated with lower overconfidence. Linear mixed-effect models also showed a
statistically significant effect of AOT-Reasons on overconfidence except in Study 2 such
that higher AOT-Reasons scores were associated with lower overconfidence.

Additionally, AOT-Reasons had a statistically significant effect on the probability
judgments subjects made for their preferred answers in all studies. Similar to its effect on
overconfidence, scoring high on AOT-Reasons were associated with lower probability
judgments for the preferred options. It seems that writing reasons against preferred
answers lowers subjects’ unwarranted confidence, which partially explains lower Brier scores, as this accuracy measure utilizes subjects’ probability judgments as well.

The correlations between AOT-Beliefs measure and overconfidence was less consistent than those between AOT-Reasons and overconfidence. Using either overconfidence measure, AOT-Beliefs was correlated negatively with overconfidence in only 4 studies, and in only 2 studies, this correlation was statistically significant (Table 3). Similarly, linear mixed-effect models showed a significant effect of AOT-Beliefs on overconfidence in only two studies (Table 4). Thus, even though it seemed that higher AOT-Beliefs scores were associated with lower overconfidence, this relation was not consistent over 7 studies.

**Interventions for Myside Bias**

In Studies 4 and 5, we required subjects to list at least one reason against their preferred answer in half of the questions (Twoside condition). In Study 4, we observed that the Twoside condition had higher AOT-Reasons than the control condition, wrote more reasons than those in the control condition. We also observed that the Twoside condition had lower, and thus better, Brier scores and showed lower overconfidence than the control condition, but these results were not statistically significant (see Table 5 for a summary of the comparison t-tests run). We similarly observed a statistically significant effect of Twoside condition on AOT-Reasons such that compared to the control condition, the AOT-Reasons scores were higher in Twoside condition (Table 4). These results show
that our experimental manipulation was successful in getting people to consider alternative reasons, which in turn helped them lower their overconfidence.

Furthermore, we observed an effect of the Twoside condition on Brier scores, but this result was not statistically significant. Given that we found a situational usefulness of writing alternative reasons in our task, we hypothesized that our Twoside instructions could also be most effective when subjects did not answer the questions correctly. This hypothesis was confirmed. When subjects answered the questions incorrectly, the Twoside condition had a good effect on subjects’ Brier scores such that the Twoside instructions led to lower Brier scores. This effect was reversed when subjects knew the correct answers. The effect of the Twoside condition on Brier scores became statistically non-significant, and the direction of this effect was in the opposite direction. It seemed that writing opposite reasons when one knew the correct answer, hurt her Brier scores.

This result makes sense if we consider how thinking about alternative reasons lead to better judgments. The Twoside intervention makes other answer options more salient by forcing people to generate reasons that do not support their preferred answers, which in turn, decreases the amount of confidence people might have assigned to their initially favored answers. This mechanism is helpful when subjects do not know the correct answer, as it decreases unwarranted confidence subjects might have. However, it seems that this intervention overcorrects subjects’ confidence when they know the correct answer, and therefore leading to worse Brier scores. If the correct answer is known, the assigned probability to this answer should be 1. Any lower probability judgment will lead
to non-optimal Brier scores. Therefore, if the correct answer is known, there is no need for the subject to seek alternative reasons as to why they might be wrong. Just like the AOT-Reasons measure, this intervention is most useful when used in scenarios of uncertainty and low information conditions.

Despite the success of the Twoside intervention in Study 4, we did not observe any effect of a similar intervention in Study 5, where we used numerical estimation questions rather than categorical probability judgment questions. As we mentioned before, this was probably due to a more general failure of the experimental design (as we also did not observe any of the other relationships outside of the intervention) rather than just specific to the instructions we used in this study.

**Training Adults in Actively-Open Minded Thinking**

We made two attempts at training adults in AOT. Study 6 used a between-subject design while Study 7 used a within-subject design to test the effect of training on people’s AOT-Reasons scores, AOT-Beliefs scores, Brier scores, and overconfidence. Table 5 summarizes the comparison results from these two studies.

In Study 6 we observed that subjects who went through the AOT training had higher AOT-Reasons scores and lower Brier scores than those who did not go through the training, but this difference between the two conditions was not statistically significant. Using either overconfidence measure, we did not observe lower overconfidence in training condition subjects as compared to the control condition subjects either. However,
subjects in the training condition did write significantly more reasons than those in the control condition.

Even though these results showed that training adults in AOT is possible via a short training, it did not lead to any statistically significant changes in either AOT measures, Brier scores, or overconfidence. Therefore, we repeated the study with an improved training module, where we tried to encourage subjects to generate more alternative reasons, put the concepts they learned into practice, and provided more feedback. The results in Study 7 showed that subjects who went through the training module improved their AOT-Reasons and AOT-Beliefs scores significantly. We also found some support for the hypothesis that going through the training would reduce overconfidence from the first round to the second more compared to just doing the same task twice, yet this result was statistically significant only when we used the overconfidence measure that compared subjects’ probability judgments for their preferred options to the percentage of their correct answers. We also observed an improvement in training condition subjects’ Brier scores as compared to the control condition subjects, even though this difference in Brier scores changes was not statistically significant overall. However, similar to the results we observed in the intervention study, subjects seemed to benefit from the training the most when they did not know the correct answer, while there was no effect of the training on Brier scores when subjects knew the correct answer.
In sum, we were able to train our subjects in AOT by using an hour long online training module, and the training had some success in reducing overconfidence, improving Brier scores, and AOT-Beliefs and AOT-Reasons scores.

**Future Directions**

In this dissertation, we devised a behavioral AOT measure based on subjects’ reasons, which had some success in predicting subjects’ Brier scores and overconfidence in probabilistic judgment tasks. One problem with this measure was that it was not as easy to score as we thought it would be. One of us (BG) had to go back and check the classifications of reasons by subjects, and in some cases the reasons were misclassified, so these had to be corrected. This was very time consuming. However, this misclassification problem could be solved by giving subjects a short and simple exercise before they start the task, where they classify reasons and receive feedback on their performance.

One might also argue that the fact that AOT-Reasons predicted Brier scores only when subjects did not answer the questions prevents this measure from being a more general behavioral measure. However, we would argue that this AOT measure is useful where it is most necessary. In real world situations, it is rare that the correct answer or solution to a decision problem can be easily found, so people more often operate in an uncertain world. In these circumstances, AOT-Reasons measure could be useful in assessing the quality of one’s thinking process while making the decision and before the outcome is known.
A related criticism of the study concerns the questions used over 7 studies. We used questions with known answers, and had to instruct subjects not to gather information from outside sources. It was a difficult process to find questions in domains where the majority of subjects had enough knowledge to generate reasons, and it took us multiple tries to find a good mixture of questions. Even then, subjects seemed to struggle with some questions, and especially those that involved submitting point estimates and confidence interval estimates. In future studies, it would be interesting to see how the behavioral measure and the training we devised would predict subjects’ performance in other tasks that involve personal, organizational, or political forecasting type problems. It might also be useful to include another measure such as decision satisfaction to see whether people who consider more alternative reasons are more satisfied with their decisions after the outcome in the long run.

The final version of our training module was successful in changing AOT behavior in subjects, and improving their performance in questions where subjects did not know the correct answer. Given that this module was administered online and on average took one hour to complete, the results are promising. Our training was successful in encouraging subjects to consider alternative reasons, which was its main goal. Additionally, it affected their probability judgments to a lesser extent even though subjects received no specific instructions about making probability judgments during the training. However, we do not know how long the effect of this kind of training would last. It is possible that the effects of this online training would last shorter than the
training sessions Perkins et al. (1986), Baron et al. (1985), or Graumlich and Baron (1991) designed, since these decision making courses took over the course of several weeks. Future studies will need to investigate how lasting the effects of our online AOT training are. However, our training module could be a useful training tool for organizations who want to improve quality and accuracy of decision making of their employees.

**Conclusion**

Our studies investigated the individual differences in AOT by using belief and behavioral measures. The behavioral measure was successful in predicting Brier scores and overconfidence, especially in cases where subjects did not know the correct answer, but it also furthered our understanding of how AOT improves the accuracy of probabilistic judgments. The belief measure was also associated with Brier scores and overconfidence, especially starting with studies where we added three “search” items to the existing AOT scale.

In Chapter 2, we also looked at the effect of a myside bias intervention, which required people to list reasons against their preferred answers. This intervention was most successful in increasing subjects’ Brier scores, when they did not know the correct answer. Finally, we tested two different versions of short AOT training modules. The training increased subjects’ AOT-Beliefs and AOT-Reasons scores, and improved their Brier scores when subjects did not answer the questions correctly. We also observed some effect of training on subjects’ overconfidence.
Table 3. This table summarizes the results from correlations run over 7 studies. “+” indicates statistically significant results in the expected direction; “(+)” indicates statistically non-significant results in the expected direction; “—” indicates statistically significant results in the opposite direction; “(—)” indicated statistically non-significant results in the opposite direction; and blank cells are those where a specific analysis was not run. Study 5 is not included since no relations were observed. Study 7 was a within-subject design study, and we report the results from both rounds. R1 corresponds to the first round of the study, while R2 corresponds to the second round of the study.

<table>
<thead>
<tr>
<th>Correlations</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1A</td>
</tr>
<tr>
<td>AOT-Beliefs</td>
<td></td>
</tr>
<tr>
<td>AOT-Reasons</td>
<td></td>
</tr>
<tr>
<td>Brier scores</td>
<td>(+)</td>
</tr>
<tr>
<td>Best estimate error (Besterror)</td>
<td></td>
</tr>
<tr>
<td>Interval estimate error (Overconf)</td>
<td></td>
</tr>
<tr>
<td>mean Brier scores</td>
<td></td>
</tr>
<tr>
<td>corr. coef. between AOT-Reasons and Brier scores</td>
<td>(+)</td>
</tr>
<tr>
<td>Overconfidence (Over1)</td>
<td></td>
</tr>
<tr>
<td>Overconfidence (Over1)</td>
<td></td>
</tr>
<tr>
<td>Overconfidence (Over2)</td>
<td></td>
</tr>
<tr>
<td>Overconfidence (Over2)</td>
<td></td>
</tr>
</tbody>
</table>

108
Table 4. This table summarizes the results from linear mixed effects models run over 7 studies. “+” indicates statistically significant effects in the expected direction; “(+)” indicates statistically non-significant effects in the expected direction; “—” indicates statistically significant effects in the opposite direction; “(—)” indicates statistically non-significant effects in the opposite direction; and blank cells are those where a specific model was not run. Study 5 is not included since no relations were observed. Study 7 was a within-subject design study, and we report the results from both rounds. R1 corresponds to the first round of the study, while R2 corresponds to the second round of the study. We do not include the random effects variables in this table (see results section of the corresponding study for the full models). The “data” column indicates whether we used the full data from that study or data with only correct or incorrect answers.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Data</th>
<th>Dependent Variable</th>
<th>Independent Variable(s)</th>
<th>1A</th>
<th>1B</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
<th>7 (R1)</th>
<th>7 (R2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>Brier scores</td>
<td>AOT-Reasons</td>
<td>(+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(+)</td>
<td>(+)</td>
<td>(+)</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>Brier scores</td>
<td>Correct</td>
<td>(+)</td>
<td>(+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AOT-Reasons*Correct</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>incorrect answers</td>
<td>Brier scores</td>
<td></td>
<td>AOT-Reasons</td>
<td>(+)</td>
<td>(+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct answers</td>
<td>Brier scores</td>
<td></td>
<td>AOT-Reasons</td>
<td>(—)</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Brier scores</td>
<td></td>
<td>AOT-Beliefs</td>
<td>(+)</td>
<td>(—)</td>
<td>(+)</td>
<td>(+)</td>
<td>(+)</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Brier scores</td>
<td></td>
<td>Twoside</td>
<td>(+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>Dependent Variable</td>
<td>Independent Variable(s)</td>
<td>1A</td>
<td>1B</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>7 (R1)</td>
<td>7 (R2)</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------------</td>
<td>--------------------------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>--------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Brier scores</td>
<td>Correct</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Twoside</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct*Twoside</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>incorrect answers</td>
<td>Brier scores</td>
<td>Twoside</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct answers</td>
<td>Brier scores</td>
<td>Twoside</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Correct</td>
<td>AOT-Reasons</td>
<td>(—)</td>
<td>(+)</td>
<td>(—)</td>
<td>(—)</td>
<td>(—)</td>
<td>(—)</td>
<td>(—)</td>
<td>(—)</td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Best estimate error (Besterror)</td>
<td>AOT-Reasons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Best estimate error (Besterror)</td>
<td>AOT-Beliefs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Interval estimate error (Overconf)</td>
<td>AOT-Reasons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Interval estimate error (Overconf)</td>
<td>AOT-Beliefs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>AOT-Reasons</td>
<td>Twoside</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>No. of reasons for the preferred answer (Arb)</td>
<td>Twoside</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>Dependent Variable</td>
<td>Independent Variable(s)</td>
<td>1A</td>
<td>1B</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7 (R1)</td>
<td>7 (R2)</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------</td>
<td>-------------------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>all</td>
<td>Overconfidence ((\text{Over2}))</td>
<td>AOT-Reasons</td>
<td>+</td>
<td>+</td>
<td>(—)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Overconfidence ((\text{Over2}))</td>
<td>AOT-Beliefs</td>
<td>(—)</td>
<td>(—)</td>
<td>(+)</td>
<td>(+)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Probability assigned to preferred answer ((\text{Prob}))</td>
<td>Twoside</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Probability assigned to preferred answer ((\text{Prob}))</td>
<td>Twoside</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AOT-Reasons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Probability assigned to preferred answer ((\text{Prob}))</td>
<td>No. of reasons for the preferred answer ((\text{Arb}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Probability assigned to preferred answer ((\text{Prob}))</td>
<td>AOT-Reasons</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Probability assigned to preferred answer ((\text{Prob}))</td>
<td>Correct</td>
<td>(+)</td>
<td>(—)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>Probability assigned to preferred answer ((\text{Prob}))</td>
<td>AOT-Reasons</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No. of reasons for the preferred answer ((\text{Arb}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Twoside</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5. This table summarizes the results from t-tests run in Studies 4, 6, and 7. “+” indicates statistically significant effects in the expected direction; “(+)” indicates statistically non-significant effects in the expected direction; “—” indicates statistically significant effects in the opposite direction; “(—)” indicates statistically non-significant effects in the opposite direction; and blank cells are those where a specific model was not run. Study 5 is not included since no statistically significant results were observed.

<table>
<thead>
<tr>
<th>Hypotheses Tested</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects in the experimental condition have higher AOT-Reasons scores (Studies 4 and 6) or improved their AOT-Reasons scores more (Study 7) than those in the control condition.</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Subjects in the experimental condition have higher AOT-Beliefs scores (Study 6) or improved their AOT-Beliefs scores more (Study 7) than those in the control condition.</td>
<td></td>
</tr>
<tr>
<td>When subjects did not know the correct answer, those in the experimental condition improved their Brier scores more than those in the control condition.</td>
<td></td>
</tr>
<tr>
<td>When subjects knew the correct answer, those in the experimental condition improved their Brier scores more than those in the control condition.</td>
<td></td>
</tr>
<tr>
<td>Subjects in the experimental condition have lower Brier scores (Studies 4 and 6) or improved their Brier scores more (Study 7) than those in the control condition.</td>
<td></td>
</tr>
<tr>
<td>Subjects in the experimental condition wrote more reasons than those in the control condition.</td>
<td></td>
</tr>
<tr>
<td>Using Over1 as the overconfidence measure, subjects in the experimental condition were less overconfident (Studies 4 and 6) or their overconfidence diminished from R1 to R2 more (Study 7) than those in the control condition.</td>
<td></td>
</tr>
<tr>
<td>Using Over2 as the overconfidence measure, subjects in the experimental condition were less overconfident (Studies 4 and 6) or their overconfidence diminished from R1 to R2 more (Study 7) than those in the control condition.</td>
<td></td>
</tr>
</tbody>
</table>
Pilot Studies 1A and 1B

The main purpose of this study was to look at the individual differences in actively open-minded thinking, with a focus on people’s ability to be able to come up with arguments against their favored judgments. Additionally we looked at different types of prediction questions so that we could ask subjects questions that would appeal to their knowledge base. We divided all the questions we came up with into two studies, and aimed for half of our recruited subjects to do one set of questions, and the other half a second set of the questions. Therefore, we report the method and results from these studies together.

Method

Subjects in Pilot Study 1A. Thirty-one subjects participated in this study. The subjects’ age ranged from 25 to 78 (Median = 51); 71% were female. The subjects were from a panel of about 1200 people who volunteered to do studies for pay on the Internet over the last 15 years, through advertising, links from various web sites, and word of mouth. They were mostly Americans, varying considerably in age, income, and education level, but with women over-represented. Subjects who did not take previous studies seriously had been removed over the years. The panel was divided into three groups in order to use different samples for closely related studies. Subjects were paid $5 for this study (through PayPal). To be paid, they had to provide full identification associated with a single e-mail address. A pool of about 200 subjects was notified by e-mail when the study was ready, and the study was removed after 24 hours.
**Subjects in Pilot Study 1B.** Forty-two subjects participated in this study. The subjects’ age ranged from 22 to 70 (Median = 48.5); 69% were female. The make-up of the subjects in this study was similar to the one in Pilot Study 1A since the subjects were from the same panel. Subjects were paid $5 for this study, and once the e-mail notification about the availability of the study was sent out, subjects could access it for 24 hours.

**Questions in Pilot Study 1A.** Twenty-five questions with three-choice options were presented to subjects in the same order. We used a fixed order for this study to reduce extraneous variance. The questions were chosen from domains where subjects would be able to list both for and against arguments. The first ten questions asked subjects to guess which U.S. metropolitan area was the largest. The cities were chosen from a list on the Internet (“Cities and metropolitan areas,” n. d.). The next five questions were syllogism questions modified to have three options from Baron et al. (2015), and was part of a list of new Cognitive Reflection Task (CRT) questions. The last ten questions were matrix questions picked from Carter and Russell (2007). Appendix B-2 lists the questions and answers presented to the subjects.

**Questions in Pilot Study 1B.** Subjects answered 22 questions with three choice options A, B, and C in the same order. The first ten questions asked subjects which U.S. metropolitan area was the largest. The cities were chosen from a list on the Internet (“Cities and metropolitan areas,” n. d.). The next four questions were syllogism questions modified to have three choice options from Baron et al. (2015) and were part of a list of
new CRT questions. The next three questions were also CRT questions chosen from the
same paper. The last five questions were word analogy questions picked from an SAT/

**Procedure.** Subjects did the task on their own computers over the Internet. They
first read the instructions regarding the task (see Appendix C-2). The instructions and the
full study can be found on http://finzi.psych.upenn.edu/~baron/ex/bg/args/args1a.html
(Pilot Study 1A) and  http://finzi.psych.upenn.edu/~baron/ex/bg/args/args1b.html (Pilot
Study 1B).

The subjects were told that they would be asked a number of difficult questions,
each with three possible answers A, B, and C. They were informed that the experimenter
was interested in how they thought about answers. Once they read the instructions, they
could enter their age, gender, and e-mail address, and proceed with the study. Subjects
answered one question on each page. On each page, they saw a short note telling them
more about what the question is asking and were given definitions of terms where
necessary (see Appendix C-2). They were also given some example reasons. Below the
short instruction the subjects saw the question with three options, and underneath that
they were given ten note spaces to list their reasons for and against their preferred answer.
They were not required to use all ten spaces but they had to write at least one reason.
After listing reasons, they were asked to indicate their preferred answer, and then they
were asked to state the probability that their preferred answer was correct. The subjects
were given choices for probabilities, and were instructed to choose the one that was
closest to what they think. The options for probability judgments were 0, 1, 5, 10, 20, 30, 40, 50, 60, 70, 90, 95, 99, and 100 (in percent). After completing the page, they could move onto the next page by clicking a button where they could classify their reasons as being for or against the choice options. Once they completed the classification task, they moved onto the next page, again by clicking a button, where they were presented with a new question and they repeated the same procedure for the remaining questions.

After answering all questions, the subjects were presented with an Actively Open-Minded (AOT) scale (see Table 1). In this survey, subjects read eight statements about beliefs about thinking, and were asked to indicate their agreement with each statement on a 5-point scale (1 = completely agree, 5 = completely disagree). Once they completed the survey, they clicked a button to submit their responses.

Results

Very few subjects wrote more than one reason, which did not give us enough data to investigate individual differences in actively open-minded thinking. Looking at the average number of reasons written by subject for each question showed that the average number of reasons written for each question was smaller than 2. Subjects also reported that the questions were too difficult, which we confirmed by calculating their accuracy scores. Here we report some results for both studies as these studies were the same except for the questions asked.

We first calculated actively open-minded thinking scale scores for each subject based on the eight-item AOT scale subjects took at the end of the survey. We first subtracted 3 from each score so that subjects’ responses now ranged from -2 to +2 for
each item. Items 4 through 7 were reverse-scored items, so we multiplied subjects’
responses for these items with -1. Finally, we took the arithmetic mean of the eight items
to come up with a singular AOT Scale score for each subject. The subjects’ AOT scale
scores ranged from -0.625 to 1.5 in Pilot Study 1A ($M_{P1A} = 0.609$), and from -0.5 to 1.5 in
Pilot Study 1B ($M_{P1B} = 0.542$). We hypothesized that if someone scores high on the AOT
scale, they should have better calibration. A Pearson’s product-moment correlation test
between AOT scale scores and calibration scores for each subject confirmed our
hypothesis ($r = -0.42, t(29) = -2.5, p = 0.02$ in Pilot Study 1A; $r = -0.33, t(40) = -2.2, p =
0.03$ in Pilot Study 1B).

We also thought that people who score high on the AOT scale would also be more
accurate. We used Brier scores to assess the accuracy of the subjects. To test the
correlation between actively open-minded thinking and accuracy, we ran a Pearson’s
product-moment correlation between AOT-Beliefs and average Brier scores for each
subject. We did not find any support for our hypothesis ($r = 0.03, t(29) = 0.19, p = 0.85$ in
Pilot Study 1A; $r = 0.08, t(40) = 0.50, p = 0.62$ in Pilot Study 1B).

Next we calculated a new measure by looking at the reasons subjects listed for
each question, which we call AOT-Reasons. According to this measure, subjects gained 1
point for giving a reason that is for an option other than their preferred option or a reason
that is against their preferred option. They did not get any points for giving a reason that
supports their preferred option. We expected this measure to be positively correlated with
AOT-Beliefs. A Pearson’s product moment correlation showed a negative correlation
between these two measures ($r = -0.44, t(29) = -2.64, p = 0.01$, two-tailed, in Pilot Study
1A; \( r = -0.25, t(40) = -1.64, p = 0.11 \), two-tailed, in Pilot Study 1B). These unexpected results might be due to the fact that subjects were not able to write many reasons for questions since they found them to be difficult.

Discussion

In these two studies we looked at individual differences in actively open-minded thinking as well as what kinds of questions might be suitable for studies to test the validity of a new actively open-minded measure based on subjects’ reasoning. Because we did not require subjects to list more than one reason, many subjects did not list more than one reason per question. Additionally, the majority of the reasons listed by subjects were those which explained why their preferred option was the correct answer. This made it difficult for us to assess individual differences in actively open-minded thinking by either the attitudinal or the behavioral measures. We did not observe any correlations between these two measures, and correlational tests run between these measures and accuracy scores did not yield any meaningful results either.

The task difficulty probably exacerbated subjects’ poor performance in this study. Because we told our subjects not to look up information on the Web, even if our subjects were able to utilize actively open-minding thinking principles, they had to rely on their own knowledge base, which most of the time was lacking for the questions they had to answer. In Pilot Study 1A, for only 2 questions out of 25 they had to answer, the average percentage of giving a correct answer was above 75%; and only for 10 questions at least half of the subjects were able to answer the questions correctly, and similarly, in Pilot
Study 1B, for only 2 questions out of 20 they had to answer, the average percentage of giving a correct answer was above 75%; and only for 12 questions at least half of the subjects were able to answer the questions correctly (see Appendix B-2 for the percentage of correct answers for each question). Therefore, we decided to repeat these studies with some modifications and changes to the questions.

**Pilot Study 2**

The main purpose of this study was to systematically choose questions to be used in Studies 2 and 3.

**Method**

**Subjects.** Eighty-eight subjects participated in the study. The subjects’ age ranged from 22 to 78 (Median = 49.5); 73.8% were female. The make-up of the subjects in this study were the same as those in previous studies. Subjects were paid $6 for their participation (through PayPal). A panel of about 200 subjects was notified by email when the study was ready, and the study was removed when there were about 85 subjects, and thus aiming for 90.

**Questions.** There were 114 questions with three answer options A, B, and C. In 20 of the questions, subjects were asked to pick the most populous city. A second set of 20 questions asked subjects to pick the most populous country, while a third set of 20 questions asked them to pick the most populous U.S. state. A fourth set of 20 questions asked subjects to pick the highest ranking university. For the university ranking questions, the 2014 Academic Ranking of the World Universities was used. This ranking is also know as Shanghai Ranking, and is published by Shanghai Ranking Consultancy. A
fifth set of 20 questions asked subjects to pick the movie that had the most recent release date. The last 14 questions were items chosen from a list of new CRT questionnaire, which consisted of logic and arithmetic questions (Baron et al., 2015). A full list of questions and answers can be found in Appendix B-2.

**Procedure.** Subjects did the task on their own computers over the Internet. They first read the instructions regarding the task (see Appendix C-2). The instructions and the full study can be found on http://finzi.psych.upenn.edu/~baron/ex/bg/args/args3.html.

Each subject answered 60 random questions picked out of 114 questions in randomized order (see Appendix B-2 for a list of all questions used in this study). On each page the subjects saw one question with three answer options A, B, and C. They were first asked to pick their preferred answer, and then were asked to state the probability that each answer could be correct. The subjects were given choices for probabilities, and were instructed to choose the one that was closest to what they thought. The options for probability judgments were 0, 1, 5, 10, 20, 30, 40, 50, 60, 70, 90, 95, 99, and 100 (in percent).

**Results**

To pick the questions to be used in later studies, we ran a biserial correlation between the variable Correct (where 1 = subject answered the question correctly, 0 = subject answered the question incorrectly) and the subject means of variable Correct to assess discrimination and fitted a Rasch model on Correct to assess difficulty of the panel of questions. Appendix B-2 shows the discrimination and difficulty coefficients for each
chosen question. We tried to pick questions that were on the more difficult side of the
difficulty scale as well as those that had better discrimination.

Pilot Study 3

The main purpose of this study was to test the possibility of using numerical
estimation questions in Studies 3 and 5.

Method

Subjects. Ninety-eight subjects participated in the study. The subjects’ age ranged
from 19 to 78 (Median = 50.5); 72.4% were female. The make-up of the subjects in this
study were the same as those in previous studies. Subjects were paid $4 for their
participation (through PayPal). A panel of about 200 subjects was notified by email when
the study was ready, and the study was removed when we had about 100 subjects.

Questions. There were 40 questions that asked for numerical answers. A full list
of questions can be found in Appendix B-2. The questions were made up based on the
data available on government websites, national newspapers, and research centers.

Procedure. Subjects did the task on their own computers over the Internet. They
first read the instructions regarding the task (see Appendix C-2). The instructions and the
full study can be found on http://finzi.psych.upenn.edu/~baron/ex/bg/args/args6.html.

Each subjects answered 40 questions in randomized order. Subjects saw one
question on each page (e.g., “What percentage of U.S. citizens do not have health
insurance?”) and were asked to submit a point estimate. After entering their best estimate
they were told to give two more numbers, a lower number, below their best judgments,
and an upper number above it. They were also instructed to choose these two numbers so that there would be a 90% chance that the right answer is in the interval between them. After entering a lower and upper number, they were reminded that the two numbers they put in meant that they thought there was a 90% chance that the right answer was between their upper and lower numbers, and if they did not think this was the case, they were advised to adjust their numbers. After they answered all the questions, they were given an AOT scale (see Table 2).

**Results**

**Selecting Questions for Study 3**

Since subjects were asked to make point estimates and give us confidence intervals for their point estimates, we had to come up with a scoring rule that would accommodate our data. Jose and Winkler (2009) discusses a linear scoring rule for evaluating quantile estimates that would be preferable over the more commonly used scoring rules, but their scoring rule is designed as an incentive to get experts to give the most accurate quantile estimates. However, our aim in this study was not to incentivize people to give us more accurate estimates, but score subjects’ performance to pick questions for the next two studies. Therefore, we used a scoring rule that fit this purpose. The scoring rule would also need to be viable for percentage questions. Finally, we wanted a scoring rule that would be able to assess both under- and overconfidence together such that we would be able to use this scoring rule while analyzing the data from later studies. Therefore, we came up with a scoring rule, which we call CIscore. We next describe how we thought about this scoring rule and how to calculate it.
We thought that we could compare the given alpha (0.05) with what the alpha would be if the true value were at one end of the confidence interval, which we will call the inferred alpha. In particular, we wanted to use a logistic function instead of a normal distribution to get the inferred alpha. The reason for using a logistic function is that this function is linear in log odds, and thus makes the calculations simpler.

\[ y(x) = \frac{1}{1 + e^{-x}} \]

Let us suppose that we look up alpha levels in a logistic function instead of a normal one, such that \( x \) is the numerical continuum, with 0 in the middle. The 90% confidence interval would be between where \( y(x)_l = 0.05 \) and \( y(x)_h = 0.95 \). The confidence interval corresponding to \( X_0 \), the true value or the correct answer, would be the value of \( y \) at \( x = X_0 \). If we transform the probability scale, the vertical axis in the graph of the logistic function to log odds, \( \log(y/(1-y)) \), we find that the new function is a linear function of \( x \). Therefore, the log odds of the deviation of \( X_0 \) from \( X_h \), upper boundary of the confidence interval (or \( X_l \), lower boundary of the confidence interval) is a linear function of the difference between \( X_0 \) and \( X_h \) (or \( X_l \)). Thus, we measure accuracy of the interval as,

\[
\frac{(X_h - \text{true value})}{(X_h - \text{best estimate})} \text{ if true value} \geq \text{best estimate},
\]

\[
\frac{(\text{true value} - X_l)}{(X_l - \text{best estimate})} \text{ if true value} \leq \text{best estimate}.
\]

If a subject sets their confidence intervals correctly, the mean of these differences should be 0.
To pick the questions we looked at the means of CI\(\text{score}\) for each question, and excluded the questions where the subjects’ CI\(\text{scores}\) seemed out of proportion. This left us with 25 questions to be used for Study 3 (see Appendix B).

**Selecting Questions for Study 5**

For selecting questions to be used in Study 5, we used absolute log error to measure the accuracy of subjects’ best estimates, which we call Besterror, and an alternative accuracy score to measure their confidence intervals, which we call Overconf. Below we describe the reasoning and calculation of these two accuracy scores.

As mentioned above, Besterror was the absolute log error of subject’s best estimate (best), which we calculated as follows:

\[
\log\left(\frac{\text{best}}{X_0}\right), \text{ where } X_0 \text{ is the correct answer.}
\]

We used a slightly different accuracy score formula for percentage questions. We thought that the best way to measure the accuracy of the central tendency on a percentage question was to use log odds, similar to the CI\(\text{score}\) explained above. However, using log odds might be problematic in some cases on percentage questions. Let us consider a subject with 100\% as best estimate for a question. If we were to calculate the log odds for this estimate, we would get infinity. Therefore, we introduced a small correction suggested by Ariely et al. (2000) and Baron et al. (2014) and used extensively in literature to fix this problem. Specifically we rescaled the odds ratio before taking log odds so that it was between 0.01 and 0.99 instead of 0 to 1.00, such that

\[
\log\left(\frac{\frac{0.98 \cdot \text{best}}{100 + 0.01}}{\frac{0.99 - 0.98 \cdot \text{best}}{100}}\right)
\]
This approach is advantageous because log odds considers the distance from both 0 and 1.00, so it does not necessitate an arbitrary decision about which end of the scale to use. Additionally, this approach gives full weight to errors near the ends of the scale.

Our new accuracy score, Overconf, which used subjects’ confidence interval judgments, looked at how serious subjects’ errors would be given where their best estimate and the boundaries of their confidence interval judgments fell. Let us consider the case where the true value is higher than subject’s best estimate. Next, consider where the upper confidence interval bound (X_h) falls. If the subject’s best estimate is below the true value, then the error is not very serious. If it is just above the best estimate, then that is a very serious error. We can, then, compute the ratio \((X_0-X_h)/(X_0-\text{best})\), and raise this ratio to 0 if it is less than 0 so that we eliminate the cases in which the true value is within the confidence interval. Then, we can do the analogous calculation for when the true value is less than subject’s best estimate, \((X_l-X_0)/(\text{best}-X_0)\). This way all the scores are between 0 and 1. The idea with this score is that AOT should reduce overconfidence, and not necessarily affect underconfidence, so Overconf does not assess underconfidence. However, one can develop an alternative measure of underconfidence along the same lines very easily.

To pick the questions for Study 5, we first looked at the Cronbach’s alpha values of Overconf and Besterror to see if there were any items that were reverse-correlated. This resulted in the exclusion of six items. Next, we looked at the means of item means of Overconf and Besterror, and picked questions that had the lowest means in both scoring methods. This left us with 20 questions. Finally, we modified the wording of
some of these questions to make them clearer (see Appendix B-2 for a list of questions used in Study 5).
APPENDIX B-1: QUESTIONS AND ANSWERS USED IN REPORTED STUDIES

Study 1A

**CITY QUESTIONS:** Which is largest?

<table>
<thead>
<tr>
<th></th>
<th>A. City Name 1</th>
<th>B. City Name 2</th>
<th>C. City Name 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A. Chicago, IL</td>
<td>B. Washington, D. C.</td>
<td>C. Seattle, WA</td>
</tr>
<tr>
<td>2</td>
<td>A. Philadelphia, PA</td>
<td>B. Boston, MA</td>
<td>C. Baltimore, MD</td>
</tr>
<tr>
<td>3</td>
<td>A. Phoenix, AZ</td>
<td>B. Atlanta, GA</td>
<td>C. Louisville, KY</td>
</tr>
<tr>
<td>4</td>
<td>A. San Francisco, CA</td>
<td>B. Dallas, TX</td>
<td>C. Portland, OR</td>
</tr>
<tr>
<td>5</td>
<td>A. Baltimore, MD</td>
<td>B. Columbus, OH</td>
<td>C. Las Vegas, NV</td>
</tr>
<tr>
<td>6</td>
<td>A. Las Vegas, NV</td>
<td>B. Milwaukee, WI</td>
<td>C. St. Louis, MO</td>
</tr>
<tr>
<td>7</td>
<td>A. Indianapolis, IN</td>
<td>B. Cincinnati, OH</td>
<td>C. Tampa, FL</td>
</tr>
<tr>
<td>8</td>
<td>A. Nashville, TN</td>
<td>B. Orlando, FL</td>
<td>C. Milwaukee, WI</td>
</tr>
<tr>
<td>9</td>
<td>A. Austin, TX</td>
<td>B. Dallas, TX</td>
<td>C. Atlanta, GA</td>
</tr>
<tr>
<td>10</td>
<td>A. San Antonio, TX</td>
<td>B. Hartford, CT</td>
<td>C. Austin, TX</td>
</tr>
<tr>
<td>11</td>
<td>A. Albany, NY</td>
<td>B. Cincinnati, OH</td>
<td>C. Salt Lake City, UT</td>
</tr>
<tr>
<td>12</td>
<td>A. Denver, CO</td>
<td>B. Reno, NV</td>
<td>C. Raleigh, NC</td>
</tr>
<tr>
<td>13</td>
<td>A. Detroit, MI</td>
<td>B. Houston, TX</td>
<td>C. Cleveland, OH</td>
</tr>
<tr>
<td>14</td>
<td>A. Minneapolis, MN</td>
<td>B. Denver, CO</td>
<td>C. Phoenix, AZ</td>
</tr>
<tr>
<td>15</td>
<td>A. Seattle, WA</td>
<td>B. San Diego, CA</td>
<td>C. Baltimore, MD</td>
</tr>
<tr>
<td>16</td>
<td>A. Pittsburgh, PA</td>
<td>B. Portland, OR</td>
<td>C. Sacramento, CA</td>
</tr>
<tr>
<td>17</td>
<td>A. Honolulu, HI</td>
<td>B. Anchorage, AK</td>
<td>C. Little Rock, AR</td>
</tr>
<tr>
<td>18</td>
<td>A. Syracuse, NY</td>
<td>B. Tucson, AZ</td>
<td>C. Knoxville, TN</td>
</tr>
<tr>
<td>19</td>
<td>A. Boston, MA</td>
<td>B. Washington, D. C.</td>
<td>C. Miami, FL</td>
</tr>
<tr>
<td>20</td>
<td>A. Philadelphia, PA</td>
<td>B. Baltimore, MD</td>
<td>C. Minneapolis, MN</td>
</tr>
</tbody>
</table>

**ANSWERS:**

**Study 1B**

**Calorie Content Questions:** Most calories per serving?

<table>
<thead>
<tr>
<th></th>
<th>A. Bread, whole wheat or multi-grain</th>
<th>B. Almonds</th>
<th>C. Sweet corn, cooked</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A. Lentils, cooked</td>
<td>B. Brown rice, cooked</td>
<td>C. Spaghetti, whole-wheat, cooked</td>
</tr>
<tr>
<td>2</td>
<td>A. Turnip greens, boiled</td>
<td>B. Green peas, cooked</td>
<td>C. Artichoke, cooked</td>
</tr>
<tr>
<td>3</td>
<td>A. Carrot, raw</td>
<td>B. Potato, with skin, cooked</td>
<td>C. Bread, rye</td>
</tr>
<tr>
<td>4</td>
<td>A. Banana</td>
<td>B. Oat bran muffin</td>
<td>C. Strawberries</td>
</tr>
</tbody>
</table>

**Fiber Content Questions:** Most fiber per serving?

<table>
<thead>
<tr>
<th></th>
<th>A. Cheese</th>
<th>B. Pasta</th>
<th>C. Salad dressing</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>A. Bacon</td>
<td>B. Beef</td>
<td>C. Sausage, pork</td>
</tr>
<tr>
<td>7</td>
<td>A. Dates</td>
<td>B. Potato salad</td>
<td>C. Brown rice</td>
</tr>
<tr>
<td>8</td>
<td>A. Hard boiled eggs</td>
<td>B. Oats</td>
<td>C. Orange juice</td>
</tr>
<tr>
<td>9</td>
<td>A. Avocado</td>
<td>B. Peanut butter</td>
<td>C. Milk chocolate</td>
</tr>
</tbody>
</table>

**Fat Content Questions:** Most fat per serving?

<table>
<thead>
<tr>
<th></th>
<th>A. Egg, fried</th>
<th>B. Cod, fried in batter</th>
<th>C. Olives, in brine</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>A. Grilled lamb chops</td>
<td>B. Cream cheese</td>
<td>C. Duck roast</td>
</tr>
<tr>
<td>12</td>
<td>A. Pancake</td>
<td>B. Cheesecake</td>
<td>C. Tomato and cheese pizza</td>
</tr>
<tr>
<td>13</td>
<td>A. Canned salmon</td>
<td>B. Canned tuna, in oil</td>
<td>C. Canned sardines</td>
</tr>
<tr>
<td>14</td>
<td>A. Roasted and salted peanuts</td>
<td>B. Chestnuts</td>
<td>C. Walnuts</td>
</tr>
</tbody>
</table>

**Protein Content Questions:** Most protein per serving?

<table>
<thead>
<tr>
<th></th>
<th>A. Lean ground beef</th>
<th>B. Chicken breast</th>
<th>C. Ground turkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>A. Lamb</td>
<td>B. Steak, lean cut</td>
<td>C. Tofu</td>
</tr>
<tr>
<td>17</td>
<td>A. Greek yoghurt</td>
<td>B. Cottage cheese</td>
<td>C. Swiss cheese</td>
</tr>
<tr>
<td>18</td>
<td>A. Hummus</td>
<td>B. Black beans</td>
<td>C. Edamame beans</td>
</tr>
<tr>
<td>19</td>
<td>A. Shrimp</td>
<td>B. Tuna</td>
<td>C. Cod</td>
</tr>
</tbody>
</table>

**Answers:**

### Study 2

**CITY POPULATION QUESTIONS:** Which is the most populous city proper? A city proper is an area contained within city limits. A city proper may not include suburbs.

<table>
<thead>
<tr>
<th></th>
<th>A. London, United Kingdom</th>
<th>B. Moscow, Russia</th>
<th>C. New York City, U.S.A</th>
<th>DIFFICULTY SCORE</th>
<th>DISCRIMINATION SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A. Moscow, Russia</td>
<td></td>
<td>C. New York City, U.S.A</td>
<td>3.35</td>
<td>0.53</td>
</tr>
<tr>
<td>2</td>
<td>A. Madrid, Spain</td>
<td>B. Baghdad, Iraq</td>
<td>C. Paris, France</td>
<td>2.92</td>
<td>0.52</td>
</tr>
<tr>
<td>3</td>
<td>A. London, United Kingdom</td>
<td>B. Los Angeles, U.S.A</td>
<td>C. Seoul, South Korea</td>
<td>0.3</td>
<td>0.48</td>
</tr>
</tbody>
</table>

**COUNTRY POPULATION QUESTIONS:** Which is the most populous country?

<table>
<thead>
<tr>
<th></th>
<th>A. Poland</th>
<th>B. Israel</th>
<th>C. Canada</th>
<th>DIFFICULTY SCORE</th>
<th>DISCRIMINATION SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>A. Poland</td>
<td>B. Israel</td>
<td>C. Canada</td>
<td>2.93</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>A. France</td>
<td>B. Egypt</td>
<td>C. Australia</td>
<td>1.45</td>
<td>0.55</td>
</tr>
<tr>
<td>6</td>
<td>A. Germany</td>
<td>B. Brazil</td>
<td>C. Poland</td>
<td>-1.01</td>
<td>0.52</td>
</tr>
<tr>
<td>7</td>
<td>A. Canada</td>
<td>B. Japan</td>
<td>C. The United States of America</td>
<td>-0.58</td>
<td>0.48</td>
</tr>
<tr>
<td>8</td>
<td>A. Turkey</td>
<td>B. United Kingdom</td>
<td>C. Ukraine</td>
<td>1.34</td>
<td>0.46</td>
</tr>
<tr>
<td>9</td>
<td>A. Australia</td>
<td>B. Turkey</td>
<td>C. Canada</td>
<td>0.95</td>
<td>0.49</td>
</tr>
</tbody>
</table>

**UNIVERSITY RANKING QUESTIONS:** Which is the highest ranking university in 2014? The ranking is entirely academic and research oriented. It does not include athletics.

<table>
<thead>
<tr>
<th></th>
<th>A. Princeton University</th>
<th>B. Oxford University (UK)</th>
<th>C. Massachusetts Institute of Technology</th>
<th>DIFFICULTY SCORE</th>
<th>DISCRIMINATION SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>A. Princeton University</td>
<td>B. Oxford University (UK)</td>
<td>C. Massachusetts Institute of Technology</td>
<td>2.76</td>
<td>0.69</td>
</tr>
<tr>
<td>11</td>
<td>A. Duke University</td>
<td>B. Cornell University</td>
<td>C. New York University</td>
<td>0</td>
<td>0.48</td>
</tr>
</tbody>
</table>
**Movie Release Date Questions:** Which is the movie that has the most recent release date?

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>A. V for Vendetta</td>
<td>B. The Matrix</td>
<td>C. Se7en</td>
<td>0.31</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>A. Twelve Monkeys</td>
<td>B. Annie Hall</td>
<td>C. Stand By Me</td>
<td>-0.63</td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>A. The Wizard of Oz</td>
<td>B. Strangers on a Train</td>
<td>C. A Clockwork Orange</td>
<td>-0.67</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>A. The Usual Suspects</td>
<td>B. Alien</td>
<td>C. Some Like It Hot</td>
<td>-1.15</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Logic Questions:** What can we conclude from these two statements?

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>In a box, no blue things are triangular, and no triangular things are large.</td>
<td>A. No blue things are large. blue things are not large. conclude anything about blue things and large things.</td>
<td>B. Some</td>
<td>C. We can’t</td>
<td>-0.51</td>
<td>0.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>In a box, some red things are square, and some square things are large.</td>
<td>A. Some red things are large. things are large. conclude anything about red things and large things.</td>
<td>B. All red</td>
<td>C. We can’t</td>
<td>-1.16</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**State Population Questions:** Which is the most populous state?

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>A. West Virginia</td>
<td>B. Rhode Island</td>
<td>C. Oregon</td>
<td>0.68</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>A. Massachusetts</td>
<td>B. Maryland</td>
<td>C. Pennsylvania</td>
<td>0.29</td>
<td>0.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Arithmetic Questions:**

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>Soup and salad cost $5.50 in total. The soup costs a dollar more than the salad. How much does the salad cost?</td>
<td>A. $4.50</td>
<td>B. $3.50</td>
<td>C. 2.31</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Answers:**

### Study 3

<table>
<thead>
<tr>
<th><strong>Question</strong></th>
<th><strong>Answer</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>What percentage of American employees can receive paid sick leave?</td>
<td>61</td>
</tr>
<tr>
<td>What is the average annual household income of Broadway theatergoers? (in U.S. dollars)</td>
<td>201500</td>
</tr>
<tr>
<td>How many construction workers in India are women? (in millions)</td>
<td>8</td>
</tr>
<tr>
<td>How many tanning salons are there in the U.S.?</td>
<td>14000</td>
</tr>
<tr>
<td>How many automobiles were sold in the U.S. in 2014? (in millions)</td>
<td>16.5</td>
</tr>
<tr>
<td>How many inches of snow did Boston, MA, get this past winter season (2014-2015)?</td>
<td>108.6</td>
</tr>
<tr>
<td>What percentage of U.S. adults are vegetarians?</td>
<td>5</td>
</tr>
<tr>
<td>What's the average cost for a middle-income family to raise a child born in 2013 until age 18? (in U.S. dollars)</td>
<td>245340</td>
</tr>
<tr>
<td>How much money did the movies released in 2014 make? (in billion U.S. dollars)</td>
<td>9.69</td>
</tr>
<tr>
<td>What percentage of U.S. citizens do not have health insurance?</td>
<td>12.9</td>
</tr>
<tr>
<td>How many original TV shows aired in the U.S. in 2014?</td>
<td>328</td>
</tr>
<tr>
<td>How many pounds of Haas avocados (the avocados one normally purchases at the U.S. grocery stores) were sold in the U.S. in 2014? (in billions)</td>
<td>1.9</td>
</tr>
<tr>
<td>According to the 2013 American community Survey, how many U.S. residents speak a language other than English at home? (in millions)</td>
<td>61.8</td>
</tr>
<tr>
<td>How much does college education cost per year on average in all institutions in the U.S.? (in U.S. dollars)</td>
<td>33047</td>
</tr>
<tr>
<td>What is the average math SAT score in the U.S., out of 800?</td>
<td>513</td>
</tr>
<tr>
<td>QUESTION</td>
<td>ANSWER</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>16 What is the average one-way commute time in the U.S. in minutes?</td>
<td>25.4</td>
</tr>
<tr>
<td>17 In what percentage of U.S. households does someone play video games?</td>
<td>67</td>
</tr>
<tr>
<td>18 What's the average age of someone who plays video games with some regularity?</td>
<td>30</td>
</tr>
<tr>
<td>19 What percentage of women in the U.S. play video games?</td>
<td>47</td>
</tr>
<tr>
<td>20 What was the median gross rent between 2009-2013 in New York City, NY? (in U.S. dollars)</td>
<td>1200</td>
</tr>
<tr>
<td>21 What's the U.S. per capita red meat consumption for 2014 in pounds?</td>
<td>101.4</td>
</tr>
<tr>
<td>22 What's the average age of a white female Broadway theatergoer?</td>
<td>44</td>
</tr>
<tr>
<td>23 What's the average human life expectancy (in years) at birth in the world according to 2010 data?</td>
<td>67.2</td>
</tr>
<tr>
<td>24 How many books do American adults (ages 18+) on average read per year?</td>
<td>12</td>
</tr>
<tr>
<td>25 What is the direct distance between San Francisco and Tokyo, in miles?</td>
<td>5140</td>
</tr>
</tbody>
</table>
### Study 4

**City Population Questions:** Which is the most populous city proper? A city proper is an area contained within city limits. A city proper may not include suburbs.

<table>
<thead>
<tr>
<th></th>
<th>A. London, United Kingdom</th>
<th>B. Moscow, Russia</th>
<th>C. New York City, The United States of America</th>
<th>Difficulty Score</th>
<th>Discrimination Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A. London, United Kingdom</td>
<td>B. Moscow, Russia</td>
<td>C. New York City, The United States of America</td>
<td>3.35</td>
<td>0.53</td>
</tr>
<tr>
<td>2</td>
<td>A. Madrid, Spain</td>
<td>B. Baghdad, Iraq</td>
<td>C. Paris, France</td>
<td>2.92</td>
<td>0.52</td>
</tr>
<tr>
<td>3</td>
<td>A. London, United Kingdom</td>
<td>B. Los Angeles, The United States of America</td>
<td>C. Seoul, South Korea</td>
<td>0.3</td>
<td>0.48</td>
</tr>
</tbody>
</table>

**Country Population Questions:** Which is the most populous country?

<table>
<thead>
<tr>
<th></th>
<th>A. Poland</th>
<th>B. Israel</th>
<th>C. Canada</th>
<th>Difficulty Score</th>
<th>Discrimination Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>A. Poland</td>
<td>B. Israel</td>
<td>C. Canada</td>
<td>2.93</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>A. France</td>
<td>B. Egypt</td>
<td>C. Australia</td>
<td>1.45</td>
<td>0.55</td>
</tr>
<tr>
<td>6</td>
<td>A. Germany</td>
<td>B. Brazil</td>
<td>C. Poland</td>
<td>-1.01</td>
<td>0.52</td>
</tr>
<tr>
<td>7</td>
<td>A. Canada</td>
<td>B. Japan</td>
<td>C. The United States of America</td>
<td>-0.58</td>
<td>0.48</td>
</tr>
<tr>
<td>8</td>
<td>A. Turkey</td>
<td>B. United Kingdom</td>
<td>C. Ukraine</td>
<td>1.34</td>
<td>0.46</td>
</tr>
<tr>
<td>9</td>
<td>A. Belgium</td>
<td>B. Poland</td>
<td>C. Ireland</td>
<td>-0.28</td>
<td>0.46</td>
</tr>
<tr>
<td>10</td>
<td>A. Australia</td>
<td>B. Turkey</td>
<td>C. Canada</td>
<td>0.95</td>
<td>0.49</td>
</tr>
</tbody>
</table>

**University Ranking Questions:** Which is the highest ranking university in 2014? The ranking is entirely academic and research oriented. It does not include athletics.

<table>
<thead>
<tr>
<th></th>
<th>A. Princeton University</th>
<th>B. Oxford University (UK)</th>
<th>C. Massachusetts Institute of Technology</th>
<th>Difficulty Score</th>
<th>Discrimination Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>A. Princeton University</td>
<td>B. Oxford University (UK)</td>
<td>C. Massachusetts Institute of Technology</td>
<td>2.76</td>
<td>0.69</td>
</tr>
<tr>
<td>12</td>
<td>A. Duke University</td>
<td>B. Cornell University</td>
<td>C. New York University</td>
<td>0.00</td>
<td>0.48</td>
</tr>
</tbody>
</table>
**Movie Release Date Questions:** Which is the movie that has the most recent release date?

<table>
<thead>
<tr>
<th></th>
<th>Movie 1</th>
<th>Movie 2</th>
<th>Movie 3</th>
<th>Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>A. V for Vendetta</td>
<td>B. The Matrix</td>
<td>C. Se7en</td>
<td>0.31</td>
<td>0.51</td>
</tr>
<tr>
<td>14</td>
<td>A. Twelve Monkeys</td>
<td>B. Annie Hall</td>
<td>C. Stand By Me</td>
<td>-0.63</td>
<td>0.57</td>
</tr>
<tr>
<td>15</td>
<td>A. The Wizard of Oz</td>
<td>B. Strangers on a Train</td>
<td>C. A Clockwork Orange</td>
<td>-0.67</td>
<td>0.66</td>
</tr>
<tr>
<td>16</td>
<td>A. The Usual Suspects</td>
<td>B. Alien</td>
<td>C. Some Like It Hot</td>
<td>-1.15</td>
<td>0.47</td>
</tr>
</tbody>
</table>

**Logic Questions:** What can we conclude from these two statements?

<table>
<thead>
<tr>
<th></th>
<th>Statement 1</th>
<th>Statement 2</th>
<th>Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>In a box, no blue things are triangular, and no triangular things are large.</td>
<td>A. No blue things are large.</td>
<td>-0.51</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>B. Some blue things are not large.</td>
<td>C. We can’t conclude anything about blue things and large things.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>In a box, some red things are square, and some square things are large.</td>
<td>A. Some red things are large.</td>
<td>-1.16</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>B. All red things are large.</td>
<td>C. We can’t conclude anything about red things and large things.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**State Population Questions:** Which is the most populous state?

<table>
<thead>
<tr>
<th></th>
<th>State 1</th>
<th>State 2</th>
<th>State 3</th>
<th>Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>A. West Virginia</td>
<td>B. Rhode Island</td>
<td>C. Oregon</td>
<td>0.68</td>
<td>0.47</td>
</tr>
<tr>
<td>20</td>
<td>A. Massachusetts</td>
<td>B. Maryland</td>
<td>C. Pennsylvania</td>
<td>0.29</td>
<td>0.46</td>
</tr>
</tbody>
</table>

**Answers:**

<table>
<thead>
<tr>
<th>QUESTION</th>
<th>ANSWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  What percentage of American employees can receive paid sick leave?</td>
<td>61</td>
</tr>
<tr>
<td>2  What is the average annual household income of Broadway theatergoers? (in U.S. dollars)</td>
<td>201500</td>
</tr>
<tr>
<td>3  How many tanning salons are there in the U.S.?</td>
<td>14000</td>
</tr>
<tr>
<td>4  How many inches of snow did Boston, MA, get this past winter season (2014-2015)?</td>
<td>108.6</td>
</tr>
<tr>
<td>5  What percentage of U.S. adults are vegetarians (someone who does not eat meat or seafood)?</td>
<td>5</td>
</tr>
<tr>
<td>6  What's the average cost for a middle-income family to raise a child born in 2013 until age 18? (in U.S. dollars)</td>
<td>245340</td>
</tr>
<tr>
<td>7  What percentage of U.S. citizens do not have health insurance?</td>
<td>12.9</td>
</tr>
<tr>
<td>8  According to the 2013 American community Survey, how many U.S. residents speak a language other than English at home? (in millions)</td>
<td>61.8</td>
</tr>
<tr>
<td>9  How much does college education cost per year on average in all institutions in the U.S.? (in U.S. dollars)</td>
<td>33047</td>
</tr>
<tr>
<td>10 What is the average math SAT score in the U.S., out of 800?</td>
<td>513</td>
</tr>
<tr>
<td>11 What is the average one-way commute time in the U.S. in minutes?</td>
<td>25.4</td>
</tr>
<tr>
<td>12 In what percentage of U.S. households does someone play video games?</td>
<td>67</td>
</tr>
<tr>
<td>13 What's the average age of someone who plays video games with some regularity?</td>
<td>30</td>
</tr>
<tr>
<td>14 What was the median gross rent between 2009-2013 in New York City, NY? (in U.S. dollars)</td>
<td>1200</td>
</tr>
<tr>
<td>15 What's the average age of a white female Broadway theatergoer?</td>
<td>44</td>
</tr>
<tr>
<td>16 How many people in the U.S. died in motor vehicle crashes in 2013?</td>
<td>32719</td>
</tr>
<tr>
<td>17 What's the average human life expectancy (in years) at birth in the world according to 2010 data?</td>
<td>67.2</td>
</tr>
<tr>
<td>18 How many business bachelor’s degrees were granted in the U.S. in 2011-2012?</td>
<td>366800</td>
</tr>
<tr>
<td>19 How many education bachelor’s degrees were granted in the U.S. in 2011-2012?</td>
<td>106800</td>
</tr>
<tr>
<td>20 What is the direct distance between San Francisco and Tokyo, in miles?</td>
<td>5140</td>
</tr>
</tbody>
</table>
APPENDIX B-2: QUESTIONS AND ANSWERS USED IN PILOT STUDIES

Pilot Study 1A

**CITY QUESTIONS:** Which is largest?

<table>
<thead>
<tr>
<th></th>
<th>A. Chicago, IL</th>
<th>B. Washington, D. C.</th>
<th>C. Seattle, WA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A. Philadelphia, PA</td>
<td>B. Boston, MA</td>
<td>C. Baltimore, MD</td>
</tr>
<tr>
<td>2</td>
<td>A. Phoenix, AZ</td>
<td>B. Atlanta, GA</td>
<td>C. Louisville, KY</td>
</tr>
<tr>
<td>3</td>
<td>A. San Francisco, CA</td>
<td>B. Dallas, TX</td>
<td>C. Portland, OR</td>
</tr>
<tr>
<td>4</td>
<td>A. Baltimore, MD</td>
<td>B. Columbus, OH</td>
<td>C. Las Vegas, NV</td>
</tr>
<tr>
<td>5</td>
<td>A. Las Vegas, NV</td>
<td>B. Milwaukee, WI</td>
<td>C. St. Louis, MO</td>
</tr>
<tr>
<td>6</td>
<td>A. Indianapolis, IN</td>
<td>B. Cincinnati, OH</td>
<td>C. Tampa, FL</td>
</tr>
<tr>
<td>7</td>
<td>A. Nashville, TN</td>
<td>B. Orlando, FL</td>
<td>C. Milwaukee, WI</td>
</tr>
<tr>
<td>8</td>
<td>A. Austin, TX</td>
<td>B. Dallas, TX</td>
<td>C. Atlanta, GA</td>
</tr>
<tr>
<td>10</td>
<td>A. San Antonio, TX</td>
<td>B. Hartford, CT</td>
<td>C. Austin, TX</td>
</tr>
</tbody>
</table>

**LOGIC QUESTIONS:**

11. All flowers have petals. Roses have petals.
   A. Roses are flowers.
   B. Some roses are flowers.
   C. We cannot conclude anything about roses and flowers.

12. All mammals walk. Whales are mammals.
   A. Whales walk.
   B. Whales do not walk.
   C. We cannot conclude whether whales walk.

13. All things that have a motor need oil. Automobiles need oil.
   A. Some things with motors are automobiles.
   B. Automobiles have a motor.
   C. We cannot conclude anything about automobiles and motors.

14. All living things need water. Roses need water.
   A. Roses are living things.
   B. Some living things are roses.
   C. We cannot conclude anything about roses and living things.

15. All vehicles have wheels. Boats are vehicles.
   A. Boats do not have wheels.
   B. Boats have wheels.
   C. We cannot conclude anything about boats and wheels.
**Matrix Questions:** Which fits best? (The images are listed at the end of this table.)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>
17.

18.
25.
**Pilot Study 1B**

<table>
<thead>
<tr>
<th><strong>CITY QUESTIONS:</strong> Which is largest?</th>
<th><strong>% CORRECT</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>26 A. Albany, NY</td>
<td>B. Cincinnati, OH</td>
</tr>
<tr>
<td>27 A. Denver, CO</td>
<td>B. Reno, NV</td>
</tr>
<tr>
<td>28 A. Detroit, MI</td>
<td>B. Houston, TX</td>
</tr>
<tr>
<td>29 A. Minneapolis, MN</td>
<td>B. Denver, CO</td>
</tr>
<tr>
<td>30 A. Seattle, WA</td>
<td>B. San Diego, CA</td>
</tr>
<tr>
<td>31 A. Pittsburgh, PA</td>
<td>B. Portland, OR</td>
</tr>
<tr>
<td>32 A. Honolulu, HI</td>
<td>B. Anchorage, AK</td>
</tr>
<tr>
<td>33 A. Syracuse, NY</td>
<td>B. Tucson, AZ</td>
</tr>
<tr>
<td>34 A. Boston, MA</td>
<td>B. Washington, D. C.</td>
</tr>
<tr>
<td>35 A. Philadelphia, PA</td>
<td>B. Baltimore, MD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>LOGIC QUESTIONS:</strong></th>
<th></th>
</tr>
</thead>
</table>
| 36 In a box, some red things are square, and some square things are large. | A. Some red things are large.  
B. All red things are large.  
C. We can’t conclude anything about red things and large things. | 62 |
| 37 In a box, no green things are round, and all round things are large. | A. No green things are large.  
B. Some green things are not large.  
C. We can’t conclude anything about green things and large things. | 60 |
| 38 In a box, no blue things are triangular, and no triangular things are large. | A. No blue things are large.  
B. Some blue things are not large.  
C. We can’t conclude anything about blue things and large things. | 57 |
| Jack is looking at Anne but Anne is looking at George. Jack is married but George is not. Is a married person looking at an unmarried person? | A. Yes  
B. No  
C. Cannot be determined | 12 |
COGNITIVE REFLECTION TASK (CRT) QUESTIONS:

<table>
<thead>
<tr>
<th>Question</th>
<th>A.</th>
<th>B.</th>
<th>C. Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>40. If it takes 2 nurses 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients?</td>
<td>200</td>
<td>2</td>
<td>Other 57</td>
</tr>
<tr>
<td>41. Soup and salad cost $5.50 in total. The soup costs a dollar more than the salad. How much does the salad cost?</td>
<td>$4.50</td>
<td>$3.50</td>
<td>Other 86</td>
</tr>
<tr>
<td>42. Sally is making sun tea. Every hour, the concentration of the tea doubles. If it takes 6 hours for the tea to be ready, how long would it take for the tea to reach half of the final concentration?</td>
<td>3 hours</td>
<td>5 hours</td>
<td>Other 48</td>
</tr>
</tbody>
</table>

WORD ANALOGY QUESTIONS

<table>
<thead>
<tr>
<th>Question</th>
<th>A.</th>
<th>B.</th>
<th>C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>43. to : too :: loot:</td>
<td>steal</td>
<td>toot</td>
<td>lute 57</td>
</tr>
<tr>
<td>44. haste : waste :: _____ : crowd</td>
<td>fast</td>
<td>three</td>
<td>group 57</td>
</tr>
<tr>
<td>45. _____ : rainfall :: condensation : humidity</td>
<td>ground</td>
<td>cloud</td>
<td>forecast NA</td>
</tr>
<tr>
<td>46. no : know :: steal : _____</td>
<td>rob</td>
<td>steel</td>
<td>don’t 88</td>
</tr>
<tr>
<td>47. essay : thesis :: article :: _____</td>
<td>book</td>
<td>fiction</td>
<td>topic 52</td>
</tr>
</tbody>
</table>

ANSWERS:


35 This question did not have one correct answer and was excluded from data analyses.
CITY POPULATION QUESTIONS: Which is the most populous city proper? A city proper is an area contained within city limits. A city proper may not include suburbs.

<table>
<thead>
<tr>
<th></th>
<th>A. Shanghai, China</th>
<th>B. Delhi, India</th>
<th>C. Istanbul, Turkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A. Moscow, Russia</td>
<td>B. Cairo, Egypt</td>
<td>C. Beijing, China</td>
</tr>
<tr>
<td>2</td>
<td>A. Tokyo, Japan</td>
<td>B. Sao Paulo, Brazil</td>
<td>C. Mexico City, Mexico</td>
</tr>
<tr>
<td>3</td>
<td>A. Seoul, South Korea</td>
<td>B. Tokyo, Japan</td>
<td>C. Istanbul, Turkey</td>
</tr>
<tr>
<td>4</td>
<td>A. London, United Kingdom</td>
<td>B. Moscow, Russia</td>
<td>C. New York City, The United States of America</td>
</tr>
<tr>
<td>5</td>
<td>A. Hong Kong, China</td>
<td>B. Mexico City, Mexico</td>
<td>C. Baghdad, Iraq</td>
</tr>
<tr>
<td>6</td>
<td>A. Los Angeles, The United States of America</td>
<td>B. Berlin, Germany</td>
<td>C. Madrid, Spain</td>
</tr>
<tr>
<td>7</td>
<td>A. Ankara, Turkey</td>
<td>B. Singapore, Singapore</td>
<td>C. Seoul, South Korea</td>
</tr>
<tr>
<td>8</td>
<td>A. London, United Kingdom</td>
<td>B. Berlin, Germany</td>
<td>C. Toronto, Canada</td>
</tr>
<tr>
<td>9</td>
<td>A. Rio de Janeiro, Brazil</td>
<td>B. Los Angeles, The United States of America</td>
<td>C. Ankara, Turkey</td>
</tr>
<tr>
<td>10</td>
<td>A. Tokyo, Japan</td>
<td>B. Singapore, Singapore</td>
<td>C. Moscow, Russia</td>
</tr>
<tr>
<td>11</td>
<td>A. London, United Kingdom</td>
<td>B. Los Angeles, The United States of America</td>
<td>C. Seoul, South Korea</td>
</tr>
<tr>
<td>12</td>
<td>A. Beijing, China</td>
<td>B. Mexico City, Mexico</td>
<td>C. Shanghai, China</td>
</tr>
<tr>
<td>13</td>
<td>A. Mexico City, Mexico</td>
<td>B. Rio de Janeiro, Brazil</td>
<td>C. Berlin, Germany</td>
</tr>
<tr>
<td>14</td>
<td>A. Baghdad, Iraq</td>
<td>B. New York City, The United States of America</td>
<td>C. Tokyo, Japan</td>
</tr>
<tr>
<td>15</td>
<td>A. New York City, The United States of America</td>
<td>B. Tokyo, Japan</td>
<td>C. Madrid, Spain</td>
</tr>
<tr>
<td></td>
<td>A. Sao Paulo, Brazil</td>
<td>B. Seoul, South Korea</td>
<td>C. Delhi, India</td>
</tr>
<tr>
<td>---</td>
<td>---------------------</td>
<td>----------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>16</td>
<td>A. Madrid, Spain</td>
<td>B. Baghdad, Iraq</td>
<td>C. Paris, France</td>
</tr>
<tr>
<td>17</td>
<td>A. Cairo, Egypt</td>
<td>B. Hong Kong, China</td>
<td>C. London, United Kingdom</td>
</tr>
<tr>
<td>18</td>
<td>A. Los Angeles, The United States of America</td>
<td>B. Istanbul, Turkey</td>
<td>C. Moscow, Russia</td>
</tr>
<tr>
<td>19</td>
<td>A. Singapore, Singapore</td>
<td>B. Delhi, India</td>
<td>C. Madrid, Spain</td>
</tr>
</tbody>
</table>

**Country Population Questions:** Which is the most populous country?

<table>
<thead>
<tr>
<th></th>
<th>A. The United States of America</th>
<th>B. Russia</th>
<th>C. China</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>A. Spain</td>
<td>B. Argentina</td>
<td>C. Japan</td>
</tr>
<tr>
<td>22</td>
<td>A. Russia</td>
<td>B. Canada</td>
<td>C. India</td>
</tr>
<tr>
<td>23</td>
<td>A. Ukraine</td>
<td>B. Italy</td>
<td>C. Finland</td>
</tr>
<tr>
<td>24</td>
<td>A. Poland</td>
<td>B. Israel</td>
<td>C. Canada</td>
</tr>
<tr>
<td>25</td>
<td>A. Iran</td>
<td>B. Portugal</td>
<td>C. Italy</td>
</tr>
<tr>
<td>26</td>
<td>A. Australia</td>
<td>B. Turkey</td>
<td>C. Canada</td>
</tr>
<tr>
<td>27</td>
<td>A. Greece</td>
<td>B. Sweden</td>
<td>C. Iceland</td>
</tr>
<tr>
<td>28</td>
<td>A. Norway</td>
<td>B. Denmark</td>
<td>C. Germany</td>
</tr>
<tr>
<td>29</td>
<td>A. Iceland</td>
<td>B. Switzerland</td>
<td>C. Finland</td>
</tr>
<tr>
<td>30</td>
<td>A. Israel</td>
<td>B. Iran</td>
<td>C. Egypt</td>
</tr>
<tr>
<td>31</td>
<td>A. Austria</td>
<td>B. Hungary</td>
<td>C. Greece</td>
</tr>
<tr>
<td>32</td>
<td>A. Belgium</td>
<td>B. Poland</td>
<td>C. Ireland</td>
</tr>
<tr>
<td>33</td>
<td>A. Germany</td>
<td>B. Brazil</td>
<td>C. Poland</td>
</tr>
<tr>
<td>34</td>
<td>A. Turkey</td>
<td>B. United Kingdom</td>
<td>C. Ukraine</td>
</tr>
<tr>
<td></td>
<td>A. Country</td>
<td>B. Country</td>
<td>C. Country</td>
</tr>
<tr>
<td>---</td>
<td>-------------</td>
<td>------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>36</td>
<td>A. Italy</td>
<td>B. Ireland</td>
<td>C. Austria</td>
</tr>
<tr>
<td>37</td>
<td>A. France</td>
<td>B. Egypt</td>
<td>C. Australia</td>
</tr>
<tr>
<td>38</td>
<td>A. Portugal</td>
<td>B. Brazil</td>
<td>C. Turkey</td>
</tr>
<tr>
<td>39</td>
<td>A. Mexico</td>
<td>B. Argentina</td>
<td>C. Canada</td>
</tr>
<tr>
<td>40</td>
<td>A. Canada</td>
<td>B. Japan</td>
<td>C. The United States of America</td>
</tr>
</tbody>
</table>

**State Population Questions:** Which is the most populous state?

<table>
<thead>
<tr>
<th></th>
<th>A. State</th>
<th>B. State</th>
<th>C. State</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>A. Texas</td>
<td>B. California</td>
<td>C. Alaska</td>
</tr>
<tr>
<td>42</td>
<td>A. Arizona</td>
<td>B. Colorado</td>
<td>C. Nevada</td>
</tr>
<tr>
<td>43</td>
<td>A. Wyoming</td>
<td>B. South Dakota</td>
<td>C. Vermont</td>
</tr>
<tr>
<td>44</td>
<td>A. New York</td>
<td>B. Ohio</td>
<td>C. Florida</td>
</tr>
<tr>
<td>45</td>
<td>A. Illinois</td>
<td>B. Connecticut</td>
<td>C. Rhode Island</td>
</tr>
<tr>
<td>46</td>
<td>A. Washington</td>
<td>B. Florida</td>
<td>C. Maryland</td>
</tr>
<tr>
<td>47</td>
<td>A. Tennessee</td>
<td>B. Kentucky</td>
<td>C. Oklahoma</td>
</tr>
<tr>
<td>48</td>
<td>A. Virginia</td>
<td>B. North Carolina</td>
<td>C. Michigan</td>
</tr>
<tr>
<td>49</td>
<td>A. New Jersey</td>
<td>B. Ohio</td>
<td>C. Mississippi</td>
</tr>
<tr>
<td>50</td>
<td>A. Massachusetts</td>
<td>B. Maryland</td>
<td>C. Pennsylvania</td>
</tr>
<tr>
<td>51</td>
<td>A. Georgia</td>
<td>B. Missouri</td>
<td>C. Montana</td>
</tr>
<tr>
<td>52</td>
<td>A. New Mexico</td>
<td>B. Kansas</td>
<td>C. Utah</td>
</tr>
<tr>
<td>53</td>
<td>A. Indiana</td>
<td>B. Utah</td>
<td>C. Delaware</td>
</tr>
<tr>
<td>54</td>
<td>A. Michigan</td>
<td>B. Iowa</td>
<td>C. New Jersey</td>
</tr>
<tr>
<td>55</td>
<td>A. Minnesota</td>
<td>B. Pennsylvania</td>
<td>C. South Carolina</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>A. Wisconsin</td>
<td>B. Alabama</td>
<td>C. Indiana</td>
</tr>
<tr>
<td>57</td>
<td>A. West Virginia</td>
<td>B. Rhode Island</td>
<td>C. Oregon</td>
</tr>
<tr>
<td>58</td>
<td>A. Vermont</td>
<td>B. Idaho</td>
<td>C. Hawaii</td>
</tr>
<tr>
<td>59</td>
<td>A. Maine</td>
<td>B. Alaska</td>
<td>C. Louisiana</td>
</tr>
<tr>
<td>60</td>
<td>A. Arizona</td>
<td>B. Arkansas</td>
<td>C. Massachusetts</td>
</tr>
</tbody>
</table>

**University Ranking Questions:** Which is the highest ranking university in 2014? The ranking is entirely academic and research oriented. It does not include athletics.
<table>
<thead>
<tr>
<th></th>
<th>A.</th>
<th>B.</th>
<th>C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>76</td>
<td>A. Rutgers University</td>
<td>B. New York University</td>
<td>C. University of Chicago</td>
</tr>
<tr>
<td>77</td>
<td>A. Northwestern University</td>
<td>B. University of Pennsylvania</td>
<td>C. University of California, Los Angeles</td>
</tr>
<tr>
<td>78</td>
<td>A. Stanford University</td>
<td>B. Princeton University</td>
<td>C. California Institute of Technology</td>
</tr>
<tr>
<td>79</td>
<td>A. Yale University</td>
<td>B. University of Chicago</td>
<td>C. Duke University</td>
</tr>
<tr>
<td>80</td>
<td>A. Arizona State University</td>
<td>B. Rutgers University</td>
<td>C. New York University</td>
</tr>
</tbody>
</table>

**Movie Release Date Questions:** Which is the movie that has the most recent release date?

<table>
<thead>
<tr>
<th></th>
<th>A.</th>
<th>B.</th>
<th>C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>A. The Godfather</td>
<td>B. The Good, The Bad, and The Ugly</td>
<td>C. 12 Angry Men</td>
</tr>
<tr>
<td>82</td>
<td>A. One Flew Over the Cuckoo’s Nest</td>
<td>B. Full Metal Jacket</td>
<td>C. Scarface</td>
</tr>
<tr>
<td>83</td>
<td>A. The Princess Bride</td>
<td>B. Groundhog Day</td>
<td>C. Rocky</td>
</tr>
<tr>
<td>84</td>
<td>A. Roman Holiday</td>
<td>B. The Big Sleep</td>
<td>C. The Hustler</td>
</tr>
<tr>
<td>85</td>
<td>A. Twelve Monkeys</td>
<td>B. Annie Hall</td>
<td>C. Stand By Me</td>
</tr>
<tr>
<td>86</td>
<td>A. The Wizard of Oz</td>
<td>B. Strangers on a Train</td>
<td>C. A Clockwork Orange</td>
</tr>
<tr>
<td>87</td>
<td>A. Psycho</td>
<td>B. The Shining</td>
<td>C. Singin’ in The Rain</td>
</tr>
<tr>
<td>88</td>
<td>A. V for Vendetta</td>
<td>B. The Matrix</td>
<td>C. Se7en</td>
</tr>
<tr>
<td>89</td>
<td>A. Fargo</td>
<td>B. The Silence of the Lambs</td>
<td>C. Goodfellas</td>
</tr>
<tr>
<td>90</td>
<td>A. It’s a Wonderful Life</td>
<td>B. Casablanca</td>
<td>C. Breakfast at Tiffany’s</td>
</tr>
<tr>
<td>91</td>
<td>A. Trainspotting</td>
<td>B. The Sixth Sense</td>
<td>C. Jurassic Park</td>
</tr>
<tr>
<td>92</td>
<td>A. Gone with the Wind</td>
<td>B. Ben-Hur</td>
<td>C. The Grapes of Wrath</td>
</tr>
<tr>
<td>93</td>
<td>A. The Usual Suspects</td>
<td>B. Alien</td>
<td>C. Some Like It Hot</td>
</tr>
<tr>
<td>94</td>
<td>A. Pulp Fiction</td>
<td>B. Fight Club</td>
<td>C. Dead Poets Society</td>
</tr>
<tr>
<td>95</td>
<td>A. Rear Window</td>
<td>B. Jaws</td>
<td>C. 2001: A Space Odyssey</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>A. Reservoir Dogs</td>
<td>B. Blade Runner</td>
<td>C. Chinatown</td>
</tr>
<tr>
<td>97</td>
<td>A. The Terminator</td>
<td>B. Die Hard</td>
<td>C. Braveheart</td>
</tr>
<tr>
<td>98</td>
<td>A. Brokeback Mountain</td>
<td>B. Slumdog Millionaire</td>
<td>C. A Beautiful Mind</td>
</tr>
<tr>
<td>99</td>
<td>A. Gladiator</td>
<td>B. The Truman Show</td>
<td>C. Million Dollar Baby</td>
</tr>
<tr>
<td>100</td>
<td>A. Elf</td>
<td>B. Iron Man</td>
<td>C. Billy Elliot</td>
</tr>
</tbody>
</table>

**LOGIC QUESTIONS:** What can we conclude from these two statements?

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>All flowers have petals. Roses have petals. A. Roses are flowers. B. Flowers are roses. C. None of the above.</td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>All mammals walk. Whales are mammals. A. Whales walk. B. Whales do not walk. C. None of the above.</td>
<td></td>
</tr>
<tr>
<td>103</td>
<td>All things that have a motor need oil. Automobiles need oil. A. Automobiles do not have a motor. B. Automobiles have a motor. C. None of the above.</td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>All living things need water. Roses need water. A. Roses are living things. B. Roses are not living things. C. None of the above.</td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>All vehicles have wheels. Boats are vehicles. A. Boats have wheels. B. Boats do not have wheels. C. None of the above.</td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>In a box, some red things are square, and some square things are large. A. Some red things are large. B. All red things are large. C. We can’t conclude anything about red things and large things.</td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>In a box, no green things are round, and all round things are large. A. No green things are large. B. Some green things are not large. C. We can’t conclude anything about green things and large things.</td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>In a box, no blue things are triangular, and no triangular things are large. A. No blue things are large. B. Some blue things are not large. C. We can’t conclude anything about blue things and large things.</td>
<td></td>
</tr>
</tbody>
</table>
ARITHMETIC AND OTHER QUESTIONS:

109 If it takes 2 nurses 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients?

A. 200 B. 2 C. Other

110 Soup and salad cost $5.50 in total. The soup costs a dollar more than the salad. How much does the salad cost?

A. $4.50 B. $3.50 C. Other

111 Sally is making sun tea. Every hour, the concentration of the tea doubles. If it takes 6 hours for the tea to be ready, how long would it take for the tea to reach half of the final concentration?

A. 3 hours B. 5 hours C. Other

112 Jack is looking at Anne but Anne is looking at George. Jack is married but George is not. Is a married person looking at an unmarried person?

A. Yes B. No C. Cannot be determined

113 Ann’s father has a total of five daughters: Lala, Lele, Lili, Lolo, and ___. What is the name of the fifth daughter?

A. Lulu B. Lora C. Other

114 On the side of a boat hangs a ladder with six rungs. Each rung is one foot from the next one, and the bottom rung is resting on the surface of the water. The tide rises at a rate of one foot an hour. How long will take the water to reach the top rung?

A. 5 hours B. 6 hours C. Never

ANSWERS:


151
<table>
<thead>
<tr>
<th>NUMBER</th>
<th>QUESTION</th>
<th>ANSWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>How many business bachelor's degrees were granted in the U.S. 2011-2012?</td>
<td>109000</td>
</tr>
<tr>
<td>2</td>
<td>How many psychology bachelor's degrees were granted in the U.S. in 2011-2012?</td>
<td>366800</td>
</tr>
<tr>
<td>3</td>
<td>How many health professions and related programs bachelor's degrees were granted in the U.S. in 2011-2012?</td>
<td>163400</td>
</tr>
<tr>
<td>4</td>
<td>How many education bachelor's degrees were granted in the U.S. in 2011-2012?</td>
<td>106800</td>
</tr>
<tr>
<td>5</td>
<td>What percentage of American employees can receive paid sick leave?</td>
<td>61</td>
</tr>
<tr>
<td>6</td>
<td>How many employees of Aetna, a health insurance company, have taken the company's free yoga classes so far?</td>
<td>13000</td>
</tr>
<tr>
<td>7</td>
<td>What is the average annual household income of Broadway theatergoers? (in U.S. dollars)</td>
<td>201500</td>
</tr>
<tr>
<td>8</td>
<td>How many construction workers in India are women? (in millions)</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>How many tanning salons are there in the U.S.?</td>
<td>14000</td>
</tr>
<tr>
<td>10</td>
<td>How many automobiles were sold in the U.S. in 2014? (in millions)</td>
<td>16.5</td>
</tr>
<tr>
<td>11</td>
<td>How many inches of snow did Boston, MA, get this past winter season (2014-2015)?</td>
<td>108.6</td>
</tr>
<tr>
<td>12</td>
<td>What percentage of U.S. adults are vegetarians?</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>How much profit did the American insurance industry make in 2014? (in billion U.S. dollars)</td>
<td>338</td>
</tr>
<tr>
<td>14</td>
<td>What's the average cost for a middle-income family to raise a child born in 2013 until age 18? (in U.S. dollars)</td>
<td>245340</td>
</tr>
<tr>
<td>15</td>
<td>How much money did the movies released in 2014 make? (in billion U.S. dollars)</td>
<td>9.69</td>
</tr>
<tr>
<td>Question</td>
<td>Answer</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>16 What percentage of U.S. citizens do not have health insurance?</td>
<td>12.9</td>
<td></td>
</tr>
<tr>
<td>17 How many original TV shows aired in the U.S. in 2014?</td>
<td>328</td>
<td></td>
</tr>
<tr>
<td>18 How many pounds of Haas avocados (the avocados one normally purchased at the U.S. grocery stores) were sold in the U.S. in 2014? (in billions)</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>19 How many U.S. households own dogs?</td>
<td>43346000</td>
<td></td>
</tr>
<tr>
<td>20 How many U.S. households own cats?</td>
<td>36117000</td>
<td></td>
</tr>
<tr>
<td>21 How many U.S. households own horses?</td>
<td>1780000</td>
<td></td>
</tr>
<tr>
<td>According to the 2013 American community survey, how many U.S. residents speak a language other than English at home? (in millions)</td>
<td>61.8</td>
<td></td>
</tr>
<tr>
<td>22 How much does college education cost per year on average in all institutions in the U.S.? (in U.S. dollars)</td>
<td>33047</td>
<td></td>
</tr>
<tr>
<td>23 What is the average math SAT score in the U.S., out of 800?</td>
<td>513</td>
<td></td>
</tr>
<tr>
<td>24 What is the average one-way commute time in the U.S. in minutes?</td>
<td>25.4</td>
<td></td>
</tr>
<tr>
<td>25 In what percentage of U.S. households does someone play video games?</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>26 What's the average age of someone who plays video games with some regularity?</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>27 What percentage of women in the U.S. play video games?</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>28 What was the median gross rent between 2009-2013 in New York City, NY? (in U.S. dollars)</td>
<td>1200</td>
<td></td>
</tr>
<tr>
<td>29 In 2013 for how many people in the U.S. influenza was the cause of death?</td>
<td>3697</td>
<td></td>
</tr>
<tr>
<td><strong>QUESTION</strong></td>
<td><strong>ANSWER</strong></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------------------------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>31 In 2013 for how many people in the U.S. Alzheimer's Disease was the cause of death?</td>
<td>84767</td>
<td></td>
</tr>
<tr>
<td>32 How many marriages were there in the U.S. in 2012?</td>
<td>2131000</td>
<td></td>
</tr>
<tr>
<td>33 How many divorces and annulments were there in the U.S. in 2012?</td>
<td>851000</td>
<td></td>
</tr>
<tr>
<td>34 In 2013 how many unmarried women gave birth?</td>
<td>1595873</td>
<td></td>
</tr>
<tr>
<td>35 What's the U.S. per capita red meat consumption for 2014 in pounds?</td>
<td>101.4</td>
<td></td>
</tr>
<tr>
<td>36 What's the average age of a white female Broadway theatergoer?</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>37 How many people in the U.S. died in motor vehicle crashes in 2013?</td>
<td>32719</td>
<td></td>
</tr>
<tr>
<td>38 What's the average human life expectancy (in years) at birth in the world according to 2010 data?</td>
<td>67.2</td>
<td></td>
</tr>
<tr>
<td>39 How many books do American adults (ages 18+) on average read per year?</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>40 What is the direct distance between San Francisco and Tokyo, in miles?</td>
<td>5140</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX C-1: STUDY INSTRUCTIONS FOR SUBJECTS

STUDY 1A

This study contains 20 questions, each with three possible answers, A, B, and C. We are interested in how you think about the answers. After each question, we ask you to make notes about your thoughts. Each note should give one, and only one, reason for or against one of the three possible answers. You have space for 6 notes, which should be plenty; you don't have to use all 6, but you must give at least two reasons. Please refrain from writing non-English phrases or putting in statements such as "I don't know" or "I'm not sure" that are not actually reasons. We would like you to say briefly why you think an answer might be correct or wrong.

Then we ask for your preferred answer and the probabilities for any answer being the correct choice. For the probability questions, note that 100% means that you are completely certain that an answer is correct. But if you just guess, you have about a 33% chance of being correct. We give you choices for the probability; choose the one that is closest to what you think. Your probability estimates for the three answers should add up to about 100% (no more than 110%, no less than 90%).

On the next page, after the reasons, we ask you to classify each of your reasons, telling us whether it is for, or against, Answer A, Answer B, or Answer C.

The study ends with one additional page of questions about thinking.

On Each Question Page:

In this study, you are given three city names, and are asked to determine which of the three cities is located in the most populous metropolitan areas of the United States. According to Wikipedia, a metropolitan area is "a region consisting of a densely populated urban area and its less-populated surrounding territories, sharing industry, infrastructure, and housing". Your task is to select the city that is in the most populated urban area.

Examples of reasons: 'It looks bigger on a map'; 'It is a national capital'; 'It has two airports'; 'I have heard more about it'; 'It has more sports teams'. Any one of these can be made negative, for example, 'It looks small on a map'.

STUDY 1B

This study contains 20 questions, each with three possible answers, A, B, and C. We are interested in how you think about the answers. After each question, we ask you to make
notes about your thoughts. Each note should give one, and only one, reason for or against
one of the three possible answers. You have space for 6 notes; you don't have to use all 6,
but you must give at least two reasons. Please refrain from putting in statements such as
"I don't know" or "I'm not sure" that are not actually reasons. We would like you to say
briefly why you think an answer might be correct or incorrect.

Then we ask for your preferred answer and the probabilities for any answer being the
correct choice. For the probability questions, note that 100% means that you are
completely certain that an answer is correct. But if you just guess, you have about a 33%
chance of being correct. We give you choices for the probability; choose the one that is
closest to what you think. Your probabilities for the three answers should add up to about
100% (no more than 110%, no less than 90%).

In this study, you are given three food names or descriptions, and you are asked to
determine which of the three foods listed have the highest calorie content, fiber content,
fat content, or protein content. Please pay attention to what the question is specifically
asking for.
Examples of reasons are: 'tastes sweet'; 'vegetables are usually high fiber'; 'meat usually
has fat'; 'servings are small'.

On the next page, after the reasons, we ask you to classify each of your reasons, telling us
whether it is for, or against, Answer A, Answer B, or Answer C.

The study ends with one additional page of questions about thinking.

For Calorie Content Items:

For the following questions, please pick the food with the highest CALORIE content per
serving.

For Fiber Content Items:

For the following questions below, please pick the food that has the most amount of
FIBER per 100 grams (~ 3.53 ounces).

For Fat Content Items:

For the following questions, please pick the food with the highest FAT content per 100
grams (~ 3.53 ounces).

For Protein Content Items:
For the following questions, please pick the food with the highest PROTEIN content per serving.

STUDY 2

This study contains 20 questions, each with three possible answers, A, B, and C. Please pay attention to what each question is asking before you answer.

We are interested in how you think about the answers. After each question, we ask you to make notes about your thoughts. Each note should give one, and only one, reason for or against one of the three possible answers. You have space for 6 notes; you don't have to use all 6, but you must give at least two reasons. Please refrain from putting in statements such as "I don't know" or "I'm not sure" that are not actually reasons. We would like you to say briefly why you think an answer might be correct or incorrect.

Here is an example question:

Please pick the most populous country below.
1. Norway
2. Denmark
3. Germany

Example Reasons:
Germany has a lot of cities.
Denmark has Greenland, which is very big.
Greenland is not really populated.

Then we ask for your preferred answer and the probabilities for any answer being the correct choice. For the probability questions, note that 100% means that you are completely certain that an answer is correct. But if you just guess, you have about a 33% chance of being correct. We give you choices for the probability; choose the one that is closest to what you think. Your probabilities for the three answers should add up to about 100% (no more than 110%, no less than 90%).

On the next page, after the reasons, we ask you to classify each of your reasons, telling us whether it is for, or against, Answer A, Answer B, or Answer C.

Example Reason Classification:
Germany has a lot of cities (For C)
Denmark has Greenland, which is very big. (For B)
Greenland is not really populated. (Against B)
The study ends with one additional page of questions about thinking.

STUDY 3

This study contains 20 questions about numbers. Try to answer the questions as accurately as you can without getting any outside or online help.

We are also interested in how you think about the answers. After each question, we ask you to make notes about your reasons. Each note should give one, and only one, reason. You have space for 5 notes. You don't have to use all 5, but you must give at least two reasons.

Below is an example question with possible answers and reasons to help you get a better idea about what this study will be like.

Question: How many cups of coffee on average are consumed by Americans consume every day?
Let's assume that you gave the following answer.

Best estimate: 300,000,000
Lower number: 250,000,000
Higher number: 400,000,000

Next, you list some reasons/factors that affect your answers. Then you classify each reason as to its relation to your answer. For example:

Reasons:

The U.S. population is around 300 million. (classification: why my answer makes sense)
People work long hours in this country. (why my answer makes sense)
Not everyone drinks/likes coffee. (why my answer might be too high)
Each coffee drinker is likely to drink several cups. (why it might be too low)
Maybe my figure about the population is way off. (why it might be too high or too low)

(The correct number is 400 million cups a day.)

The study ends with one additional page of questions about thinking.

STUDY 4
This study contains 20 questions, each with three possible answers, A, B, and C. Please pay attention to what each question is asking before you answer. Try to answer the questions as accurately as you can without getting any outside or online help.

We are also interested in how you think about the answers. After each question, we ask you to make notes about your thoughts. Each note should give one, and only one, reason for or against one of the three possible answers. You have space for 5 notes; you don't have to use all 5, but you must give at least two reasons. On some pages, you will also be asked to put in different kinds of reasons, so please pay attention to the instructions that appear in blue.

Here is an example question:

Please pick the most populous country below.
1. Norway
2. Denmark
3. Germany

Example Reasons:
Germany has a lot of cities.
Denmark has Greenland, which is very big.
Greenland is not really populated.

Then we ask for your preferred answer and the probabilities for any answer being the correct choice. For the probability questions, note that 100% means that you are completely certain that an answer is correct. But if you just guess, you have about a 33% chance of being correct. We give you choices for the probability; choose the one that is closest to what you think. Your probabilities for the three answers should add up to about 100% (no more than 110%, no less than 90%).

After you put in your probability judgments, we ask you to classify each of your reasons, telling us whether it is for, or against, Answer A, Answer B, or Answer C, or none in particular.

Example Reason Classification:
Germany has a lot of cities (For C)
Denmark has Greenland, which is very big. (For B)
Greenland is not really populated. (Against B)

The study ends with one additional page of questions about thinking.

STUDY 5
This study contains 20 questions about numbers. Try to answer the questions as accurately as you can without getting any outside or online help.

We are also interested in how you think about the answers. After each question, we ask you to make notes about your reasons. Each note should give one, and only one, reason. You have space for 5 notes. You don't have to use all 5, but you must give at least two reasons.

Below is an example question with possible answers and reasons to help you get a better idea about what this study will be like.

Question: How many cups of coffee on average are consumed by Americans consume every day?
Let's assume that you gave the following answer.

Best estimate: 300,000,000
Lower number: 250,000,000
Higher number: 400,000,000

Next, you list some reasons/factors that affect your answers. Then you classify each reason as to its relation to your answer. For example:

Reasons:

The U.S. population is around 300 million. (classification: why my answer makes sense)
People work long hours in this country. (why my answer makes sense)
Not everyone drinks/likes coffee. (why my answer might be too high)
Each coffee drinker is likely to drink several cups. (why it might be too low)
Maybe my figure about the population is way off. (why it might be too high or too low)

(The correct number is 400 million cups a day.)

The study ends with one additional page of questions about thinking.

STUDY 6

This study contains 20 questions about numbers. Try to answer the questions as accurately as you can without getting any outside or online help. You will be asked to submit your best estimate, a lower bound that is lower than your best estimate, and a higher bound, that is higher than your best estimate. Some questions tell you in which unit your answer should be, so please pay attention to these.
We are also interested in how you think about the answers. After each question, we ask you to make notes about your reasons. Each note should give one, and only one, reason. You have space for 5 notes. You don't have to use all 5, but you must give at least two reasons.

Below is an example question with possible answers and reasons to help you get a better idea about what this study will be like.

Question: How many cups of coffee on average are consumed by Americans every day?

Let's assume that you gave the following answers.

Best estimate: 300,000,000
Lower number: 250,000,000
Higher number: 400,000,000

Next, you list some reasons/factors that affect your answers. Then you classify each reason as to its relation to your answer. For example:

Reasons:

The U.S. population is around 300 million. (classification: why my answer makes sense)
People work long hours in this country. (why my answer makes sense)
Not everyone drinks/likes coffee. (why my answer might be too high)
Each coffee drinker is likely to drink several cups. (why it might be too low)
Maybe my figure about the population is way off. (why it might be too high or too low)
(The correct number is 400 million cups a day.)

The study ends with one additional page of questions about thinking.

STUDY 7

First Part Instructions:

This study contains 20 questions, each with three possible answers, A, B, and C. Please pay attention to what each question is asking before you answer. Try to answer the questions as accurately as you can without getting any outside or online help.

We are interested in how you think about the answers. After each question, we ask you to make notes about your thoughts. Each note should give one, and only one, reason for or against one of the three possible answers. You have space for 5 notes; you don't have to use all 5, but you must give at least two reasons.
Here is an example question:

Please pick the most populous country below.
1. Norway
2. Denmark
3. Germany
Example Reasons:
Germany has a lot of cities.
Denmark has Greenland, which is very big.
Greenland is not really populated.

We ask you to classify each of your reasons, telling us whether it is for, or against, Answer A, Answer B, or Answer C, or 'other'. Notice that if an answer is 'for B or C' then you can classify it as 'against A', because these are the only three options. If a reason does not help in distinguishing one option from another, use the 'other' response.

Example Reason Classification:
Germany has a lot of cities (For C)
Denmark has Greenland, which is very big. (For B)
Greenland is not really populated. (Against B)

Finally we ask for your preferred answer and the probabilities for any answer being the correct choice. For the probability questions, note that 100% means that you are completely certain that an answer is correct. But if you just guess, you have about a 33% chance of being correct. We give you choices for the probability; choose the one that is closest to what you think. Your probabilities for the three answers should add up to about 100% (no more than 110%, no less than 90%).

The study ends with one additional page of questions about thinking.

Second Part Instructions:

This is the same study as you did before. We think you may do better when you do it a second time.

Again, we ask you to classify your answers as being for A, B, C or against one of these, or 'other'. A reason that is 'against A' is the same as one that is 'for B or C'. If a reason does not help in distinguishing one option from another, use the 'Other' response.
Again we ask for your preferred answer and the probabilities for any answer being the correct choice. For the probability questions, note that 100% means that you are completely certain that an answer is correct. But if you just guess, you have about a 33% chance of being correct. We give you choices for the probability; choose the one that is closest to what you think. Your probabilities for the three answers should add up to about 100% (no more than 110%, no less than 90%).

The study ends with one additional page of questions about thinking.
PILOT STUDY 1

Introduction Page:

This study contains a number of difficult questions, each with three possible answers, A, B, and C. We are interested in how you think about the answers. After each question, we ask you to make notes about your thoughts. Each note should give one, and only one, reason for or against one of the three possible answers. You have space for 10 notes, which should be plenty; you don't have to use all 10, but you must give at least one reason.

Then we ask for your preferred answer, and the probability that you are correct. For the probability question, note that 100% means that you are completely certain. But if you just guess, you have about a 33% chance of being correct. We give you choices for the probability; choose the one that is closest to what you think.

On the next page, we ask you to classify each of your reasons, telling us whether it is for, or against, Answer A, Answer B, or Answer C.

The remaining parts of the study contain problems in logic, arithmetic, visual analogies (with patterns), and/or verbal analogies.

For City Items:

In the first part of the study, you are given three city names, and are asked to determine which of the three cities is located in the most populous metropolitan areas of the United States. According to Wikipedia, a metropolitan area is "a region consisting of a densely populated urban area and its less-populated surrounding territories, sharing industry, infrastructure, and housing". Your task is to select the city that is in the most populated urban area.

Examples of reasons: 'It looks bigger on a map'; 'It is a national capital'; 'It has two airports'; 'I have heard more about it'; 'It has more sports teams'. Any one of these can be made negative, for example, 'It looks small on a map'.

For Logic Items:

This part is about logic. All questions give two statements and ask what we can conclude if both of these are true.
Examples of reasons are: “The conclusion is true”; “I know another example like this”; “I can think of an example like this where the conclusion is false”; “The two statements fit together to yield the conclusion.”

For Cognitive Reflection Task (CRT) Items:

This part is about math problems. Examples of reasons are: ’This is X minus Y (where X and Y are numbers in the problem’; ’Next number in the sequence’.

For Matrix Items:

This part is about geometric matrices. Examples of reasons are: “This is similar to another figure,” “... completes a sequence”; “... fills in the a pattern”; “The shapes get larger as they go from left to right and from top to bottom”; “In each row, the first two images are parts of the last image.”

For Word Analogy Items:

This part is about word analogies. Examples of reasons are: “Same relation as A and B (the first two terms”’; “Similar to C”; “Same category as C”; “The words are synonyms, are so are A and B”; “The relation between A and B is the same as that between C and D.”

PILOT STUDY 2

This study contains 60 questions, each with three possible answers, A, B, and C. It is a preliminary study to test the items themselves.

We ask for your preferred answer and the probabilities for any answer being the correct choice. For the probability questions, note that 100% means that you are completely certain that an answer is correct. If you just guess, you have about a 33% chance of being correct. We give you choices for the probability so that you don't have to enter the numbers yourself; choose the one that is closest to what you think. Your probabilities for the three answers should add up to about 100% (no more than 110%, no less than 90%).

Please pay attention to what the question is specifically asking for.

PILOT STUDY 3

This study contains 40 questions about numbers. Try to answer the questions as accurately as you can without getting any outside or online help.

The study ends with one additional page of questions about thinking.
APPENDIX D-1: TRAINING MODULE FOR STUDY 6

The following pages include the training module for subjects in the experimental condition.
Training in actively open-minded thinking

Before you start the training please enter your email address again and then press NEXT. (Please use the one with which you registered.)

Training in actively open-minded thinking

The purpose of these exercises is to teach you a way of thinking about thinking. We have evidence that good thinking is "actively open-minded", in a way that we will explain. People use actively open-minded thinking (AOT) to varying degrees, and most people will think this way in some situations and not others. In most situations, AOT helps thinkers get the best answer to the question they are thinking about.

First, we will teach you some concepts we use to talk about all kinds of thinking. Then we will explain the characteristics of AOT by using these concepts, and we will give you a chance to test your understanding. Finally, we will give you some problems to think about so that you can test your own thinking.

Part I: Thinking about beliefs

Thinking is what you do when you don't know what to do, what to believe, or what goal to pursue.

Thinking involves searching for possibilities, evidence, and goals. The possibilities are outcomes that could occur (or possible answers to the question at hand). The evidence is anything you think of that strengthens or weakens one answer or another. The goals are the standards or criteria that you use to evaluate the possibilities in the light of the evidence.

Thinking also involves making inferences from the things your search has found, that is, drawing conclusions from them. The conclusions may be revised several times during an episode of thinking. An "episode" can take anywhere from a few seconds to several years, and people can think as individuals or as a group.

When you think about what to believe, you have only a single goal, truth. Answers to the original question are possibilities. You can also make up a new question that will help you answer the original question, such as "What could possibly happen?" Then the answers to this question are also possibilities. The evidence consists of things that you think of or things that you find outside yourself, by asking people, by reading, and so on. Evidence can strengthen or weaken your belief in different possibilities.

Thinking about beliefs usually does not lead to a conclusion of which we are 100% certain. Most of the time, we want to establish a strength of belief. Often we want to know whether we believe something strongly enough to act on it, even if we are not certain. In some situations, you can assign a number to your strength of belief, from 0 to 100%, indicating your degree of certainty. 100% means that you are absolutely sure that the belief is true. 0% that you are sure it is false, and 50% that it is equally likely to be true or false. We sometimes interpret these numbers as probabilities. My probability that the sun will rise tomorrow is 100%. My probability that I will see tomorrow's sunrise is about 10%.
An exercise in classifying thoughts: What percentage of Americans have pets?

Here are some examples of a person thinking about the question. See if you can classify the statements as **possibilities** (a possible answer to the question), **evidence** (some fact or belief that makes some answers more or less likely), or statements that **search for possibilities or evidence** (a statement that something is being sought), or **conclusion** (a statement of the effect of evidence). Sometimes a statement can fall into more than one classification.

<table>
<thead>
<tr>
<th>Classify the statements</th>
<th>Possibility</th>
<th>Evidence</th>
<th>Conclusion</th>
<th>Search for Possibility</th>
<th>Search for Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>It could be a majority of Americans.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A lot of people I know have cats or dogs.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>But owning a pet is very costly.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I guess it is possible that not many Americans own pets.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many Americans work, so that would make pet ownership difficult.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A lot of people live in apartments or rent their living space, which puts restrictions to their pet ownership.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>But what about fish and bird owners?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>There is a large market for pet products.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>It is possible that half of the Americans own pets.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What else should I consider before making an estimate?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What other animals could be considered pets?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I guess I would say that 60% of Americans own pets.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Notice that the question was "What percentage of Americans has a pet?" This is an open-ended question that allows search for possibilities.

Below, you will first see how you classified each statement or question. Underneath that are the correct classification of the statements. Please compare your answers to the correct ones to see what you might have gotten wrong.

Your classifications are:

- It could be a majority of the Americans.
- A lot of people I know have cats or dogs.
- But owning a pet is very costly.
- I guess it is possible that not many Americans own pets.
- Many Americans work, so that would make pet ownership difficult.
- A lot of people live in apartments or don't own their houses, which puts restrictions to their pet ownership.
- But what about fish and bird owners?
- There is a large market for pet products.
- It is possible that half of the Americans own pets.
- What else should I consider before making an estimate?
- What other animals could be considered pets?
- I guess I would say that 60% of Americans own pets.

The correct answers are:

- It could be a majority of the Americans. **POSSIBILITY**
- A lot of people I know have cats or dogs. **EVIDENCE**
- But owning a pet is very costly. **EVIDENCE**
- I guess it is possible that not many Americans own pets. **POSSIBILITY**
- Many Americans work, so that would make pet ownership difficult. **EVIDENCE**
- A lot of people live in apartments or don't own their houses, which puts restrictions to their pet ownership. **EVIDENCE**
- But what about fish and bird owners? **EVIDENCE**
- There is a large market for pet products. **EVIDENCE**
- It is possible that half of the Americans own pets. **POSSIBILITY**
- What else should I consider before making an estimate? **SEARCH FOR POSSIBILITY**
- What other animals could be considered pets? **SEARCH FOR EVIDENCE**
- I guess I would say that 60% of Americans own pets. **CONCLUSION**
Part II: Myside Bias

Part II: Myside bias

Many of the mistakes that people make in thinking result from a bias toward whatever conclusion they initially favored, which we call "myside bias". This bias can operate in many ways:

- Failing to look for other possibilities.
- Looking only for reasons favoring the initial possibility.
- Failing to look for evidence against it.
- Finding evidence against the favorite but (partially or fully) ignoring it.
- When drawing conclusions, giving too much weight to evidence favoring the favorite.
- Or not giving enough weight to evidence against the favorite or for another possibility.

Here are some simple examples.
Which has a larger population, California or Canada?

Respondent 1 answers:

You've got to be kidding. Canada is a whole country. It must be larger. And I can think of many big cities in Canada: Montreal, Ottawa, Toronto, Vancouver. And it looks much bigger on a map. I'm 100% sure it is Canada.

Respondent 2 answers:

You've got to be kidding. Canada is a whole country. It must be larger. And I can think of many big cities in Canada: Montreal, Ottawa, Toronto, Vancouver. Hmm, but California has San Francisco, Los Angeles, and San Diego, for a start. And I bet these places like San Jose are just as big as Vancouver. And I think Los Angeles is huge. It is a big metropolitan area with freeways and everything. But Canada looks much bigger on a map. But part of the reason it looks bigger is that it is farther north and things there get stretched out on a map. And most of it is uninhabited. But a lot of California is uninhabited too. I'm not sure. I'd say 50/50.

Notice that Respondent 2 has thought of reasons on the other side. Respondent 1 exhibits an example of myside bias of an extreme sort. Respondent 2 is an example of AOT.

Response 2 is an example of actively open-minded thinking (AOT), because the respondent looks actively for reasons against her initial conclusion and then takes them seriously, without dismissing them. If Respondent 2 ended by saying, "I'm still sure it is Canada", then this would be a clear example of under-weighing evidence against the favorite, when drawing a conclusion.
Here is another pair of examples. The question is which of the following three languages has more speakers (including those for whom it is a second language): Russian, Portuguese, Indonesian?

Respondent 1 answers:
• Russia is a huge country.
• Many people in the former Soviet Union also speak Russian even if it isn't their first language.
• Portugal is a small country.
• Russia is bigger than Indonesia. I will say Russia with high confidence.

Respondent 2 answers:
• Russia is a huge country.
• Many people in the former Soviet Union also speak Russian even if it isn't their first language.
• Portugal is a small country.
• But they speak Portuguese in Brazil, which is also a huge country.
• Russia is bigger than Indonesia.
• But most of Russia is Siberia. Maybe its population is not so big. I'm not sure. It could be any of the three. I would guess Portuguese, but not with much confidence.

Again, if Respondent 2 ended concluding "Russia with high confidence", then the thinker would be largely neglecting the evidence on the other side.

Another type of error, not illustrated here, is failing to search for evidence outside the task, for example, on the World Wide Web, or from other people. We ask you not to do that in our experiments, but in a real situation you would look for information outside of what you know yourself.

People who use AOT have better probability judgments. Those with myside bias sometimes give very high probabilities to statements that turn out to be false, or very low probabilities to statements that turn out to be true. Those with AOT do not make these mistakes so often, but they still give high probabilities to true statements and low probabilities to false ones.

Part III: Intro to Test Your Own Thinking

Part III: Test Your Own Thinking

On the next few pages we will give you 6 problems to think about. The questions topics include country population, movie release dates, and socio-economic facts. We ask you to list your thoughts, entering them in a box. Then you give your best guess about the answer and assign probabilities to each of the options. You have the options to skip 3 items, but please make sure that you answer at least 3 items.
Part III: Q1

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 6 seconds
Click Count: 0 clicks

Which of the three following countries has the highest population?
Vietnam
Germany
Iran

Answer the question to the best of your abilities. Please don't look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.

Which of the three following countries has the highest population?

- Vietnam
- Germany
- Iran
- Skip question

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iran</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Part III: Q2

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three following countries has the highest population?

Indonesia
Brazil
Pakistan

Answer the question to the best of your abilities. Please don't look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.

Which of the following countries has the highest population?

- Indonesia
- Brazil
- Pakistan
- Skip question

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indonesia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pakistan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Part III: Q3

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three following countries has the highest population?

Russia
Nigeria
Bangladesh

Answer the question to the best of your abilities. Please don't look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.

☐ Russia
☐ Nigeria
☐ Bangladesh
☐ Skip question

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nigeria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangladesh</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Part III: Explain

Here is an example of AOT for the first item: Vietnam, Germany, Iran. (The answer is Vietnam.)
- I know that Germany is the largest country in Europe. It must be large. [Evidence for Germany]
- Germany has many famous cities, such as Berlin, Munich, Frankfurt, Bonn, Dresden, Leipzig, and Hamburg. [More evidence for Germany]
- Vietnam is narrow. Evidence against Vietnam.
- Vietnam is along the coast, and many people live along the coast. Evidence for Vietnam.
- It has the Mekong delta. Evidence for Vietnam.
- Iran is very large, probably as large as Germany. Evidence for Iran.
- I'm not sure. I will say Germany with a probability of 40%, and 35% for Vietnam, and 25% for Iran.

Of course, my side bias could look like this:
- I know that Germany is the largest country in Europe. It must be large.
- Germany has many famous cities, such as Berlin, Munich, Frankfurt, Bonn, Dresden, Leipzig, and Hamburg.
- Vietnam is narrow.
- I will conclude Germany with high probability.

Part III: Q4

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three movies has the latest release date?
Gone with the Wind
The Godfather
Twelve Angry Men

Answer the question to the best of your abilities. Please don't look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.
Part II: Q5

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gone with the Wind</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Godfather</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Twelve Angry Men</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Part III: Q5

Which of the three movies has the latest release date?

- Casablanca
- Singin' in the Rain
- Breakfast at Tiffany's

Answer the question to the best of your abilities. Please don't look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.
What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casablanca</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singin' in the Rain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breakfast at Tiffany's</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Part III: Q6

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three movies has the latest release date?

The Jurassic Park
Jaws
The Terminator

Answer the question to the best of your abilities. Please don’t look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.
Part II: Q1

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Jurassic Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jaws</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Terminator</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Part III: R1

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which country has the largest population

Between these options: $(q:/QID21/ChoiceGroup/AllChoices?displayLogic=0),
you picked: $(q:/QID21/ChoiceGroup/SelectedChoices),
based on your notes:

$(q:/QID38/Choice?TextEntryValue)

The correct answer is: Vietnam

Which category did your responses fall in:

- [ ] I failed to look hard enough for counter-evidence (I now see that there are arguments that I didn’t think of)
- [ ] I looked for counter-evidence but didn’t find any (This is not an example of my bias, because you tried)
- [ ] Under-weighted counter-evidence (I thought of evidence against my favorite but did NOT take it seriously enough)
- [ ] Over-weighted evidence (I took evidence for my favorite too seriously)
- [ ] I knew it (I already knew the answer so I did NOT have to think)
- [ ] Good AOT (I thought of counter-evidence and used it appropriately)
Part III: R2

These page timer metrics will not be displayed to the recipient.
Field Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which country has the largest population

Between these options: $\{q://QID87/ChoiceGroup/AllChoices?displayLogic=0\}$, you picked: $\{q://QID87/ChoiceGroup/SelectedChoice\}$, based on your notes:

$\{q://QID88/ChoiceTextEntryValue\}$

The correct answer is: Indonesia

Which category did your responses fall in:

- Failed to look hard enough for counter-evidence (I now see that there are arguments that I didn’t think of)
- Looked for counter-evidence but didn’t find any (This is not an example of my side bias, because you tried)
- Under-weighted counter-evidence (I thought of evidence against my favorite but did NOT take it seriously enough)
- Over-weighted evidence (I took evidence for my favorite too seriously)
- Knew it (I already knew the answer so I did NOT have to think)
- Good AOT (I thought of counter-evidence and used it appropriately)
Part III: R3

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which country has the largest population

Between these options: ${q://QID91/ChoiceGroup/AllChoices?displayLogic=0},
you picked: ${q://QID91/ChoiceGroup/SelectedChoices},

based on your notes:

${q://QID90/ChoiceTextEntryValue}

The correct answer is: Nigeria

Which category did your responses fall in:

- Failed to look hard enough for counter-evidence (I now see that there are arguments that I didn’t think of)
- Looked for counter-evidence but didn’t find any (This is not an example of my side bias, because you tried)
- Under-weighted counter-evidence (I thought of evidence against my favorite but did NOT take it seriously enough)
- Over-weighted evidence (I took evidence for my favorite too seriously)
- Knew it (I already knew the answer so I did NOT have to think)
- Good ADT (I thought of counter-evidence and used it appropriately)
Part III: R4

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three movies has the latest release date?

Between these options: [QID95/ChoiceGroup/AllChoices?displayLogic=0], you picked: [QID95/ChoiceGroup/SelectedChoices].

Based on your notes:
[QID94/ChoiceTextEntryValue]

The correct answer is: The Godfather

Which category did your responses fall in:

- Failed to look hard enough for counter-evidence (I now see that there are arguments that I didn’t think of)
- Looked for counter-evidence but didn’t find any (This is not an example of my side bias, because you tried)
- Under-weighted counter-evidence (I thought of evidence against my favorite but did NOT take it seriously enough)
- Over-weighted evidence (I looked evidence for my favorite too seriously)
- Knew it (I already knew the answer so I did NOT have to think)
- Good ADT (I thought of counter-evidence and used it appropriately)
Part III: R5

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three movies has the latest release date?

Between these options: $[q://QID148/ChoiceGroup/AllChoices?displayLogic=0], you picked: $[q://QID148/ChoiceGroup/SelectedChoices], based on your notes:

$[q://QID145/ChoiceTextEntryValue]

The correct answer is: Breakfast at Tiffany's

Which category did your responses fall in:

- Failed to look hard enough for counter-evidence (I now see that there are arguments that I didn't think of)
- Looked for counter-evidence but didn't find any (This is not an example of my side bias, because you tried)
- Under-weighted counter-evidence (I thought of evidence against my favorite but did NOT take it seriously enough)
- Over-weighted evidence (I took evidence for my favorite too seriously)
- Knew it (I already knew the answer so I did NOT have to think)
- Good ADT (I thought of counter-evidence and used it appropriately)
Part III: R6

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 6 seconds
Click Count: 0 clicks

Which of the three movies has the latest release date?

Between these options: \$\{q:/QID120/ChoiceGroup/AllChoices?displayLogic=0\},
you picked: \$\{q:/QID120/ChoiceGroup/SelectedChoices\},

based on your notes:
\$\{q:/QID119/ChoiceTextEntryValue\}

The correct answer is: The Jurassic Park

Which category did your responses fall in:

- Failed to look hard enough for counter-evidence (I now see that there are arguments that I didn't think of)
- Looked for counter-evidence but didn't find any (This is not an example of my side bias, because you tried)
- Under-weighted counter-evidence (I thought of evidence against my favorite but did NOT take it seriously enough)
- Over-weighted evidence (I took evidence for my favorite too seriously)
- Knew it (I already knew the answer so I did NOT have to think)
- Good AOT (I thought of counter-evidence and used it appropriately)
Feedback

Thank you for taking your time to learn about AOT

We would like to know what you think of the experience.

Please indicate how much you agree or disagree with the following statements.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOT will help me make better probability judgments</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>AOT is an easy concept to understand</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>AOT is an easy concept to apply</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Learning about AOT was a waste of my time</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I often used AOT before this training session</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I will apply AOT in the future</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I have no intention to use AOT</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

Do you have any comments for us?

[Blank space for comments]
APPENDIX D-2: TRAINING MODULE FOR STUDY 7

The following pages include the training experimental condition subjects.
Introduction

These page timer metrics will not be displayed to the recipient.
First Click: 111.175 seconds
Last Click: 111.175 seconds
#QuestionText, TimingPageSubmit#: 0 seconds
#QuestionText, TimingClickCount#: 1 clicks

Welcome to the training part of our study. In this section we will train you in good thinking principles and give you some exercises to apply these principles. When answering questions, be honest so that you can benefit the most from this training. Please read carefully as you will be asked questions regarding the information given to you. To receive payment, you need to follow all the instructions, so please take your time and don’t rush through the training.

Before you start the training enter your email address and then press NEXT. (Please use the one with which you registered.)

These page timer metrics will not be displayed to the recipient.
First Click: 111.174 seconds
Last Click: 111.174 seconds
#QuestionText, TimingPageSubmit#: 0 seconds
#QuestionText, TimingClickCount#: 1 clicks

What percentage of Americans have pets? (Please put in a number between 0 and 100.)

Please tell us how you came up with your answer. Write as many reasons as you can.

Please answer the question again. You may change your answer from your initial one or put in the same answer.
What percentage of Americans have pets? (Please put in a number between 0 and 100.)
Which of the following three languages has more speakers (including those for whom it is a second language)?

- [ ] Russian
- [ ] Portuguese
- [ ] Indonesian

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian</td>
<td>0 %</td>
</tr>
<tr>
<td>Portuguese</td>
<td>0 %</td>
</tr>
<tr>
<td>Indonesian</td>
<td>0 %</td>
</tr>
<tr>
<td>Total</td>
<td>0 %</td>
</tr>
</tbody>
</table>

Please tell us how you came up with your answer. Write as many reasons as you can.


Please answer the question again. You may change your answer from your initial one or put in the same answer.

Which of the following three languages has more speakers (including those for whom it is a second language)?

- [ ] Russian
- [ ] Portuguese
- [ ] Indonesian

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian</td>
<td>0 %</td>
</tr>
<tr>
<td>Portuguese</td>
<td>0 %</td>
</tr>
<tr>
<td>Indonesian</td>
<td>0 %</td>
</tr>
<tr>
<td>Total</td>
<td>0 %</td>
</tr>
</tbody>
</table>
Training in actively open-minded thinking

The purpose of these exercises is to teach you a way of thinking about thinking. We have evidence that good thinking is "actively open-minded", in a way that we will explain. People use actively open-minded thinking (AOT) to varying degrees, and most people will think this way in some situations and not others. In most situations, AOT helps thinkers get the best answer to the question they are thinking about.

In this training we will explain the characteristics of AOT, and we will give you a chance to test your understanding. Finally, we will give you some problems to think about so that you can test your own thinking.

Part 1: Thinking about beliefs

Thinking is what you do when you don’t know what to do, what to believe, or what goal to pursue.

When you think about what to believe, you have only a single goal, truth. Answers to the original question are possibilities. You can also make up a new question that will help you answer the original question, such as "What could possibly happen?" Then the answers to this question are also possibilities. The evidence consists of things that you think of or things that you find outside yourself, by asking people, by reading, and so on. Evidence can strengthen or weaken your belief in different possibilities.

Thinking about beliefs usually does not lead to a conclusion of which we are 100% certain. Most of the time, we want to establish a strength of belief. Often we want to know whether we believe something strongly enough to act on it, even if we are not certain. In some situations, you can assign a number to your strength of belief, from 0 to 100%, indicating your degree of certainty. 100% means that you are absolutely sure that the belief is true, 0% that you are sure it is false, and 50% that it is equally likely to be true or false. We sometimes interpret these numbers as probabilities. My probability that the sun will rise tomorrow is 100%. My probability that I will see tomorrow's sunrise is about 10%.

Actively Open-Minded Thinking

We said earlier that good thinking is actively open-minded thinking. What do we mean by that?

Actively open-minded thinking consists of:

- Search for possibilities and evidence that is thorough enough for the task at hand
- Fair judgment of other possibilities and evidence other than the ones we initially favor
- Confidence that is assigned appropriately based on the quantity and the quality of thinking done.

AOT helps us correct mistakes before we make them. It also helps to make good judgments about probabilities. With AOT, we believe more strongly in true conclusions and less strongly in false ones.
Part II: Myside Bias

These page timer metrics will not be displayed to the recipient.
First Click: 111.173 seconds
Last Click: 111.173 seconds
#QuestionText, TimingPageSubmit#: 0 seconds
#QuestionText, TimingClickCount#: 1 clicks

Part II: Myside bias

Many of the mistakes that people make in thinking result from a bias toward whatever conclusion they initially favored, which we call "myside bias". It is a bias toward whatever conclusion you favor at the moment, your pet conclusion. This bias can operate in many ways:

- Failing to look for other possibilities
- Looking only for reasons favoring your pet conclusion
- Failing to look for evidence against it
- Finding evidence against your pet conclusion but (partially or fully) ignoring it.
- When drawing conclusions, giving too much weight to evidence for your pet conclusion, or too little weight to evidence against it or for something else

These page timer metrics will not be displayed to the recipient.
First Click: 111.173 seconds
Last Click: 111.173 seconds
#QuestionText, TimingPageSubmit#: 0 seconds
#QuestionText, TimingClickCount#: 1 clicks

Next we will take a look at the two questions you answered in the beginning of the training. You will see how two different individuals responded to the same questions you saw. We will also show you your responses so that you can self-evaluate your responses.

The first question you saw was:
What percentage of Americans have pets?

Respondent 1 answers:
My answer is 40%. Owning pets is very costly, and time consuming. A lot of Americans work, so many people don't have time to care for a dog. Also a lot of people either live in apartments or rent their living space, which puts a lot of restrictions on pet ownership. I'm 100% confident that my answer is correct.

Respondent 2 answers:
My answer is 40%. Owning pets is very costly, and time consuming. A lot of Americans work, so many people don't have time to care for a dog. Also a lot of people either live in apartments or rent their living space, which puts a lot of restrictions on pet ownership. Hmm, but these restrictions are usually for dogs and cats. What about people who own fish and birds? And those pets are not as costly and time consuming as owning a dog or a cat is. There might also be other kinds of pets I'm not considering. There is a big market for pet products and services, so there must be many people using them. I'm no longer sure that it is as low as 40%, but I still think that it can't be so much higher. So I would say 50% of Americans own pets, and I am 75% confident that my answer is correct.

Notice that Respondent 2 has thought of reasons on the other side. Respondent 1 exhibits an example of myside bias of an extreme sort. Respondent 2 is an example of AOT.

Response 2 is an example of actively open-minded thinking (AOT), because the respondent looks actively for reasons against her initial conclusion and then takes them seriously, without dismissing them. She is also able to adjust their initial answer in the face of counteracting evidence.

If Respondent 2 ended by saying, "I'm still sure it is 40%"., then this would be a clear example of under-weighing evidence against their favored answer, when drawing a conclusion.
Now let's look at your response to the same question. Your answer to the question was $(q_i/QID207/ChoiceTextEntryValue)$ based on the following reasons:

$(q_i/QID191/ChoiceTextEntryValue)$

The correct answer is: 68%

Looking at your reasoning, would you say that your thinking was more like Respondent 1 or Respondent 2?

- More like Respondent 1
- More like Respondent 2

Take a look at your reasoning again. Which categories did your reasoning fail in? (Select all that apply.)

- Failed to look hard enough for counter-evidence (I now see that there are arguments I didn't think of)
- Looked for counter-evidence but didn't find any (This is not an example of myside bias because you tried)
- Under-weighted counter-evidence (I thought of evidence against my answer but did not take it seriously enough)
- Over-weighted evidence (I took evidence for my answer too seriously)
- Knew it (I already knew the answer so I did NOT have to think)
- Good AOT (I thought of counter-evidence and used it appropriately)

These page timer metrics will not be displayed to the recipient.
First Click: 111.172 seconds
Last Click: 111.172 seconds
#QuestionText, TimingPageSubmit#: 0 seconds
#QuestionText, TimingClickCount#: 1 clicks

Here is the second question you saw.

Which of the following three languages has more speakers (including those for whom it is a second language):

- Russian, Portuguese, Indonesian?

Respondent 1 answers:
- Russia is a huge country.
- Many people in the former Soviet Union also speak Russian even if it isn't their first language.
- Portugal is a small country.
- Russia is bigger than Indonesia. I will say Russia with high confidence.

Respondent 2 answers:
- Russia is a huge country.
- Many people in the former Soviet Union also speak Russian even if it isn't their first language.
- Portugal is a small country.
- But they speak Portuguese in Brazil, which is also a huge country.
- Russia is bigger than Indonesia. Maybe its population is not so big. I'm not sure. It could be any of the three. I would guess Portuguese, but not with much confidence.

Again, if Respondent 2 ended concluding "Russia with high confidence", then the thinker would be largely neglecting the evidence on the other side.

Another type of error, not illustrated here, is failing to search for evidence outside the task, for example, on the World Wide Web, or from other people. We ask you not to do that in our experiments, but in a real situation you would look for information outside of what you know yourself.

People who use AOT have better probability judgments. Those with myside bias sometimes give very high probabilities to statements that turn out to be false, or very low probabilities to statements that turn out to be true. Those with AOT do not make these mistakes so often, but they still give high probabilities to true statements and low probabilities to false ones.
Now let's look at your response to the same question. Your answer to the question was \( q_{208 + ChoiceGroup/SelectedChoices} \) based on the following reasons:

\( q_{103 + ChoiceTextEntryValue} \)

The correct answer is: **Russian**

Looking at your reasoning, would you say that your thinking was more like Respondent 1 or Respondent 2?

- More like Respondent 1
- More like Respondent 2

Take a look at your reasoning again. Which categories did your reasoning fall in? (Select all that apply.)

- Failed to look hard enough for counter-evidence (I now see that there are arguments I didn't think of)
- Looked for counter-evidence but didn't find any (This is not an example of myside bias because you tried)
- Under-weighed counter-evidence (I thought of evidence against my answer but did not take it seriously enough)
- Over-weighed evidence (I took evidence for my answer too seriously)
- Knew it (I already knew the answer so I did NOT have to think)
- Good AOT (I thought of counter-evidence and used it appropriately)

**Review Test**

**These page timer metrics will not be displayed to the recipient.**

First Click: 111.172 seconds
Last Click: 111.172 seconds

#QuestionText, TimingPageSubmit#: 0 seconds
#QuestionText, TimingClickCount#: 1 clicks

Before moving on to the exercise questions, let's review what we have learned. You will be asked 3 multiple choice questions testing the concepts of actively open-minded thinking and myside bias. After answering each question, you will see whether you answered the question correctly or not.

**These page timer metrics will not be displayed to the recipient.**

First Click: 111.172 seconds
Last Click: 111.172 seconds

#QuestionText, TimingPageSubmit#: 0 seconds
#QuestionText, TimingClickCount#: 1 clicks

Which of the following is NOT an example of AOT?

- Looking for possibilities before making a decision
- Making appropriate probability judgments to possibilities and evidence based on the strength of arguments
- Thinking of evidence against one's pet beliefs but not taking it seriously
- Challenging one's pet beliefs
- Evaluating evidence supporting one's pet beliefs appropriately
Which of the following is NOT an example of myside bias?

- Dismissing evidence that goes against one's pet beliefs
- Trying to look for counter-evidence but not finding any
- Looking for evidence that supports only one's pet beliefs
- Paying too much attention to the evidence that supports one's pet beliefs
- Failing to look for alternative possibilities

Sam is a rising junior in college and needs to declare a major, but she is undecided between Psychology and Chemistry. She has been thinking of reasons to help her make this decision. Here is one of them:

"I am interested in learning about how the mind works, but I also enjoy my lab courses in chemistry."

How would this reasoning affect Sam's decision?

- It should make her want to become a Psychology major more.
- It should make her want to become a Chemistry major more.
- It should not have any clear effect on her decision.

For the question:

Which of the following is NOT an example of AOT?

with the following answer options

$\{eq://QID199ChoiceGroup/AllChoices?displayLogic=0|\}$

You picked $\{eq://QID199ChoiceGroup/SelectedChoices\}$.  

The correct answer is: Thinking of evidence against one's pet beliefs but not taking it seriously

This behavior is not an example of AOT, because evidence against one's pet beliefs is useless unless it is taken seriously. This behavior is an example of myside bias.
Which of the following is NOT an example of myside bias?

with the following answer options

You picked $[q://QID200/ChoiceGroup/SelectedChoices]$. The correct answer is: **Trying to look for counter-evidence but not finding any**

This behavior is not an example of myside bias, because the person tried to find counter-evidence, yet was having difficulty coming up with any. This might happen when one does not have enough knowledge or the means to gain more knowledge about a certain topic.

Sam is a rising junior in college and needs to declare a major, but she is undecided between Psychology and Chemistry. She has been thinking of reasons to help her make this decision. Here is one them:

"I am interested in learning about how the mind works, but I also enjoy my lab courses in chemistry."

How should this reasoning affect Sam's decision?

with the following answer options

You picked $[q://QID201/ChoiceGroup/SelectedChoices]$. The correct answer is: **It should not have any clear effect on her decision.**

The evidence presented here is not enough for Sam to make a decision one way or the other. It shows that Sam has interest in both fields. This kind of evidence is called **neutral evidence**. Since this evidence does not support one choice or the other but presents support for both, Sam cannot use this piece of evidence to reach a conclusion.
You answered $\text{SC_6TYAEzdZbf57Bcx/Score}$ out of 3 questions correct. Here is a quick review of the concepts we have learned before you move onto the next section of the training.

**Actively Open Minded Thinking:**

AOT consists of searching for possibilities and evidence, and evaluating these fairly even if they go against one’s initial beliefs. People who use AOT make appropriate probability judgments such that they assign high probabilities to true statements but low probabilities to false ones.

**My-side Bias:**

Many of the errors in judgment result from a bias toward their pet beliefs, which we call my-side bias. People who display this bias neglect to look for other possibilities, or look only for reasons that favor their pet beliefs and not look for any evidence against it. They might either find evidence against their pet belief but partially or fully ignore it or might give too much weight to evidence that favor their pet beliefs.

Part III: Intro to Test Your Own Thinking

These page timer metrics will not be displayed to the recipient.

First Click: 111.169 seconds
Last Click: 111.169 seconds
#QuestionText, TimingPageSubmit#: 0 seconds
#QuestionText, TimingClickCount#: 1 clicks

Part III: Test Your Own Thinking

On the next few pages we will give you 6 problems to think about. The first 3 problems are policy problems. We ask that you answer at least 2 of these questions.

The last 3 problems are about country population, movie release dates, and socio-economic facts. You must answer all 3 questions.

You will be asked to list your thoughts, entering them in the provided space. Then you will need to choose an answer and present arguments to support your favored answer or others. For the last 3 questions you will also assign probabilities to each of the options.

Before you move on, please pick which 2 of the 3 policy problems you would like to answer.

1. How should a government handle vaccination requirements for its citizens?
2. How should government agencies handle the manufacturing and distribution of nutritional supplements and vitamins?
3. How should the U.S. government handle subsidies regarding corn production?

- 1 and 2
- 1 and 3
- 2 and 3
These page timer metrics will not be displayed to the recipient.
First Click: 111.169 seconds
Last Click: 111.169 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

How should a government handle vaccination requirements for its citizens?
Below propose at least 3 different solutions to the problem above.

Solution 1
Solution 2
Solution 3
Solution 4
Solution 5

Please take a look at the solutions you listed. Which solution is your preferred solution?

☐ Solution 1
☐ Solution 2
☐ Solution 3
☐ Solution 4
☐ Solution 5

Below list the arguments for or against your favored solution. You do not have to use all the space but must list at least 2 arguments.

Argument 1
Argument 2
Argument 3
Argument 4
Argument 5
Argument 6
Argument 7
Argument 8

Now please tell us whether the arguments you listed above are FOR or AGAINST your favored solution or neither.

<table>
<thead>
<tr>
<th>Argument</th>
<th>For my favored solution</th>
<th>Against my favored solution</th>
<th>Neither for nor against my favored solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argument 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Please take a look at the solutions you listed above. Which solution is your preferred solution?

- Solution 1
- Solution 2
- Solution 3
- Solution 4
- Solution 5

Part III: Q3

These page timer metrics will not be displayed to the recipient.
First Click: 111.159 seconds
Last Click: 111.159 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

How should government agencies handle the manufacturing and distribution of nutritional supplements and vitamins?
Below propose at least 3 different solutions to the problem above.

Solution 1
Solution 2
Solution 3
Solution 4
Solution 5

Please take a look at the solutions you listed. Which solution is your preferred solution?

- Solution 1
- Solution 2
- Solution 3
- Solution 4
- Solution 5

Below list the arguments for or against your favored solution. You do not have to use all the space but must list at least 2 arguments.

Argument 1
Argument 2
Argument 3
Argument 4
Argument 5
Argument 6
Argument 7
Argument 8
Now please tell us whether the arguments you listed above are FOR or AGAINST your favored solution or neither.

<table>
<thead>
<tr>
<th>Argument 1</th>
<th>For my favored solution</th>
<th>Against my favored solution</th>
<th>Neither for nor against my favored solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg 8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please take a look at the solutions you listed above. Which solution is your preferred solution?

- Solution 1
- Solution 2
- Solution 3
- Solution 4
- Solution 5

Part II: Q3

These page timer metrics will not be displayed to the recipient.
First Click: 111.168 seconds
Last Click: 111.168 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

How should the U.S. government handle subsidies regarding corn production?
Below propose at least 3 different solutions to the problem above.

| Solution 1 |                      | Solution 2 |                      | Solution 3 |                      | Solution 4 |                      | Solution 5 |

Please take a look at the solutions you listed. Which solution is your preferred solution?

- Solution 1
- Solution 2
- Solution 3
- Solution 4
- Solution 5
Below list the arguments for or against your favored solution. You do not have to use all the space but must list at least 2 arguments.

| Argument 1 |  |
| Argument 2 |  |
| Argument 3 |  |
| Argument 4 |  |
| Argument 5 |  |
| Argument 6 |  |
| Argument 7 |  |
| Argument 8 |  |

Now please tell us whether the arguments you listed above are FOR or AGAINST your favored solution or neither.

<table>
<thead>
<tr>
<th>Argument 1</th>
<th>For my favored solution</th>
<th>Against my favored solution</th>
<th>Neither for nor against my favored solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argument 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argument 8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please take a look at the solutions you listed above. Which solution is your preferred solution?

- [ ] Solution 1
- [ ] Solution 2
- [ ] Solution 3
- [ ] Solution 4
- [ ] Solution 5
These page timer metrics will not be displayed to the recipient.
First Click: 111.167 seconds
Last Click: 111.167 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

Which of the following countries has the most forested area in square miles?
- Brazil
- Canada
- Russia

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%)
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Canada</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Russia</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Below list the reasons for or against your favored answer or any other related evidence. You do not have to use all the space but must list at least 2 reasons. We will later show your notes, response and the correct answer.

<table>
<thead>
<tr>
<th>Reason</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

Now please tell us whether the arguments you listed above are FOR or AGAINST your favored solution or neither.

<table>
<thead>
<tr>
<th>Reason</th>
<th>For my favored answer</th>
<th>Against my favored answer</th>
<th>Neither for nor against my favored answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>●</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>●</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>●</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>●</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>●</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>●</td>
<td>●</td>
<td></td>
</tr>
</tbody>
</table>
Please answer the question again. You may change your answer or put in the same one as before.

Which of the following countries has the most forested area in square miles?

- Brazil
- Canada
- Russia

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%)

<table>
<thead>
<tr>
<th>Brazil</th>
<th>0 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>0 %</td>
</tr>
<tr>
<td>Russia</td>
<td>0 %</td>
</tr>
<tr>
<td>Total</td>
<td>a  %</td>
</tr>
</tbody>
</table>

Part III: Q5

These page timer metrics will not be displayed to the recipient.
First Click: 111.167 seconds
Last Click: 111.167 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

Which of the U.S. metropolitan areas below has the highest percentage of gay population?

- New York, NY
- Austin, TX
- Los Angeles, CA

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%)

<table>
<thead>
<tr>
<th>New York, NY</th>
<th>0 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austin, TX</td>
<td>0 %</td>
</tr>
<tr>
<td>Los Angeles, CA</td>
<td>0 %</td>
</tr>
<tr>
<td>Total</td>
<td>a  %</td>
</tr>
</tbody>
</table>

Below list the reasons for or against your favored answer or any other related evidence. You do not have to use all the space but must list at least 2 reasons. We will later show your notes, response and the correct answer.

Reason 1
Reason 2
Reason 3
Reason 4
Reason 5
Reason 6
Reason 7
Reason 8
Now please tell us whether the arguments you listed above are FOR or AGAINST your favored solution or neither.

<table>
<thead>
<tr>
<th>Reason</th>
<th>For my favored answer</th>
<th>Against my favored answer</th>
<th>Neither for nor against my favored answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reason 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reason 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reason 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reason 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reason 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reason 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reason 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reason 8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please answer the question again. You may change your answer or put in the same one as before.

Which of the U.S. metropolitan areas below has the highest percentage of gay population?

- [ ] New York, NY
- [ ] Austin, TX
- [ ] Los Angeles, CA

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York, NY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austin, TX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Los Angeles, CA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Part III: Q8

These page timer metrics will not be displayed to the recipient.

First Click: 111.167 seconds
Last Click: 111.167 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

According to a 2013 report by World Health Organization, which of the following countries has the highest life expectancy at birth?

- [ ] The United States of America
- [ ] Sweden
- [ ] Italy

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>The U.S.A.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sweden</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Below list the reasons for or against your favored answer or any other related evidence. You do not have to use all the space but must list at least 2 reasons. We will later show your notes, response and the correct answer.

Reason 1
Reason 2
Reason 3
Reason 4
Reason 5
Reason 6
Reason 7
Reason 8

Now please tell us whether the arguments you listed above are FOR or AGAINST your favored solution or neither.

<table>
<thead>
<tr>
<th>Reason</th>
<th>For my favored answer</th>
<th>Against my favored answer</th>
<th>Neither for nor against my favored answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reason 1</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Reason 2</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Reason 3</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Reason 4</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Reason 5</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Reason 6</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Reason 7</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Reason 8</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Please answer the question again. You may change your answer or put in the same one as before.

According to a 2013 report by World Health Organization, which of the following countries has the highest life expectancy at birth?

- ☐ The United States of America
- ☐ Sweden
- ☐ Italy

What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

| The U.S.A. | 0 %
| Sweden     | 0 %
| Italy      | 0 %
| Total      | 0 % |
These page timer metrics will not be displayed to the recipient.
First Click: 111.167 seconds
Last Click: 111.167 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

How should a government handle vaccination requirements for its citizens?
Below propose at least 3 different solutions to the problem above.

You listed the following solutions:
Solution 1: $(q://QID85/ChoiceTextEntryValue/1)$
Solution 2: $(q://QID85/ChoiceTextEntryValue/2)$
Solution 3: $(q://QID85/ChoiceTextEntryValue/3)$
Solution 4: $(q://QID85/ChoiceTextEntryValue/4)$
Solution 5: $(q://QID85/ChoiceTextEntryValue/5)$

Your answer was $(q://QID249/ChoiceGroup/SelectedChoices)$. You gave the following arguments:
$(q://QID251/ChoiceTextEntryValue/1)$
$(q://QID251/ChoiceTextEntryValue/2)$
$(q://QID251/ChoiceTextEntryValue/3)$
$(q://QID251/ChoiceTextEntryValue/4)$
$(q://QID251/ChoiceTextEntryValue/5)$
$(q://QID251/ChoiceTextEntryValue/6)$
$(q://QID251/ChoiceTextEntryValue/7)$

Take a look at your solutions and arguments again. Which categories did they fall in? (Select all that apply.)
- Failed to list alternative solutions (I see that I listed solutions that are too similar)
- Failed to look hard enough for counter-evidence (I now see that there are arguments I didn't think of)
- Looked for counter-evidence but didn't find any (This is not an example of my side bias because you tried)
- My reasons for the view I do no support are not as strong as those that support my view (I now see that I chose weaker arguments for the opposing view)
- Good AOT (I thought of counter-evidence and used it appropriately)
These page timer metrics will not be displayed to the recipient.
First Click: 111.166 seconds
Last Click: 111.166 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

How should government agencies handle the manufacturing and distribution of nutritional supplements and vitamins?

Below propose at least 3 different solutions to the problem above.

You listed the following solutions:

Solution 1: $(q/QID93/ChoiceTextEntryValue/1)$
Solution 2: $(q/QID93/ChoiceTextEntryValue/2)$
Solution 3: $(q/QID93/ChoiceTextEntryValue/3)$
Solution 4: $(q/QID93/ChoiceTextEntryValue/4)$
Solution 5: $(q/QID93/ChoiceTextEntryValue/5)$

Your answer was $(q/QID253/ChoiceGroup/SelectedChoices)$.

You gave the following arguments:

$(q/QID255/ChoiceTextEntryValue/1)$
$(q/QID255/ChoiceTextEntryValue/2)$
$(q/QID255/ChoiceTextEntryValue/3)$
$(q/QID255/ChoiceTextEntryValue/4)$
$(q/QID255/ChoiceTextEntryValue/5)$
$(q/QID255/ChoiceTextEntryValue/6)$

Take a look at your solutions and arguments again. Which categories did they fall in? (Select all that apply.)

- Failed to list alternative solutions (I see that I listed solutions that are too similar)
- Failed to look hard enough for counter-evidence (I now see that there are arguments I didn’t think of)
- Looked for counter-evidence but didn’t find any (This is not an example of myside bias because you tried)
- My reasons for the view I do no support are not as strong as those that support my view (I now see that I chose weaker arguments for the opposing view)
- Good AOT (I thought of counter-evidence and used it appropriately)
How should the U.S. government handle subsidies regarding corn production?

Below propose at least 3 different solutions to the problem above.

You listed the following solutions:

Solution 1: $(q://QID257/ChoiceTextEntryValue1$)
Solution 2: $(q://QID257/ChoiceTextEntryValue2$)
Solution 3: $(q://QID257/ChoiceTextEntryValue3$)
Solution 4: $(q://QID257/ChoiceTextEntryValue4$)
Solution 5: $(q://QID257/ChoiceTextEntryValue5$)

Your answer was $(q://QID258/ChoiceGroup/SelectedChoices)$. 

You gave the following arguments:

$(q://QID259/ChoiceTextEntryValue1$)
$(q://QID259/ChoiceTextEntryValue2$)
$(q://QID259/ChoiceTextEntryValue3$)
$(q://QID259/ChoiceTextEntryValue4$)
$(q://QID259/ChoiceTextEntryValue5$)
$(q://QID259/ChoiceTextEntryValue6$)
$(q://QID259/ChoiceTextEntryValue7$)
$(q://QID259/ChoiceTextEntryValue8$)

Take a look at your solutions and arguments again. Which categories did they fall into? (Select all that apply.)

- Failed to list alternative solutions (I see that I listed solutions that are too similar)
- Failed to look hard enough for counter-evidence (I now see that there are arguments I didn’t think of)
- Looked for counter-evidence but didn’t find any (This is not an example of my side bias because you tried)
- My reasons for the view I do no support are not as strong as those that support my view (I now see that I chose weaker arguments for the opposing view)
- Good AOT (I thought of counter-evidence and used it appropriately)
Part III: R4

These page timer metrics will not be displayed to the recipient.
First Click: 111.166 seconds
Last Click: 111.166 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

Which of the following countries has the most forested area in square miles?
Between these options, $(q://QID39/ChoiceGroup/AllChoices?displayLogic=0)$, you picked: $(q://QID21/ChoiceGroup/SelectedChoices)$.

based on your notes:
$(q://QID303/ChoiceTextEntryValue/1)$
$(q://QID303/ChoiceTextEntryValue/2)$
$(q://QID303/ChoiceTextEntryValue/3)$
$(q://QID303/ChoiceTextEntryValue/4)$
$(q://QID303/ChoiceTextEntryValue/5)$
$(q://QID303/ChoiceTextEntryValue/6)$
$(q://QID303/ChoiceTextEntryValue/7)$
$(q://QID303/ChoiceTextEntryValue/8)$

The correct answer is: Russia

Take a look at your reasoning again. Which categories did your reasoning fail in? (Select all that apply.)
- Failed to look hard enough for counter-evidence (I now see that there are arguments I didn't think of)
- Looked for counter-evidence but didn't find any (This is not an example of my side bias because you tried)
- Under-weighted counter-evidence (I thought of evidence against my answer but did not take it seriously enough)
- Over-weighted evidence (I took evidence for my answer too seriously)
- Know it (I already knew the answer so I did NOT have to think)
- Good AOT (I thought of counter-evidence and used it appropriately)

Part III: R5

These page timer metrics will not be displayed to the recipient.
First Click: 111.166 seconds
Last Click: 111.166 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

Which of the U.S. metropolitan areas below has the highest percentage of gay population?
Between these options, $(q://QID98/ChoiceGroup/AllChoices?displayLogic=0)$, you picked: $(q://QID91/ChoiceGroup/SelectedChoices)$.

based on your notes:
$(q://QID305/ChoiceTextEntryValue/1)$
$(q://QID305/ChoiceTextEntryValue/2)$
$(q://QID305/ChoiceTextEntryValue/3)$
$(q://QID305/ChoiceTextEntryValue/4)$
$(q://QID305/ChoiceTextEntryValue/5)$
$(q://QID305/ChoiceTextEntryValue/6)$
$(q://QID305/ChoiceTextEntryValue/7)$
$(q://QID305/ChoiceTextEntryValue/8)$

The correct answer is: Austin, TX
Take a look at your reasoning again. Which categories did your reasoning fall in? (Select all that apply.)

- Failed to look hard enough for counter-evidence (I now see that there are arguments I didn't think of)
- Looked for counter-evidence but didn't find any (This is not an example of myside bias because you tried)
- Under-weighted counter-evidence (I thought of evidence against my answer but did not take it seriously enough)
- Over-weighted evidence (I took evidence for my answer too seriously)
- Knew it (I already knew the answer so I did NOT have to think)
- Good AOT (I thought of counter-evidence and used it appropriately)

Part III: RS

These page timer metrics will not be displayed to the recipient.

First Click: 111.165 seconds
Last Click: 111.165 seconds
Page Submit: 0 seconds
Click Count: 1 clicks

According to a 2013 report by World Health Organization, which of the following countries has the highest life expectancy at birth?

Between these options: ${q://QID144/ChoiceGroup/AllChoices?displayLogic=0}, you picked: ${q://QID146/ChoiceGroup/SelectedChoices},

based on your notes:

${q://QID307/ChoiceTextEntryValue1}
${q://QID307/ChoiceTextEntryValue2}
${q://QID307/ChoiceTextEntryValue3}
${q://QID307/ChoiceTextEntryValue4}
${q://QID307/ChoiceTextEntryValue5}
${q://QID307/ChoiceTextEntryValue6}
${q://QID307/ChoiceTextEntryValue7}
${q://QID307/ChoiceTextEntryValue8}

The correct answer is: Italy

Take a look at your reasoning again. Which categories did your reasoning fall in? (Select all that apply.)

- Failed to look hard enough for counter-evidence (I now see that there are arguments I didn't think of)
- Looked for counter-evidence but didn't find any (This is not an example of myside bias because you tried)
- Under-weighted counter-evidence (I thought of evidence against my answer but did not take it seriously enough)
- Over-weighted evidence (I took evidence for my answer too seriously)
- Knew it (I already knew the answer so I did NOT have to think)
- Good AOT (I thought of counter-evidence and used it appropriately)
Thank you for taking your time to learn about AOT.

We would like to know what you think of the experience.

Please indicate how much you agree or disagree with the following statements.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOT will help me make better probability judgments</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AOT is an easy concept to understand</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AOT is an easy concept to apply</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Learning about AOT was a waste of my time</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I often used AOT before this training session</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I will apply AOT in the future</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I have no intention to use AOT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
APPENDIX E: PRE-TEST MODULE FOR STUDY 6

The following pages include the pre-test module for control condition subjects.
Warm-up Exercises

Introduction
This part of the study looks at people's thought processes as they answer various types of questions.

On the next few pages we will give you 6 problems to think about. The questions topics include country population, movie release dates, and socio-economic facts. We ask you to list your thoughts, entering them in a box. Then you give your best guess about the answer and assign probabilities to each of the options. You have the options to skip 2 items, but please make sure that you answer at least 4 items.

What is your email address? (Please use the one with which you registered.)

---

Q1

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three following countries has the highest population?

Vietnam
Germany
Iran

Answer the question to the best of your abilities. Please don't look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.
Q2

Which of the three following countries has the highest population?

Indonesia
Brazil
Pakistan

Answer the question to the best of your abilities. Please don’t look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.
What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indonesia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pakistan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Q3

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three following countries has the highest population?
Russia
Nigeria
Bangladesh

Answer the question to the best of your abilities. Please don't look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.
What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

- Russia
- Nigeria
- Bangladesh

Q4

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three movies has the latest release date?
- Gone with the Wind
- The Godfather
- Twelve Angry Men

Answer the question to the best of your abilities. Please don't look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.
What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%).

<table>
<thead>
<tr>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gone with the Wind</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The Godfather</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Twelve Angry Men</td>
</tr>
</tbody>
</table>

Q5

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three movies has the latest release date?

Casablanca
Singin’ In the Rain
Breakfast at Tiffany's

Answer the question to the best of your abilities. Please don't look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.
What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

<table>
<thead>
<tr>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casablanca</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singin' in the Rain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breakfast at Tiffany's</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Q6

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three movies has the latest release date?
The Jurassic Park
Jaws
The Terminator

Answer the question to the best of your abilities. Please don't look up the answer. Use the text area box to record your thought process. We will later show your notes, response and the correct answer.
What is the probability that each of these options is the correct answer? (Your probabilities should add up to 100%.)

0 10 20 30 40 50 60 70 80 90 100

- The Jurassic Park
- Jaws
- The Terminator

Which country has the largest population

Between these options: $(q://QID21/ChoiceGroup/AllChoices?displayLogic=0)$, you picked: $(q://QID21/ChoiceGroup/SelectedChoices)$, based on your notes:

$(q://QID38/ChoiceTextEntryValue)$

The correct answer is: Vietnam
Which country has the largest population

Between these options: $(q://QID87/ChoiceGroup/AllChoices?displayLogic=0)$, you picked: $(q://QID87/ChoiceGroup/SelectedChoices)$, based on your notes:

$(q://QID86/ChoiceText/EntryValue)$

The correct answer is: **Indonesia**

Which country has the largest population

Between these options: $(q://QID91/ChoiceGroup/AllChoices?displayLogic=0)$, you picked: $(q://QID91/ChoiceGroup/SelectedChoices)$, based on your notes:

$(q://QID90/ChoiceText/EntryValue)$

The correct answer is: **Nigeria**
Which of the three movies has the latest release date?

Between these options: ${q://QID95/ChoiceGroup/AllChoices?displayLogic=0}, you picked: ${q://QID95/ChoiceGroup/SelectedChoices},

based on your notes:

${q://QID94/ChoiceTextEntryValue}

The correct answer is: The Godfather

Which of the three movies has the latest release date?

Between these options: ${q://QID145/ChoiceGroup/AllChoices?displayLogic=0}, you picked: ${q://QID145/ChoiceGroup/SelectedChoices},

based on your notes:

${q://QID145/ChoiceTextEntryValue}

The correct answer is: Breakfast at Tiffany's
These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds
Last Click: 0 seconds
Page Submit: 0 seconds
Click Count: 0 clicks

Which of the three movies has the latest release date?

Between these options: $(q://QID120/ChoiceGroup/AllChoices?displayLogic=0),
you picked: $(q://QID120/ChoiceGroup/SelectedChoices),
based on your notes:
$(q://QID119/ChoiceTextEntryValue)
The correct answer is: The Jurassic Park

Feedback

Do you have any comments for us?
BIBLIOGRAPHY


