# The use of alternative reasons in probabilistic judgment

Burcu Gürçay-Morris* Jonathan Baron†

We examine a behavioral measure of actively open-minded thinking (AOT), based on generation of alternative reasons or contradicting evidence. We also developed and tested a short online module to train people in actively open-minded thinking, based on the same idea. In three studies reported here, subjects made probabilistic judgments in three-choice almanac questions. Subjects were overconfident in their probability judgments, but overconfidence was lower (in two studies) in subjects who scored higher on a measure of AOT beliefs, and on trials when the behavioral measure was higher. Study 2 showed that forcing subjects to think of alternative reasons reduced overconfidence. Study 3 tested the effectiveness of new online AOT training modules we designed for adults, in a pre-test/post-test design. The training module, relative to a control condition, increased both the behavioral and belief-based measures of AOT and reduced overconfidence. Otherwise, the behavioral and belief-based AOT measures did not correlate with each other.

Keywords: overconfidence, actively open-minded thinking, probability judgment

*Microsoft, Seattle. Email: bgurcaymorris@gmail.com.
†Department of Psychology, University of Pennsylvania. Email: jonathanbaron7@gmail.com.

This article is extracted from the BG-M's 2016 doctoral dissertation in Psychology and the University of Pennsylvania, supervised by JB: "The use of alternative reasons in probabilistic judgment." It is available at https://jbaron.org/~jbaron/theses/GurcayMorrisDissertation.pdf. The three studies reported here correspond to Studies 1a and 1b, Study 4, and Study 7, respectively. The studies not reported here are consistent with the results we report, including the fact that the Aot-Beliefs scale failed to predict anything in some, yet succeeded in others. Many of the omitted studies were designed to study individual differences yet had insufficient power to detect some effects of interest.

We thank the other committee members, Barbara Mellers and Uri Simonsohn, for helpful comments. As we note, Mellers was especially helpful in the design of the training manipulation in Study 3.

*Microsoft, Seattle. Email: bgurcaymorris@gmail.com.
†Department of Psychology, University of Pennsylvania. Email: jonathanbaron7@gmail.com.

# 1 Introduction

> 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.
>
> — John Stuart Mill, *On Liberty*, 1859

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 (1985, 1988) originally 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 search for 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 (AOT) is a dispositional theory of good thinking. Dispositions are potentially malleable (Baron, 1985; Perkins, Jay, & Tishman, 1993), therefore posing a contrast with cognitive capacities, such as mental speed. Several studies show effects of AOT (measured in various ways, some not labeled as AOT) that appear to be distinct from cognitive capacities (e.g., Stanovich & West, 1997; Kokis et al., 2002; Klaczynski, 1997; Klacyznski & Gordon, 1996; see also Stanovich, West, & Toplak, 2013 for additional discussion).

Of interest here is "myside bias" (Perkins, 2019; now roughly equivalent to "confirmation bias" as discussed by Nickerson, 1998), which is a bias toward conclusions that already seem favored in a person's thinking. AOT, as a dispoition, is opposed to myside bias. A major purpose of AOT as a standard is to counteract the effects of myside bias, and that function is the focus of the present paper.

AOT is correlated with various measures of performance and a variety of traits that would seem to involve myside bias. These include: heuristics-and-biases tasks, including belief bias, that is, evaluation of syllogisms by the truth of their conclusions (Toplak, West, & Stanovich, 2014, 2016); acceptance of fake news, delusional ideation, and various beliefs based on authority and tradition rather than reflective thought (Bronstein et al., 2018; Pennycook, Cheyne, Koehler, & Fugelsang, 2020), including deontological moral judgment and acceptance of "divine command theory", the belief that morality is given to us by God, so that we should not try to understand it (Baron, Scott, Fincher, & Metz, 2015); beliefs in the supernatural and paranormal (Svedholm & Lindeman, 2013, 2018); belief in conspiracy theories, ontological confusions, and anti-science attitudes (Rizeq, Flora, & Toplak, 2020); confusion of theory and evidence (Sá, Kelley, Ho, & Stanovich, 2005); real-world outcomes such as poor financial management (Toplak, West, & Stanovich, 2016); and geopolitical forecasting performance (Mellers et al., 2015).

## 1.1 Measures of Actively Open-Minded Thinking

Many measures of AOT ask subjects 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, 2019; 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). See Baron, Gürçay, and Metz (2017) for a review of the various approaches, and Metz, Baelen, & Yu (2020), for discussion of AOT measurement in young adolescents.

In the current paper we use a short form of the AOT Belief scale (see Appendix B), similar to many other versions, and equivalent to them for practical purposes. The version we use focuses on AOT as a remedy for myside bias. Compared to the 8-item version used by Baron, Scott, Fincher, and Metz (2015), 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.

We report here a behavioral measure of AOT, based on thought listing, that is designed to assess myside bias in actual thought rather than beliefs about what kind of thought people should do. (The two are related because, presumably, people try to do what they think they should do, as argued by Baron, 1995). Our focus, like that of all past measures that we know, in on myside bias, despite the fact that one of our target measures is overconfidence. Since these studies were done, Baron (2019) has argued that high confidence when low confidence is warranted

is a major component of individual differences in AOT, on a par with myside bias. Additional items have now been added to the short scale to measure beliefs about confidence. As it happens, these items usually correlate well with those that assess myside bias.

In another approach, not based on thought listing, Stanovich and West (1997) developed a behavioral measure to evaluate myside bias called Argument Evaluation Test (AET) to evaluate such arguments. 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. A critic then presented an argument against this justification. Finally, the fictitious individual offered a rebuttal of the counterargument. Subjects were told to assume that the argument and rebuttal was both factually correct. They were asked to evaluate the strength of the rebuttal independently of their own opinions or beliefs. Subjects evaluations were then compared to a summary measure of eight expert judges. The authors estimated myside bias by trying to predict subjects' ratings from both the expert 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 in the direction of their preexisting opinions. This method is useful but may require many sets items, each with much reading for the subject to do, in order to yield a reliable measure of individual differences.

Other measures have been based on thought listing. In an early demonstration, Perkins (2019, originally 1986) asked students to write down their thoughts on issues that were "genuinely vexed and timely" and that could be discussed on the basis of knowledge that most people have, e.g., "Would providing more money for public schools significantly improve the quality of teaching and learning?" Most students gave more arguments on their favored side, "myside" thoughts, than on the other side. When the students were asked to try harder to think of arguments on each side, they thought of very few additional myside arguments but many additional otherside arguments. Left to their own devices, the students look primarily for reasons to support their initial opinion, but out of biased search rather than lack of ability or knowledge. Note that this method is focused on search rather than inference, while the AOT Belief scale is mostly about inference.

Baron (1995, following Perkins) used a similar method to assess 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. The same subjects were also asked to evaluate sets of arguments regarding abortion from fictitious students by grading

them. Subjects who exhibited myside bias in their own justifications also gave relatively higher grades to those sets that were more one-sided (with most arguments on one side). Toplak and Stanovich (2003) used a similar thought-listing measure to examine the effects of years in college on myside bias, finding that the bias was reduced in later years.

These thought listing measures were also used in studies that tried to de-bias overconfidence. Koriat, Lichtenstein, and Fischhoff (1980) and Hoch (1985) asked subjects to list reasons that was either for or against subjects' pet beliefs, asking for contrary reasons was effective even when combined with a request for supporting reasons.

Although thought listing methods seems ideal in some ways, they require time-consuming scoring by the experimenters or their helpers. Additionally, some issues used in such tasks need to appeal to a wider population such that people are familiar enough with them to have preexisting beliefs or opinions about them. We adopt a similar approach that is designed to work around these problems. We aimed here to develop a measure that could work for trivia questions and other such items, and that could be scored by the subjects themselves. (The latter aim, as we shall see, was not fully realized.)

Thus, we report results from studies where we tested individual differences in AOT on probabilistic judgment tasks using almanac questions with three possible responses, such as which of three cities is largest. Subjects assigned probabilities to the three options. They also wrote reasons and classified each reason as favoring or opposing one of the 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.

We investigated whether this behavioral AOT measure would correlate with Baron's AOT Belief Scale, subjects' accuracy, defined as Brier scores, and overconfidence. In general, we asked whether AOT was associated with lower overconfidence and higher accuracy. Haran, Ritov, and Mellers (2013) found that AOT beliefs (as measured by a scale very similar to the one we used in this study) were associated with lower overconfidence.

## 1.2 Training AOT

A second purpose of this paper is to describe (in Study 3) a brief training manipulation designed to increase AOT within the particular task we use. In the present context, this is useful only as a pilot for a more serious training study with a broader goal than one particular task and with more intensive training over a longer period. Although we think that the structure of the training may be a useful model, we use it here mainly to provide evidence about the causal relationships among the various measures we use. Study 2 is also designed for this purpose, but it

involves a manipulation of instructions with immediate effects, while Study 3 involves a test following a short delay.

We note that many studies have attempted to reduce myside bias, with some success. The clearest example is that of Perkins (2019), who tested and trained myside bias directly (as we noted). Many other studies have attempted to train thinking or decision making and have included components related to AOT in general and myside bias in particular. These are reviewed in Baron and Brown (1993), Nickerson (1988; also Nickerson, Perkins, & Smith, 1985), and Baron et al., (2017). A recent example is Mellers et al. (2015).[1]

## 2 Study 1

Study 1 is a prelminary study using questions about U.S. metropolitan area population (Appendix A). It was one of four studies testing various types of questions and was the only one that yielded overall correct response rates above chance (1/3). The other pilots used matriecs, logic questions, and questions about food content (calories and nutrients). Yet its major results were robust.

### 2.1 Method

**Subjects.** 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.

**Questions.** 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 in measurement of individual differences. 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 A lists the questions presented to the subjects.

[1]Even more recently, Sellier, Scopelliti, and Morewedge (2019) claimed to reduce "confirmation bias" but actually succeeded in reducing the use of a "positive test strategy", which is not necessarily a bias (Baron, 2008).

**Procedure.** Subjects did the task on their own computers over the Internet. They first read the instructions regarding the task. The instructions and the full study can be found on https://jbaron.org/~jbaron/ex/bg/args/args2c.html.

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. 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.

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 was the correct 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 were told 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 each page, they could click a button to move onto the the next page, 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 BG-M to ensure that subjects' reasons were classified correctly.

After answering all questions, the subjects completed the Aot-Beliefs scale (Appendix B). They 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.

## 2.2 Results

Examination of the items found that the proportion of correct responses to items 1 and 4 were outliers on the low side (6% and 3% correct, respectively). We report results for the remaining 18 items. The results are essentially the same but slightly stronger for all 20 items.

### 2.2.1 Individual Differences Measures

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. We subtracted subjects' responses from 3 so that their responses now varied between −2 and 2, and we reversed the reverse-scored items. Aot-Beliefs scores, the mean, ranged from −0.750 to 1.75. The reliability (alpha) was .76.

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 (mean, 0.276; s.d., 0.415); reliability for this measure was .93.

We also report Correct both at the item level — whether an answer was correct (1) or not (0) — and the subject means. The mean was 45% correct (s.d. across subjects, .16; alpha .30).

Prob is the maximum probability assigned to options (mean .58, s.d. .16, alpha .96).

Over is overconfidence, which is Prob minus Correct, both for items and subjects. The mean was .13 (s.d. across subjects .21, alpha .71). At the item level, of course, this measure is heavily influenced by Correct, which has a much larger variance across items.

BS is the Brier score, which is computed at the level of items, using all three probabilities for each item, $(1 - p_c)^2 + p_i^2 + p_j^2$, where $p_c$ is the probability assigned to the correct answer, and $p_i$ and $p_j$ are the probabilities assigned respectively to the other two answers. A score of 0 results from 1.00 being assigned to the correct answer and 0 to the others. The maximum score of 2 can result from assigning a probability of 1.00 to one of the incorrect answers. The mean score was .71 (s.d. .16 across subject, alpha .57).

## 2.3 Results

At the level of subjects none of the major measures correlated with each other except for those that must necessarily be true (e.g., those involving Over, Correct, Prob, and BS). In particular, Aot-Beliefs and Aot-Reasons did not correlate close to significantly with anything, including each other (r = -.06, slightly in the wrong direction).

To examine possible relationships at the level of items, we fit multi-level models with the lmer() function of the lme4 package of R (Bates et al., 2015), treating subjects and items as crossed random effects. This method allowed us to distinguish correct and incorrect responses, which of course varied as a function of both subjects

and items. In reporting most of these results, and some related results, we provide t values because they allow comparison of effects, and they are familiar. The reported values are simply the ratio of the effect to the estimated standard error, but p-values inferred from corresponding z values would be somewhat lower than they should be. Our concern is not with whether some effect is significant in an absolute sense, but whether it is clear, unclear, or clearly absent, and the t-values permit this inference. Estimates of p-values from such models require several tenuous assumptions. Most of the results we report are either clear (e.g., t values over 2.5) or clearly absent. We also report raw regression weights, which indicate the effect size in units of the dependent variable.

Predicting BS from Aot-Reasons, Correct (coded 1/0), and their interaction showed a clear interaction (coeffient .117, t=3.86). For correct responses only, Aot-Reasons was related positively to BS (.040, t=2.87). For incorrect responses, the relation was negative (−.068, t=−2.67). This makes sense because higher Aot-Reasons means more reasons against the favored answers, which would lower Prob, thus increasing the Brier score for correct answers and decreasing (improving) it for incorrect answers. Predicting Prob from Aot-Reasons, Correct and their interaction showed no interaction (.008, t=.597), but, in the absence of the interaction, Aot-Reasons was related negatively to probability (−.035, t=−4.41).

In sum, thinking of reasons against a favored options seems to reduce the probability assigned to that option, thus increasing the Brier score (so that it indicates worse performance) for correct answers and reducing it for incorrect answers. For the kinds of items we used, these effects are similar, so that there is little overall improvement in performance, in these studies. The overall benefit might be larger if the effort to think of counter-arguments were more productive when the favored option is incorrect. We shall explore this issue further in Studies 2 and 3, and Study 3 will also reconsider the apparent uselessness of the Aot-Beliefs scale.

Subjects were also overconfident. The mean for Over (intercept in the multi-level model with no predictors and random effects for subjects and items) was .127 (t=2.96).

We might expect Over to be lower when Aot-Reasons was higher, but this did not occur overall; the coefficient was .002 (slightly in the wrong direction). However, Over is highly dependent on correctness. When Correct was included as a covariate, the effect of Aot-Reasons on Over was very clear and in the expected direction (coefficient −.034, t=−4.34). This analysis is consistent with the finding that Aot-Reasons is negatively associated with Prob. Studies 2 and 3 address the question of the direction of causality in these relationships.

# 3 Study 2

The main purpose of this study was to investigate whether asking subjects to give a reason against their preferred answer or supporting a non-preferred answer would improve their Brier scores and reduce overconfidence. This experiment thus tests whether the correlations found in Study 1 could result from an effect of thinking of reasons on confidence rather than just an effect of available arguments on both confidence and the Aot-Reasons measure.

## 3.1 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 Study 1. 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. 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 https://jbaron.org/~jbaron/ex/bg/args/args5.html.

The procedure for this study was similar to that of Study 1 with the following modifications. There were two conditions: (1) a control condition where subjects would answer questions like the subjects in Study 1, 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 (Twoside) 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. 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 with the sequence of control and experimental conditions completely randomized. The other half of the subjects did the task with the first five items in the control condition, followed by items 6 through 20, consisting of ten of the questions in the experimental condition and five of them in the control condition, randomly intermixed. 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 made probability judgments about the

correctness of each option after listing their reasons.

Another way in which this study differed from Study 1 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 BG-M to ensure that subjects' reasons were classified correctly. In the case of conflict, we used her classification.

## 3.2 Results

### 3.2.1 Individual Differences Measures for Actively Open Minded Thinking

This study had only 62 usable subjects and was designed primarily for within-subject tests of asking for reasons, but we report the results for individual differences for completeness. We calculated the individual differences measures as described in Study 1. See also Table 2 for a comparison with the other studies.

The subjects' Aot-Beliefs scores ranged from −1.626 to 1.364 (M = 0.789, s.d. = 0.636), while their mean Aot-Reasons scores ranged from 0.50 to 1.40 (M = 0.82, s.d. = 0.188). A correlation across subjects between Aot-Beliefs and mean Aot-Reasons scores was slightly in the opposite direction of what would be expected (r = −0.07, t[61] = −0.55).

There was a small and statistically non-significant correlaion between mean Brier scores and Aot-Beliefs in the expected direction (r = −0.12, t[60] = −0.90, p = 0.19, one-tailed).

We could expect negative correlation between mean Aot-Reasons and mean Brier scores, with lower (better) Brier scores 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). However, we noticed that items differed considerably in this correlation. When we looked at how the correlation between Brier scores and mean Aot-Reasons behaved as the question difficulty (percept correct) varied, we found a strong correlation in the expected direction (r = −0.52, t(18)=−2.62, p=0.01, two-tailed). As the questions get more difficult, the correlation between Brier scores and AOT reasoning get stronger. Aot-Reasoning reduces confidence when the item produces more incorrect responses.

### 3.2.2 Relation between Aot-Reasons and measures based on confidence

In general, we replicated effects found (and not found) in Study 1 concerning the relation between Aot-Reasons and various measures based on confidence. We report these here, ignoring the fact that Aot-Reasons was affected by the two-side instructions.

We would expect Aot measures to correlate negatively with overconfidence across subjects. We found at best very weak correlations between Aot-Reasons and overconfidence across subjects (r=−.02 with Over, t(60) = −0.16, p=.44, one-tailed). Results for Aot-Beliefs were similar (r = −0.18 with Over1 t(60) = −1.44, p=.08, one-tailed. The remaining analyses look at items as well as subjects.

The relation between Aot-Reasons and Brier scores, which should be negative, was only slightly below zero (coefficient −0.035 with random effects for subjects and items, t=−1.36). As before, Aot-Reasons interacted with correctness (.219, t=6.40). For correct items, Aot-Reasons was associated with higher (worse) Brier scores (.062, t=4.58), and for incorrect items the association was negative (−.127, t=−4.74).

Prob (the probability assigned to the favored option) declined with Aot-Reasons (−.064, t=−7.16). There was no interaction between Aot-Reasons and correctness (.002, t=0.14).

Subjects were overconfident (with subjects and items as random effects: .095, t=2.47). Overconfidence was negatively associated with Aot-Reasons, as it should be, but very weakly (Over, −.035, t=1.41). But we found a strong relationship when Correct was included in the model, as in Study 1: (Coefficient −.062, t=−7.07).

### 3.2.3 Effect of Twoside Instructions

The main purpose of this study was to examine the effect of instructing subjects to provide reasons opposed to their favored hypothesis.

Apparently, subjects followed the instructions. The mean Aot-Reasons scores for the two-side condition (1.10) exceeded that in the control condition (0.53; with random effects for subjects and items, t=22.63 for the difference). The two-side instructions also increased the total number of reasons that subjects gave, from a mean of 2.38 to 2.59. Note that this was a smaller difference than the difference for Aot-Reasons. The upshot that the number of pro-reasons decreased as a result of two-side instructions (t=−7.34).

Two-side instructions had a very small negative (beneficial) effect on the Brier score (−.031, t=1.11). However, as we might expect, the effect was different for correct answers (.027, t=1.84) and incorrect answers (−.083, t=−2.94; with t=2.94 for the interaction between correctness and two-side.

Two-side reduce Prob, as it should (−.039, t=−4.04). Interesting, this affected appeared to be largely mediated by the effect of Two-side on Aot-Reasons. When Prob was regressed on both of these predictors, the effect of Two-side went away (−.003, t=−0.23) while the effect of Aot-Reasons remained strong (−.063, essentially unchanged from what it was without Two-side, t=5.88).

The effect of Twoside on Overconfidence was very weak (−.027, t=−1.03). Twoside reduced Overconfidence from .112 to .078. However, Overconfidence was also affected by Correct (vs. incorrect). When Correct was include in the model, the effect of Twoside was stronger and much more convincing (−.038, t=4.01). Correct was a nuisance variable; by itself it did not predict Overconfidence at all (coefficient −.014, t=−0.55).

To test for any potential carryover effects of the Two-side manipulation to control trials, we regressed the number of against-reasons subjects wrote in the control condition, Ar0, and the number of against-reasons subjects wrote in Two-side condition, Ar1, on the position of each item in the presentation order (Item), and the cumulative number of Two-side items through the current item (Twoside.cumsum) with subjects and items as crossed random effects. We did not observe any carryover effects in either regression. There was no effect of Item (0.009, t=.054) or Twoside.cumsum (−0.011, t=−0.36) on Ar0. Similarly, there was no effect of Item (0.008, t=0.85) or Twoside.cumsum (−0.013, t=0.71) on Ar1. Hence, we found no evidence for a carry-over effect.

# 4 Study 3

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 of a previous study in which training had little effect[2], we designed the training so that it included questions that would encourage the subjects to list arguments that went against their pet beliefs. In addition, the present study used a pre-test/post-test design, designed to increase statistical power by measuring pre-post differences as a function of whether or not the training was given between the pre-test and post-test.

## 4.1 Method

**Subjects.** One hundred and thirty-one subjects from the same panel as used earlier participated in the study; their ages ranged from 18 to 76 (Median = 45); 66.4% were female. Subjects who did the previous training study were

[2]The previous study is Study 6 in BG-M's dissertation. It used a between-subject design, with and without training, and found only minor effects of the training. Ignoring the training manipulation, the results concerning individual differences are much the same as those reported here. Note that the dissertation used a second measure of overconfidence, which is omitted from the present paper because of recently discovered undesirable properties (e.g., a non-monotonic relation to the standard measure).

excluded. One subject was omitted from the computation of Aot-Reasons, and any analysis that involves this measure, because this subject wrote "think so" for all reasons.

**Training.** The training module had four parts: (1) Introduction; (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.[3]

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 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 evaluate their own 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.

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

[3]The full training can https://jbaron.org/~jbaron/aot/AOTTrainingJul11.pdf.

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 evaluate their own 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 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 proceeding to the post-test.

**Procedure.** Subjects did the task on their own computers over the Internet.[4]

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 AOT scale (see Appendices A and B). 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. After the data collection was completed, the classifications of reasons were checked by one of the authors (BG-M) to ensure that subjects' reasons were classified correctly.

[4]The instructions and pre-test are avalable at https://jbaron.org/~jbaron/ex/bg/args/args10a.html for the control condition, and on https://jbaron.org/~jbaron/ex/bg/args/args10b.html for the training condition. These differ only in what they tell the subject to do next. The post-test is at https://jbaron.org/~jbaron/ex/bg/args/args10post.html.

## 4.2 Results

We used the pre-test for our analysis of individual differences, since the subject did not differ then as a function of training, and the sample was larger. Results for the post-test were not substantively different. To evaluate training effects we compared pre-post changes between the training and control group on summary measures of individual subjects.

### 4.2.1 Individual Differences Measures in Pre-test

All individual-differences measures were reasonably reliable: Aot-Beliefs .81 for alpha (.85 for Guttman's Lambda 6); Aot-Reasons .87 (.90); Correct .67 (.70); Brier score .74 (.81); Overconfidence .72 (.76); Probability .93 (.96).

The subjects' Aot-Beliefs scores ranged from −1.36 to 1.82 (mean, 0.72; s.d., 0.60. The Aot-Reasons scores ranged from 0.00 to 1.50 (mean, 0.36; s.d., 0.35). The correlation between Aot-Beliefs and mean Aot-Reasons was essentially zero, as found in Studies 1 and 2 (r = .007).

Aot-Beliefs scores were moderatately correlated with Brier scores (r=−0.31, t=−3.58) Subjects who had higher scores on Aot-Beliefs had lower (better) Brier scores than those who scored lower on Aot-Beliefs. Aot-Reasons did not correlate with Brier scores (r = −0.05, t = −0.60). Finally, we looked at how the correlation between Brier scores and mean Aot-Reasons behaved as the question difficulty varied.

As expected, subjects were overconfident on the average (mean = 0.11; t = 6.44). Overconfidence was correlated negatively with Aot-Beliefs (r=−.28, t = −3.32). However, Aot-Reasons did not correlate very well with overconfidence (r=−0.10; t=−1.16).

### 4.2.2 Analysis of Observations

Aot-Reasons and confidence varied as function of items as well as subjects, so we analyzed observations treating subjects and items as crossed random effects.

As in Studies 1 and 2, we found no overall relationship between Aot-Reasons and Brier scores (−.002, t=−0.11), but we found a strong interaction between Aot-Reasons and correctness (1=correct, 0=incorrect; 0.061, t=6.19). Aot-Reasons was positively related to Brier scores for correct responses (0.045, t=5.17) and negatively related for incorrect responses (−0.058, t=−3.73). Consistent with these results, Prob, the probability assigned to the favored response, was negatively related to Aot-Reasons (−0.040, t=−6.98).

Overconfidence (Over) at the level of items was weakly related to Aot-Reasons in the expected direction

(−0.027, t=−1.72). However, when Over was regressed on both correctness and Aot-Reasons, the latter coefficient became much stronger (−0.039), t=−6.92), as found in Studies 1 and 2 (Table 2).

### 4.2.3 Comparisons Between the Control and Training Conditions

We asked whether the actively open-minded thinking training had any effects on subjects' accuracy scores, overconfidence, and the two measures of actively open-minded thinking. In each case, we compared the pre-post changes for the two groups, training and control. Table 1 shows the results.

TABLE 1: Means for the major measures, Study 3, comparing training and control. p-values are one tailed. Some t-values have fewer than 220 df because of missing data.

| Measure | Group | Pre-test | Post-test | Change | $t_{220}$ | p |
|---|---|---|---|---|---|---|
| Aot-Beliefs | Training | 0.758 | 0.914 | 0.156 | 2.091 | .0194 |
| | Control | 0.705 | 0.674 | -0.031 | | |
| Aot-Reasons | Training | 0.374 | 0.567 | 0.193 | 7.721 | .0000 |
| | Control | 0.338 | 0.286 | -0.052 | | |
| Brier score | Training | 0.652 | 0.607 | -0.045 | 0.331 | .3777 |
| | Control | 0.634 | 0.600 | -0.034 | | |
| Correct | Training | 0.488 | 0.553 | 0.065 | 1.807 | .0367 |
| | Control | 0.530 | 0.558 | 0.028 | | |
| Probability | Training | 0.594 | 0.596 | 0.002 | 1.534 | .0641 |
| | Control | 0.646 | 0.675 | 0.029 | | |
| Overconf. | Training | 0.101 | 0.037 | -0.064 | 2.632 | .0049 |
| | Control | 0.116 | 0.116 | 0.000 | | |

The training effect on Aot-Reasons was, in essence, a manipulation check, and the manipulation was successful. The effect on Aot-Beliefs indicates that the training also led subjects to understand the value of actively open-minded thinking, to some extent, although we doubt that the effect of such a short training manipulation would last very long. Of primary interest is the effect of the training on overconfidence.

The training had little effect on Brier scores or correctness. This is not surprising, since the subjects had no opportunity to gather information. In other situations, such training may be more helpful (Mellers et al., 2015). Training did reduce the maximum probability assigned a little.

The beneficial effect of training on overconfidence may be somewhat specific to difficult problems, such as those used in this study. To examine this issue, we computed for each of the 20 items the beneficial effect of the training on overconfidence, just as we did for the overall measure of overconfidence. And we computed the

mean accuracy for each item, averaging over both pre-test and post-test, and both training conditions (range: .28 to .83, median .57). The correlation between accuracy and the benefit of training, across the 20 items, was −.471 (p=.018, one tailed). Indeed, the beneficial effect was greatest for the items with near-chance performance, and the benefit essentially disappeared when mean accuracy was .6 or higher.[5]

However, the benefit of training was consistent across all 20 items, on the whole (mean .065, $t_{19}$ = 2.88, p=.0048, one tailed). Moreover, Table 1 shows that training affected both components of overconfidence, probability and accuracy (Correct). Overconfidence is the difference between these. The (unstandardized) coefficient for the regression of change in accuracy on condition (.037) was in fact higher than that for regression of change in probability on condition (.021). A similar analysis by subject showed that the improvement in overconfidence from training was positively related to the subject's accuracy (p=.0210, two tailed, for the interaction between condition and accuracy; regression coefficients 0.27 for the training condition, −0.03 for the control condition.

At higher levels of accuracy (levels higher than those of almost all individual subjects), overconfidence is necessarily limited — because it is defined as probability minus accuracy, and both are limited by 1.0 — so it is not clear the small effect of training on probability assigned to the favored answer would lead to increased underconfidence when accuracy is high, although such an effect cannot be ruled out.

## 5 General Discussion

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.

[5]Only one item — "Which is the movie that has the most recent release date?: V for Vendetta, The Matrix, Se7en." — with an accuracy of .65 (the 8th most accruate out of 20 hence not extremely high), showed a possibly harmful effect (−.18; $t_{103}$ = .050, two tailed, uncorrected for multiple tests).

TABLE 2: Summary of results concerning AOT and Overconfidence in the three studies. T-values are in parentheses. "|Correct" means that Correct was included as a covariate.

| | Study 1 (n=69) | | Study 2 (n=62) | | Study 3 (n=131) | |
|---|---|---|---|---|---|---|
| Correlation coefficients across subjects | | | | | | |
| Aot-Beliefs with Aot-Reasons | −.07 | (−0.55) | −.07 | (−.55) | .01 | (0.08) |
| Aot-Beliefs with Overconfidence | .09 | (0.73) | −.18 | (−1.44) | −.28 | (−3.32) |
| Regression weights (unstandardized) across observations, with random effects for subjects and items | | | | | | |
| Aot-Reasons and Overconfidence | .002 | (0.07) | −.035 | (−1.41) | −.027 | (−1.72) |
| Aot-Reasons and Over.\|Correct | −.034 | (−4.34) | −.062 | (−7.07) | −.039 | (−6.92) |
| Correct (test of intercept vs. 1/3) | .45 | (2.88) | .53 | (4.49) | .51 | (4.22) |

Table 2 shows some of the major relationships that could be examined in all three studies reported here. Aot-Reasons, the tendency to think of arguments or reasons that go against an initially favored conclusion, was defined for each observation, so we report multi-level models with random effects for subjects and items. Aot-Beliefs, however, was defined for each subject, so the top part of Table 2 shows simple correlations.

We expected these two AOT measures to correlate with each other (as found by Baron, 1995, using very different measures), but they did not. We have no clear explanation for this failure, since both measures seemed to behave as expected, if only weakly, in the analysis of instruction and training effects (Studies 2 and 3). The best explanation we can suggest is that the Aot-Reasons measure was determined largely by available knowledge, in the absence of the training manipulation. Even subjects who believed that good thinking is actively open-minded did not try very hard to think of additional reasons unless instructed to do so.

The expected correlation between Aot-Beliefs and Overconfidence was not found at all in Study 1 and was not statistically significant in Study 2, although it was clealy present in Study 3. Note that samples in Studies 1 and 2 were small compared to Study 3, and the items in Study 1 were more difficult than those in Studies 2 and 3 (last row of Table 2), even after eliminating two of the items. In addition, the Aot-Beliefs scale in Study 1 was missing the last 3 items (Appendix B), which were most relevant to the questions at hand, which involved problems faced for the first time. But we really do not know why Study 1 showed no hint of the expected correlation.

Interestingly, the Aot scale we used was essentially all about myside-bias, with no questions about overconfidence as such. Later versions (e.g., Baron, 2019) explicitly included such questions.

We found no overall relation between Aot-Reasons and Overconfidence at the level of items. However, Overconfidence is the difference between Prob and Correct, and Correct is either 0 or 1. When we adjust for

this large source of variance statistically, the relation between Aot-Reasons and Overconfidence is strongly in the expected direction. However, this relationship does not imply an effect of effort to think of contrary reasons on overconfidence. Rather, both Aot-Reasons and confidence itself could be determined by the availability of reasons in memory. Thus, this relationship serves mainly as a validation of the confidence measure itself.

## 5.1 Requiring reasons on the other side

In Study 2 we required subjects to list at least one reason against their preferred answer in half of the questions (Twoside condition). This Twoside condition had higher Aot-Reasons than the control condition, and more reasons overall than the control condition. Importantly, the manipulation reduced Overconfidence, an effect that was clear when Correct was included in the model as a nuisance variable.

We also 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 the correct answer is known hurts the 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 leads to worse Brier scores.

## 5.2 Training Adults in Actively-Open Minded Thinking

In Study 3, we report the use of a training module, improved over an earlier attempt, where we tried to encourage subjects to generate more alternative reasons, put the concepts they learned into practice, and provided feedback. The results in Study 3 showed that subjects who went through the training module improved their Aot-Reasons

and Aot-Beliefs scores, and reduced their overconfidence from the first round to the second, compared to just doing the same task twice.

In sum, we were able to train our subjects in AOT by using an online training module that took less than an hour, and the training reduced overconfidence and improved Aot-Beliefs and Aot-Reasons scores.

### 5.3 Limitations and Future Directions

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-M) 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 reduced 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 could 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. Moreover, overconfidence is most insidious when a high subjective probability of being correct is paired with an incorrect belief.

A related criticism of the study concerns the questions used. 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. 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 (or less) satisfied with their decisions after the outcome in the long run.

Our training module was successful in changing AOT behavior in subjects, and improving their performance, especially 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 likely that the effects of this brief online training would last less than the extensive training sessions devised by Perkins (2019) or Baron, Badgio, and Gaskins (1986), since these decision making courses took over the course of several weeks. Future studies will need to investigate how lasting the effects of online AOT can be.

### 5.4 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.

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 a short AOT training modules. The training increased subjects' Aot-Beliefs and Aot-Reasons scores, and reduced their overconfidence.

## References

Baron, J. (1985). *Rationality and intelligence*. Cambridge, England: Cambridge University Press.

Baron, J. (1988). *Thinking and deciding* (first edition). New York: Cambridge University Press.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking & Reasoning*, *1*(3), 221–235.

Baron, J. (2008). *Thinking and deciding* (fourth edition). New York: Cambridge University Press.

Baron, J. (2009). Belief overkill in political judgments. (Special issue on Psychological Approaches to Argumentation and Reasoning, edited by L. Rips). *Informal Logic*, *29*, 368–378.

Baron, J. (2019). Actively open-minded thinking in politics. *Cognition, 188*, 8–18.

Baron, J., Badgio, P., & Gaskins, I. W. (1986). Cognitive style and its improvement: A normative approach. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 3*, pp. 173–220. Hillsdale, NJ: Erlbaum.

Baron, J., & Brown, R. V. (Eds.) (1991). *Teaching decision making to adolescents*. Hillsdale, NJ: Erlbaum.

Baron, J., Gürçay, B., & Metz, S. E. (2017). Reflection, intuition, and actively open-minded thinking. In M. Toplak & J. Weller (Eds.), *Individual differences in judgment and decision making: a developmental perspective*, pp. 107–126. Psychology Press.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51* (6), 1173–1182.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bronstein, M., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2018). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition, 8*(1), 108–117.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making, 8*(3), 188–201.

Herek, G. M., Janis, I. L., & Huth, P. (1987). Decision making during international crises. Is quality of process related to outcome? *Journal of Conflict Resolution*, *31*, 203–226.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy of predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11* (4), 719–731.

Janis, I. L. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes.* Boston: Houghton-Mifflin.

Klaczynski, P. A. (1997). Bias in adolescents' everyday reasoning, and its relationship with intellectual ability, personal theories, and self-serving motivation. *Developmental Psychology*, *33* (2), 273–283.

Klaczynski, P. A., & Gordon, D. H. (1996). Self-serving influences on adolescents' evaluation of belief-relevant evidence. *Journal of Experimental Child Psychology*, *62*, 317–339.

Kokis, J. V., Macpherson, R., Toplak, M. E., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, *83* (2002), 26-52.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1–14. https://doi.org/10.1037/xap0000040

Metz, S. E., Baelen, R. N., & Yu, A. (2020), Actively open-minded thinking in American adolescents. *Review of Education, 8*(3) 768–799. https://doi.org/10.1002/rev3.3232

Mill, J. S. (1863). *On Liberty*. Boston, MA: Ticknor and Fields. (Original work published in 1859)

Nickerson, R. S. (1988). Improving thinking through instruction. *Review of Research in Education*, *15* (1988-1989), 3-57.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220.

Nickerson, R. S., Perkins, D. N., & Smith, E. E. (1985). *The teaching of thinking*. Hillsdale, NJ: Erlbaum.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2020). On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. *Judgment and Decision Making, 15*, 476–498.

Perkins, D. (2019). Learning to reason: The influence of instruction, prompts and scaffolding, metacognitive knowledge, and general intelligence on informal reasoning about everyday social and political issues. *Judgment and Decision Making, 14*, 624–643.

Perkins, D., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. Merrill-Palmer Quarterly, 39 (1), 1–21.

Rizeq, J., Flora, D. B., & Toplak, M. E. (2020). An examination of the underlying dimensional structure of three domains of contaminated mindware: paranormal beliefs, conspiracy beliefs, and anti-science attitudes. *Thinking and Reasoning,*

Sá, W. C., Kelley, C. N., Ho, C., & Stanovich, K. E. (2005). Thinking about personal theories: Individual differences in the coordination of theory and evidence. *Personality and Individual Differences*, *38* (2005), 1149–1161.

Sellier, A.-L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing training improves decision making in the field. *Psychological Science, 30*(9), 1371–1379. really this is about positive test strategy, not conf. bias.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior beliefs and individual differences in

actively open-minded thinking. *Journal of Educational Psychology*, *89* (2), 342-357.

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

Stanovich, K. E., & West, R. F. (2008). On the failure of cognitive ability to predict myseidne and one-sided thinking biases. *Thinking and Reasoning*, *14* (2), 129–167.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22* (4), 259–264.

Svedholm-Häkkinen, A. M., Lindeman, M. (2018) Actively open-minded thinking: Development of a shortened scale and disentangling attitudes towards knowledge and people. *Thinking and Reasoning, 24*(1), 21–40.

Svedholm, A. M. & Lindeman, M. (2013). The separate roles of the reflective mind and involuntary inhibitory control in gatekeeping paranormal beliefs and the underlying intuitive confusions. *British Journal of Psychology, 104*, 303–319. https://doi.org/10.1111/j.2044-8295.2012.02118.x

Toplak, M. E., & Stanovich, K. E. (2003). Associations between myside bias on an informal reasoning task and amount of post-secondary education. *Applied Cognitive Psychology, 17*, 851–860.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental Psychology, 50*(4), 1037–1048. doi:10.1037/a0034910

Toplak, M. E., West, R. F., & Stanovich, K. E. (2016). Real-world correlates of performance on heuristics and biases tasks in a community sample. *Journal of Behavioral Decision Making, 30*, 541–554, DOI:10.1002/bdm.1973

# Appendix A: Cases used

### Study 1: "Which is largest?"

| | |
|---|---|
| Chicago IL, Washington DC, Seattle WA | Philadelphia PA, Boston MA, Baltimore MD |
| Phoenix AZ, Atlanta GA, Louisville KY | San Francisco CA, Dallas TX, Portland OR |
| Baltimore MD, Columbus OH, Las Vegas NV | Las Vegas NV, Milwaukee WI, St. Louis MO |
| Indianapolis IN, Cincinnati OH, Tampa FL | Nashville TN, Orlando FL, Milwaukee WI |
| Austin TX, Dallas TX, Atlanta GA | San Antonio TX, Hartford CT, Austin TX |
| Albany NY, Cincinnati OH, Salt Lake City UT | Denver CO, Reno NV, Raleigh NC |
| Detroit MI, Houston TX, Cleveland OH | Minneapolis MN, Denver CO, Phoenix AZ |
| Seattle WA, San Diego CA, Baltimore MD | Pittsburgh PA, Portland OR, Sacramento CA |
| Honolulu HI, Anchorage AK, Little Rock AR | Syracuse NY, Tucson AZ, Knoxville TN |
| Boston MA, Washington DC, Miami FL | Philadelphia PA, Baltimore MD, Minneapolis MN |

### 5.5 Studies 2 and 3

Questions:

| | |
|---|---|
| City | Which is the most populous city proper? A city proper is an area contained within city limits. A city proper may not include suburbs. |
| Country | Which is the most populous country? |
| State | Which is the most populous state? |
| Univ | Which is the highest ranking university in 2014? The ranking is entirely academic and research oriented. It does not include athletics. |
| Movie | Which is the movie that has the most recent release date? |
| Logic | What can we conclude from these two statements? |

Items:

| | |
|---|---|
| City | London, United Kingdom; Moscow, Russia; New York City, U.S.A. |
| Country | Poland; Israel; Canada |
| City | Madrid, Spain; Baghdad, Iraq; Paris, France |
| Univ | Princeton University; Oxford University (UK); Massachusetts Institute of Technology |
| Country | Belgium; Poland; Ireland |
| Country | France; Egypt; Australia |
| Movie | V for Vendetta; The Matrix; Se7en |
| Logic | In a box, no blue things are triangular, and no triangular things are large. What can we conclude? "No blue things are large.; Some blue things are not large.; We can't conclude anything about blue things and large things. |
| Movie | Twelve Monkeys; Annie Hall; Stand By Me |
| Movie | The Wizard of Oz; Strangers on a Train; A Clockwork Orange |
| Country | Germany; Brazil; Poland |
| Logic | In a box, some red things are square, and some square things are large. "Some red things are large.; All red things are large.; We can't conclude anything about red things and large things. |
| Country | Australia; Turkey; Canada |
| City | London, United Kingdom; Los Angeles, U.S.A.; Seoul, South Korea |
| Univ | Duke University; Cornell University; New York University |
| Country | Canada; Japan; The United States of America |
| State | West Virginia; Rhode Island; Oregon |
| Movie | The Usual Suspects; Alien; Some Like It Hot |
| Country | Turkey; United Kingdom; Ukraine |
| State | Massachusetts; Maryland; Pennsylvania |

## Appendix B: Aot-Belief scale

1. Allowing oneself to be convinced by an opposing argument is a sign of good character.
2. People should take into consideration evidence that goes against their beliefs.

3. People should revise their beliefs in response to new information or evidence.
4. Changing your mind is a sign of weakness.
5. Intuition is the best guide in making decisions.
6. It is important to persevere in your beliefs even when evidence is brought to bear against them.
7. One should disregard evidence that conflicts with one's established beliefs.
8. People should search actively for reasons why their beliefs might be wrong.
9. When we are faced with a new question, the first answer that occurs to us is usually best.
10. When faced with a new question, we should consider more than one possible answer before reaching a conclusion.
11. When faced with a new question, we should look for reasons why our first answer might be wrong, before deciding on an answer.

Last 3 note used in Study 1.