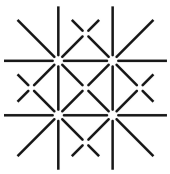
Adaptive Information Search and Judgment Strategies in Solitary and Competitive Tasks

Inauguraldissertation
zur
Erlangung der Würde
eines Doktors der Philosophie
vorgelegt der
Fakultät für Psychologie
der Universität Basel

von

Nathaniel David Phillips

aus Oregon, United States of America



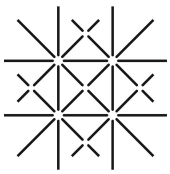
Genehmigt von der Fakultät für Psychologie

auf Antrag von

Prof. Dr. Ralph Hertwig
Prof. Dr. Jörg Rieskamp

Basel, den _____

_____
Prof. Dr. Roselind Lieb



Declaration

I, Nathaniel Phillips (born January 2, 1983, in Oregon, USA), hereby declare the following:

(i) My cumulative dissertation is based on four manuscripts**, one of which is published and three that will be submitted shortly**. I have contributed independently and substantially to all manuscripts in this dissertation and have been primarily responsible for the ideas, data collection, analyses, and writing of the papers.

(ii) I only used the resources indicated.

(iii) I marked all citations.

Berlin, August 20, 2014

Nathaniel D. Phillips

Acknowledgements

I would like to thank my two advisers, Stefan Herzog and Ralph Hertwig, for their guidance and support throughout my PhD. I am grateful to the research assistants at the former Cognitive and Decision Sciences lab at the University of Basel, Emina Canic, Lucius Caviola, Eva Guenther, Melanie Künzli, Rafael Nowak, and Samuel Senn, without whom data collection for the competitive sampling game would not have been possible.

I would also like to thank my two former office mates Dirk Wulff and Renata Suter for helping me transition both socially and academically from being a chubby, clean-shaven, wide-eyed American to the less chubby, bearded, child-sized man that I am today.

Finally, I would like to thank my siblings Heidi Trudel, Andrew Phillips, Becki Harrington and Madisen Phillips (and my new nephews David Phillips, Jack Dynamite Trudel, and Vincent Harrington) – for allowing me to run away to Europe for my PhD and supporting me over the past 3.5+ years.

Abstract

A substantial body of judgment and decision-making research focuses on decisions made under risk, where all relevant option outcome and probability information is known *a priori*. However, most real-world decision tasks are made under *uncertainty*, where such population-level information is unknown. Against this background, how can, do, and should organisms obtain and use information in order to improve their judgments and decisions under uncertainty? This dissertation addresses these questions in two distinct domains: external information search in competitive tasks (papers 1 and 2) and internal search in the context of the inner-crowd (papers 3 and 4). In paper 1, we develop a new paradigm called the Competitive Sampling Game (CSG) to study how organisms adjust search in the presence of both natural uncertainty (i.e., gamble parameters) and social uncertainty (i.e., behavior of others). The paradigm produces simulation and empirical results showing that organisms should and do dramatically reduce search in the presence of competition to almost minimal levels. In paper 2, we expand on the initial results of the CSG to show how different levels of competition drive the evolution of decision strategies. In a second domain, we address how people can improve their judgments by harnessing a diverse inner-crowd using dialectical bootstrapping. In paper 3, we apply dialectical bootstrapping to a Bayesian reasoning paradigm to show how dialectical instructions induce strategy change and how people can become more Bayesian by averaging biased non-Bayesian judgments in their inner-crowd. In paper 4, we apply the inner-crowd to a cue-based estimation task and model the effects of different estimation strategies on final estimates and confidence. Our results suggest that people can use their confidence judgments to outperform the simple average of their inner-crowd. Moreover, dialectical bootstrapping increases these effects.

Introduction

> Traditional economic theory postulates an "economic man," who, in the course of being "economic" is also "rational." This man is assumed to have knowledge of the relevant aspects of his environment which, if not absolutely complete, is at least impressively clear and voluminous. He is assumed also to have a well-organized and stable system of preferences, and a skill in computation that […] enables him to reach the highest attainable point on his preference scale. (Simon, 1955, p. 99).

If knowledge is power, then the rational human *homo economicus* assumed in classical economics is nothing short of a king in his[1] world. *Homo economicus* begins every decision with both perfect knowledge of his internal preferences and complete knowledge of all external choice options combined with their potential outcomes and probabilities. As a result of his complete *a priori* knowledge, *homo economicus* never needs to engage in any kind of information search, either internally or externally. In addition, he has immense computational capabilities that allow him to calculate the option with the highest expected value and thus optimize his decisions. For these reasons, decision making is easy for *homo economicus* – in each and every decision-making task, from deciding what to eat to whom to marry, he simply chooses the option that maximizes his expected utility. The concept of uncertainty is as foreign to *homo economicus* as fire is to a fish.

The myth of *homo economicus*

In 1955, Herbert Simon exposed *homo economicus* is an idealized fictional character – one that makes for a nice mathematical story, but cannot serve as a model for how real people do

[1] I refer to the rational economic man homo economicus as 'he' for historical continuity.

or should make decisions. Simon showed that real organisms, in contrast to *homo economicus*, never have perfect information about their environments. When a bee forages for food, it cannot know for certain how likely it is to find a patch in one corner of a field versus another (Montague, Dayan, Person, & Sejnowski, 1995). When hermit crab spots a new shell, it does not know for certain how robust the shell will be against future damage – or how many other crabs might be lurking nearby, waiting to swoop in and take it (Rotjan, Chabot, & Lewis, 2010). Instead, organisms must navigate a world of *uncertainty*, where the set of possible options, and the outcomes associated with those options, are *a priori* unknown. Moreover, real organisms are not blessed with an unconstrained cognitive system – instead, they have psychological and physiological limitations that preclude the use of complex optimization algorithms and force them to use simplifying algorithms.

Simon knew that *homo economicus* was an unattainable ideal that needed to be abandoned. In its place, he called for a new decision-making model that could make decisions under uncertainty while conforming to biological and psychological constraints. While he proposed some essential characteristics of this model (e.g., information search rules and aspiration levels), he lamented that "the distance is […] great between our present psychological knowledge of the learning and choice processes and the kinds of knowledge needed for economic and administrative theory" (Simon, 1955, p. 100).

Since Simon's early critique of *homo economicus*, there have been substantial gains in research knowledge about the psychological processes underlying judgments and decisions. Psychologists have developed a number of promising theories that describe how flesh-and-blood decision makers with physiological constraints can make good decisions in uncertain worlds. Research on heuristic decision-making strategies such as the recognition heuristic (Goldstein &

Gigerenzer, 1999) and take-the-best (Gigerenzer & Goldstein, 1999) shows how people can make good decisions based on limited information by taking advantage of certain environmental contingencies. Cognitively grounded judgment models such as the Naïve Sampling Model (Juslin et al., 2007) explain how people make judgments from exemplars learned from past experience and stored in long-term memory. Reinforcement-learning models describe how people update impressions of options over time (e.g., Gonzalez, Lerch, & Lebeire, 2003; Hertwig, Barron, Weber, & Erev, 2004), how unbiased judges can form biased opinions through incomplete feedback (Denrell, 2005), and how people select strategies for a given task (Rieskamp & Otto, 2006). Information foraging models explain how organisms can adaptively search for information in their environments given limited time and potential search costs (e.g., Stephens, Brown, & Ydenberg, 2007).

My goal in this cumulative dissertation is to continue these lines of research by adding a small piece to help fill the void left by Simon's destruction of *homo economicus*. Specifically, I try to understand how real people with a limited cognitive architecture can make good decisions in uncertain environments. I address this question in four papers that examine two distinct decision domains: information search in competitive contexts (papers 1 and 2) and the wisdom of the inner-crowd (papers 3 and 4).

From Decisions From Description to Decisions From Experience

All organisms must make decisions between options with unpredictable outcomes. Consider the decision between staying in one's current job or changing to a new job: This decision can be viewed a choice between a "sure thing" – the known happiness of one's current job – and a "risky" second option that can lead to either an increase or a decrease in happiness. Experimentally, decision-making researchers typically represent such options as monetary

gambles (see Figure 1), where each option can result in one or more potential outcomes with explicitly defined probabilities. Importantly, because all outcome and probability information for each gamble is presented to participants, these tasks are known as *decisions from description*.

Option A	Option B
$100 with probability 1.0	–$50 with probability .50
	$150 with probability .50

Figure 1: The "drosophila" of judgment and decision making research: two options represented as monetary gambles.

Since at least the 1970s, the simple descriptive monetary gamble has served as the "drosophila" of judgment and decision making, driving the majority of basic research in the field. Notably, prospect theory (Kahenman & Tversky, 1979), perhaps the most influential theory of decision making, is primarily tested using monetary gambles.[2] Prospect theory assumes that people begin a decision task by transforming each option's objective outcomes and probabilities using a utility weighting function and a probability weighting function, respectively. People are then assumed to multiply and add these transformed probabilities and utilities to calculate an expected utility for each option, and then choose the option with the highest expected utility. Among prospect theory's many predictions, which are now widely accepted in decision-making research, is that people will overweight outcomes with small probabilities and underweight outcomes with large probabilities (Kahneman & Tversky, 1979).

[2] Although prospect theory has also been applied to non-monetary, affective decisions (Pachur, Hertwig, & Wolkewitz, 2013; Rottenstreich & Hsee, 2001).

However, despite its popularity, prospect theory has a major shortcoming in that it falls prey to the false assumptions of *homo economicus* exposed by Simon (1955). Namely, prospect theory assumes that decision makers make decisions from description (Hertwig et al., 2004), where they have perfect population-level information about all options, their associated outcomes, and their probabilities. This applies to choice domains from monetary gambles to mate selection. In deciding between potential mates, prospect theory must assume that decision makers know all possible outcomes and probabilities associated with each candidate. But, of course, people in the real world do not walk around with outcome and probability labels floating above their heads, as the gambles in Figure 1 would suggest. In the real world, decision makers must navigate a world of *uncertainty*, where the set of outcomes and their associated outcomes and probabilities are *a priori* unknown.

Decisions from experience. Thankfully, organisms have a tool that can help them make better decisions in an uncertain world: exploratory *search*. People can go on dates before proposing marriage; hermit crabs can examine shells before deciding whether or not to make a move. Decisions such as these that require pre-decisional information search are known as *decisions from experience* (Hertwig et al., 2004).

How does research on decisions from experience depart from the decisions-from-description paradigm depicted in Figure 1? In the laboratory, decision-from-experience tasks are typically represented as two or more opaque options (i.e., urns or boxes) on a computer screen, where each option represents an *a priori* unknown probability distribution of outcomes. In the sampling paradigm, participants are invited to learn about these options by drawing sequential random samples from the options' underlying distributions at no financial cost. When players

decide they are ready to make a final choice, they indicate which option they want and obtain a real monetary outcome from their chosen option.

Early research on decisions from experience found that the impact of information search during decision making can turn key results from the decisions-from-description literature upside-down. Consider the role of rare events: Prospect theory assumes that people apply a non-linear weighting function to probabilities, which leads them to overweight small probability events. However, when people choose between gambles in a decisions-from-experience paradigm, they behave as if they *underweight* rare outcomes (Hertwig et al., 2004). Hertwig et al. explained this reversal from prospect theory by showing that, in decisions from experience, people do not explore options long enough to discover rare events, and thus make decisions as if the rare outcomes do not exist. This result can explain real-world examples of people apparently underweighting rare events. One striking case given by Hertwig (in press) is the housing decisions of people living in the shadow of Mount Vesuvius in the gulf of Naples, Italy. Despite repeated warnings by experts that the volcano is increasingly likely to erupt with each passing year, residents in the area firmly reject the idea that they are in real danger and refuse to move. Research on decisions from experience suggests that because the residents of Naples have never experienced an eruption in their lifetime, they act as if the threat of such a rare event did not exist.

Clearly, information search strategies and the experiences they reveal play a critical role in how people make decisions under uncertainty. So what do we know about the processes underlying decisions from experience? To guide the reader through a brief summary of the literature, I present a conceptual model of the critical steps underlying decisions from experience

in Figure 2. The model assumes that people follow a cycle of three stages: Information Search, Impression Updating / Comparison, and Stopping Decision.

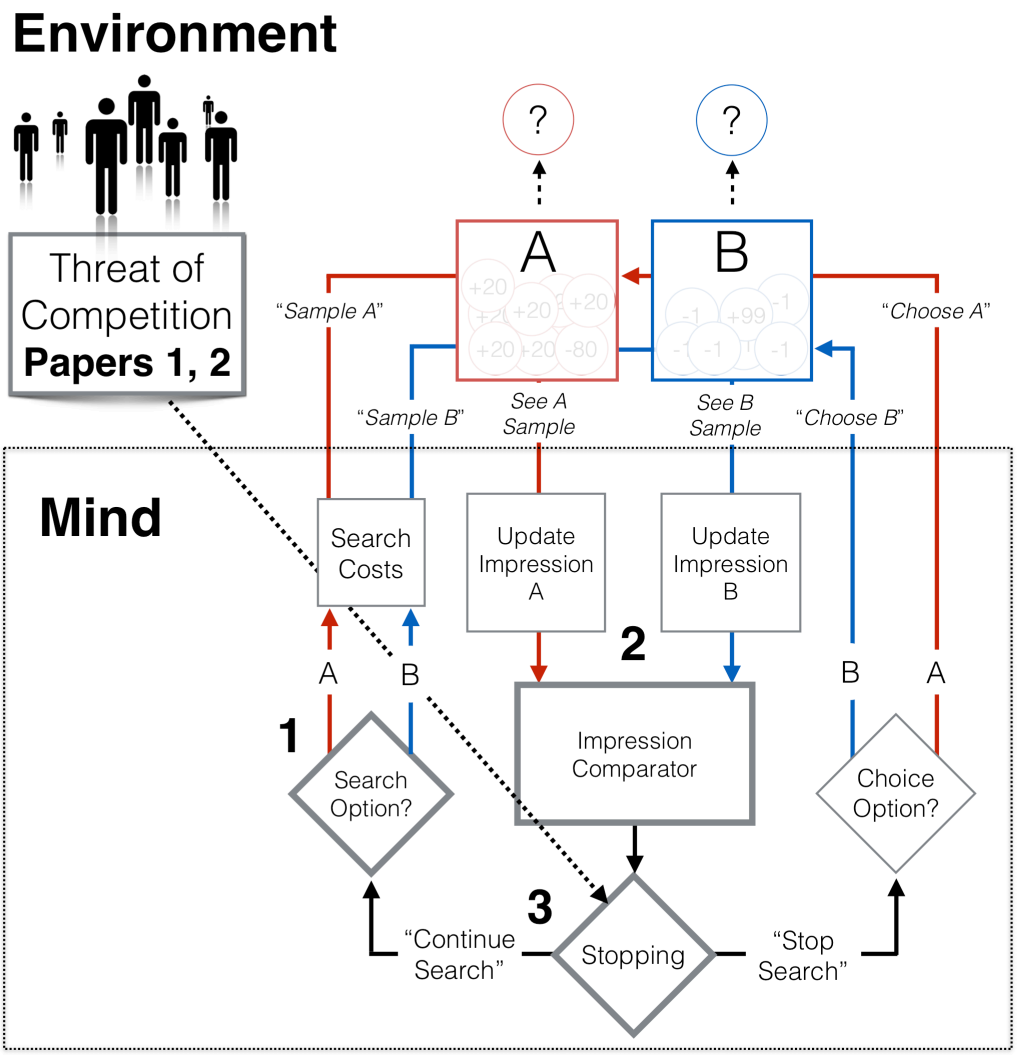


Figure 2: Conceptual model of search, impression formation, and decision making in decisions from experience. There are three key stages: (1) Information Search, (2) Impression Updating / Comparison, and (3) Stopping Rule. Decision makers cycle through these stages until the stopping rule is satisfied.

The model depiction in Figure 2 is split into a top part representing an external environment and a bottom part indicating the mind of the decision maker. A decision maker begins in the "Search" stage, where she decides which option to sample. Research suggests that most people are approximately equally likely to sample from either option (Wulff & Hertwig, 2014), but perhaps with some tendency to sample more from the option with the higher sample variance (Lejarraga, Hertwig, & Gonzalez, 2012). Engaging in search incurs some cost for the decision maker. While search costs may not be explicit (i.e., direct monetary costs), the model assumes that most if not all search incurs some cost, perhaps in the form of opportunity costs or lost time (Hertwig et al., 2014).

After deciding where to search, the decision maker observes a sample from the selected option and updates her impression of that option in the Impression Updating stage. Three impression-updating models are popular in the literature. The *value-updating* model (Hertwig et al., 2004) is a reinforcement-learning model which states that people update their impressions as a weighted average of their prior impressions and new information with a parameter that can capture recency and primacy effects. The natural mean heuristic (Hertwig & Pleskac, 2010) assumes that people store all experienced outcomes and use the running overall mean as the current impression. Finally, the *instance-based learning model* (IBL; Gonzalez et al., 2003) assumes that people store each unique outcome from an option as an "instance," and that these become more activated with each repeated occurrence but also decay over time. Under the IBL, impressions of options are formed as the blended (i.e., weighted average) of the instances of an option weighted by their probability of retrieval.

After updating their impression of the sampled option, decision makers compare their impressions of each option. In the Stopping Decision stage, they then decide whether or not to

stop search by referring to their stopping rule. If the organism decides to continue search, it will return to the Search stage and repeat the cycle. If it decides to stop search, it will leave the cycle and consume (i.e., choose) an option. More research is needed into how people select stopping rules. Several normative and descriptive stopping models have been proposed that assume that stopping decisions are a function of both search costs and the current state of the impression comparison (Busemeyer & Rapoport, 1988). Other models suggest that organisms stop search when their working memory capacity has been filled (Hertwig, in press). Moreover, a recent meta-analysis of papers on decisions from experience has found evidence for consistent individual differences in stopping rules, suggesting that different people accrue different costs during search (Wulff & Hertwig, 2014).

How does competition affect search rules in decisions under uncertainty? Papers 1 and 2 address a critical real-world search cost that has previously been ignored in the decisions-from-experience literature: the threat of *competition*. In many real-world environments, from mate selection to housing choice, organisms search for information in the presence of competitors who can consume options during the search process. While it seems clear that organisms should reduce search in the presence of competition, the *extent* to which the presence of competition should affect exploration–exploitation trade-offs remains unclear. Further, is it always better to be faster than an opponent, or is it sometimes better to be more patient and allow competitors to choose first? What characteristics of the environment affect optimal search strategies in the presence of competition? Descriptively, how do organisms jointly manage uncertainty in their natural (i.e., options) and social (i.e., degree of competition) environments?

To answer these questions, we created a new task called the Competitive Sampling Game (CSG). In this game, two players have the option to repeatedly sample from options before

making a choice on a first come, first served basis. In paper 1, we introduce this game and use both mathematical and agent-based simulation analyses to generate predictions for how people should make decisions in the task and test those predictions in an experiment. We find that real participants do indeed drastically reduce their search in the presence of competition.

Paper 1: Rivals in the dark: How competition influences search in decisions under uncertainty

Phillips, N. D., Hertwig, R., Kareev, Y., & Avrahami, J. (2014). Rivals in the dark: How competition influences search in decisions under uncertainty. *Cognition, 113*(1), 104-119.

Organisms in the real world must frequently make decisions under uncertainty, where the set of possible outcomes and the qualities associated with each outcome are *a priori* unknown. To shed light on these options, including their potential outcomes and probabilities, organisms can engage in a sequential search process. For example, humans can go on dates with potential mates, and bees can sample flowers in a patch. Because search in the real world can be costly in terms of time, energy, and missed opportunities, organisms must develop effective strategies that produce valuable information without exacting excessive costs (Pirolli, 2007). To understand how people conduct pre-decisional information search in *decisions from experience*, Hertwig et al. (2004) developed the sampling paradigm, in which decision makers are presented with several *a priori* unknown options (i.e., gambles) and can learn about them by drawing random samples. Although the sampling paradigm is certainly a better model of many real-world decisions than is the decisions-from-description paradigm, it still ignores a critical real-world search cost: the possibility of competitors consuming good options during search. For example,

the longer someone spends looking for a flat in a new city, the more likely other flat-hunters are to snap up good options. Moreover, because a flat-hunter does not know exactly how many other searchers there are and how quickly they can be expected to make a decision, competition presents an additional form of *social uncertainty* on top of uncertainty about options.

How do organisms jointly manage this social uncertainty alongside the uncertainty about options? To answer this question, we developed a new game we call the Competitive Sampling Game (CSG). In the game, two (or potentially more) decision makers conduct a simultaneous search for information about two (or potentially more) choice options (see Figure 3).

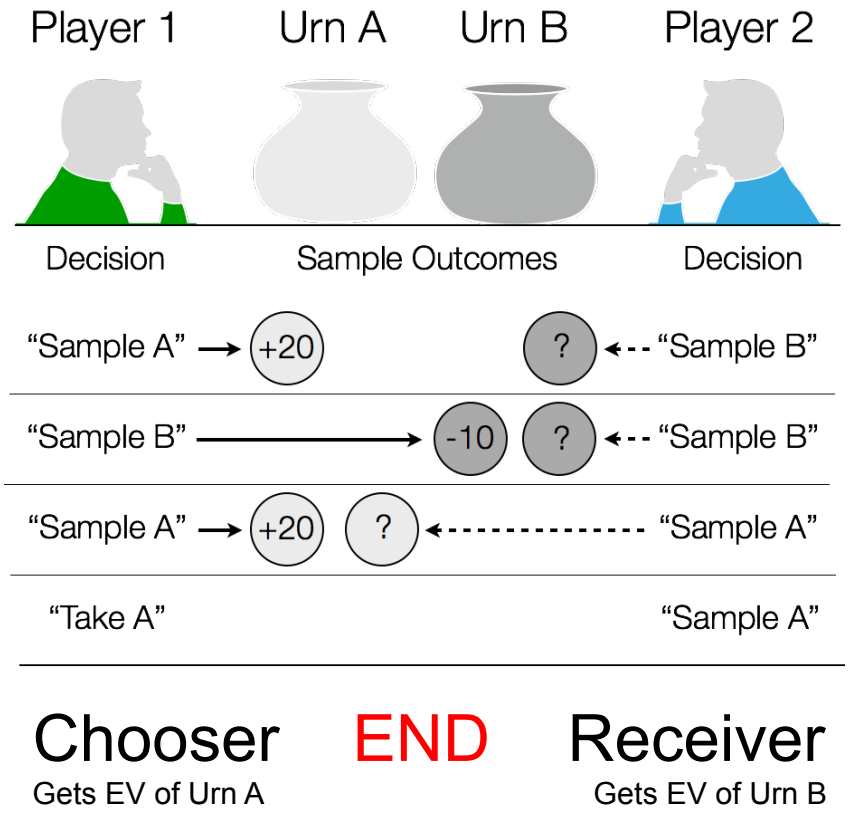


Figure 3: Diagram of the competitive sampling game (CSG). Players draw samples with replacement from urns at the same rate until one player decides to stop and choose an urn. This "chooser" gets the expected value of her chosen urn and the remaining player, the "receiver," gets the expected value of the remaining urn.

The basic rules of the CSG are as follows: Both players begin with one sample from an option of their choosing. They are then both asked if they would like to make a decision or continue sampling. Players continue the cycle of drawing samples as long as both wish to continue. As soon as one player is ready to make a decision, the faster "chooser" takes the option of his or her choice, leaving the remaining option to the "receiver." After each player obtains an option, they are rewarded with the expected value of the option's underlying probability distribution.

The CSG forces players to make a trade-off between information quality and decision speed: The more samples a player takes, the better he can estimate the options' long-term gains and the better his expected decisions will be; however, with more samples comes a higher chance that the other player will stop the game and take the better option. The CSG shares the competitive aspect of existing games of timing (Dutta & Rustichini, 1993), where two players independently decide whether or not to take an existing reward or to wait for a larger reward in the future. However, the CSG departs from these games because players must deal with uncertainty with regard not only to their social environment, but also to the choice environment. In the CSG, rewards do not increase over time by a well-defined rule (as is the case in existing games of timing); instead, what increases is the quality of choice-relevant information.

To generate normative benchmarks for the game, we used mathematical analyses and simulation methods to compare the performance of multiple strategies in the CSG. Because search costs in the CSG depend on social environments, we simulated decision strategies in three social environments representing different degrees of competition speed. We found that an omnipotent decision maker who knows exactly how long her competitor will search should take

one fewer samples than her competitor. However, if a player does not know exactly how long her opponent will search (as is the case in virtually all real-world competitive tasks), then the player should dramatically reduce search relative to her expectations of her opponent's search length. In other words, we found that it is much better to under-estimate an opponent's search length, by any amount, than to over-estimate an opponent's search length by even one sample. Moreover, we found that the only sampling strategy that guarantees obtaining the best option in no less than 50% of cases is a one-sample strategy we call the "take-good-enough, otherwise-shift" heuristic. These results suggest that participants should not only qualitatively reduce search in competitive contexts, but that they should reduce it to almost minimal amounts.

To see how much people actually reduce search in competitive contexts, we had groups of university students play the competitive sampling game for real monetary rewards. Consistent with our normative predictions, people playing in competitive contexts dramatically reduced their search – from a median of 18 in a solitary condition, to just 1 under competition. In terms of rewards, while fast choosers in the competitive condition obtained fewer rewards than solitary participants, they nonetheless outperformed slower receivers in the competitive condition. These results are consistent with our normative analysis and show that people are willing to dramatically reduce their exploration levels in the presence of competition.

We expanded these analyses by simulating the effects of different statistical environments that can either reward or punish minimal search. We show that one-sample decision-making using the "take-good-enough, otherwise-shift" heuristic will only be beneficial in environments that satisfy strict distributional criteria. We call these "one-sample friendly" environments and prove their necessary criteria (see Appendix of Manuscript 1). Next, we show why environments involving extremely rare events should dramatically shift the benefits of fast

search in competitive contexts. Because competition dramatically reduces search, we expect that people making fast decisions under competition will tend to grossly underestimate rare events (see Hertwig et al., 2004) and choose options that appear good in the short term, but have potentially large losses in the long term.

In paper 2, we use the CSG to explore how competition affects the evolution of decision strategies. We conduct a series of simulations where we evolve agents using different decision strategies in environments with varying levels of competition. Consistent with the results of paper 1, we find that – relative to agents in less competitive environments – agents in highly competitive environments evolve decision strategies that rely on far less information. Moreover, consistent with the fast and frugal heuristics program (Gigerenzer, Todd, & the ABC Research Group 1999), these results suggest that evolution favors decision speed and satisficing over absolute estimation accuracy.

Paper 2: The Janus face of Darwinian competition

Hintze, A., Phillips, N. D., Adami, C., & Hertwig, R. (2014). The Janus face of Darwinian competition.

Darwinian evolution is driven by competition. Organisms that successfully out-compete others in obtaining resources, from food to mates, will replicate their genes in the next generation, while those that fail in competition risk their genetic future. In the process, competition forces sequential generations to become better adapted to their environments. While this basic functional role of competition is a foundation in biology, it is unclear how different degrees of competition shape the cognitive capacities and decision strategies of organisms. In other words, how do information search and decision strategies evolve in environments with high versus low levels of competition? Will the cognitive architectures evolved in highly competitive

environments be systematically more complex or more accurate than those evolved in less competitive environments?

In this paper, we provide initial answers to these questions using an evolutionary simulation. In this simulation, successive generations of agents play versions of the CSG (see Phillips, Hertwig, Kareev, & Avrahami, 2014) in which each agent's fitness is defined as its outcomes in the game. Importantly, in contrast to the original version of the CSG, in this task we assign a fully realized reference option to each agent at the beginning of each game. Thus, agents are repeatedly faced with the decision of choosing the reference option, sampling from an unknown option, or choosing the unknown option. In the simulation, we distinguish between three kinds of competitive environments. In the *indirect competition* environment, each agent plays the CSG with two options. Agents play alone, and each agent's decision does not affect the choice environment of other agents. In this environment, competition happens at the population level and not at the level of individual agents. In the *direct competition* environment, agents play the CSG with three options in pairs, and each agent's decision *does* affect the choice set of other agents. Here, an agent's competitors can consume desirable options and remove them from the agent's choice set. Finally, in the *extreme competition* environment, we further increase the level of competition from the direct competition environment by reducing the number of options from three to two.

Across environments, agents evolved a high probability of choosing the reference option when the sampled difference between the reference urn and the sampled option was high. Additionally, the more variability in option outcomes (defined as the variance of option distributions), the more samples agents drew before making a decision. This result suggests that, in contrast to Phillips et al. (2014), where we assumed agents use a "fixed-N" sampling size rule,

evolution can be expected to drive organisms to dynamically adjust their sampling rules as a function of the variability of sample outcomes. We found that agents evolved dramatically different search strategies in the three competition environments. The more competitive the environment was, the less agents sampled the unknown option and the more likely they were to choose the reference option. Interestingly, the effect of option variability on search length vanished in the extreme competition environment. Here, agents evolved to make a decision after a single sample, regardless of the variability in option outcomes. In other words, in extremely competitive environments, organisms should evolve strategies that trade estimation accuracy for fast choosing. With regard to payoffs, we found that agents in the indirect competition environment evolved strategies that led to almost perfect choice performance. This is because low competition allows organisms to sample extensively and obtain very accurate estimates of option values before making a choice. In contrast, agents in the extreme competition environment chose so quickly that they obtained payoffs at near-chance level.

In conclusion, the results from this evolutionary simulation confirm the key insight from Phillips et al. (2014) that increased levels of competition should dramatically decrease search efforts in decisions under uncertainty. They further indicate that evolution should drive organisms in competitive environments to be less sensitive to outcome variability (i.e., uncertainty) than organisms in solitary environments. This means that extreme competition can inhibit the evolution of decision strategies that are sensitive to differences in environmental uncertainty.

Summary of Papers 1 and 2. These two papers add to the growing body of research on decisions from experience by introducing the CSG, a novel variant of the decisions-from-experience paradigm that allows researchers to measure the effects of competition on

information search rules. As we describe in our papers, we believe that our results have direct implications for how entities from hermit crabs to pharmaceutical companies make decisions. As a final example, consider the case of the drug Vioxx. On May 20, 1999, the US Food and Drug Administration (FDA) approved the anti-inflammatory drug Rofecoxib under the brand name Vioxx. This drug gained immediate and widespread acceptance from doctors and patients, many of whom found it the only way to get relief from arthritic pain. Doctors prescribed Vioxx to over 80 million people, and the drug quickly generated over $2.5 billion in sales revenue for its marketer Merck & Co ("Rofecoxib," 2014). However, after less than five years on the market, it became apparent that Vioxx presented a rare risk of heart attack and stroke from long-term, high-dosage use. In September 30, 2004, Merck withdrew Vioxx from the market and the company was forced to set aside $4.85 billion for legal claims from patients ("Rofecoxib," 2014). While we do not know for certain whether Merck rushed the drug to market without adequate testing, our results suggest that the highly competitive pharmaceutical industry may have pushed Merck to "bet" too quickly on a drug with short-term gains, but long-term losses. Perhaps if the pharmaceutical industry were not so competitive, companies would spend more time developing and testing drugs before bringing them to market.

In papers 1 and 2, we described how people could use external search to improve their outcomes in decisions under uncertainty. In the next section, we shift to another possible approach: harnessing the wisdom of crowds within one mind by means of dialectical bootstrapping. But before we begin, let us take a trip back to 1906, when a scientist at a fair made a notable discovery that still resonates today.

The wisdom of the inner-crowd and dialectical bootstrapping

Sir Francis Galton (1822–1911) was a prolific scientist who published over 340 papers and books in myriad fields including statistics and anthropology ("Francis Galton," 2014). Despite (or perhaps as a result of) his education, Sir Francis Galton did not believe in democracy. For Galton, the general population could not be trusted to make important election decisions. To test his beliefs, he took advantage of an opportunity that presented itself at a livestock fair he attended in 1906. The fair had a prediction contest where fairgoers were asked to estimate the weight of an ox. Each fairgoer wrote his or her estimate on a piece of paper and placed it in a jar. After everyone had made their predictions, Galton gathered the estimates and compared them to the actual weight of the ox. Having expected the estimates of these lower class fairgoers to be highly inaccurate, he was shocked to find that the group as a whole was incredibly accurate: While the true weight was 1,198, the group-mean estimate was 1,197 pounds, an error of only one pound!

Galton's discovery was not a fluke. Rather, it is a classic example of what is now known as the *wisdom of the crowds*. The wisdom of the crowds describes scenarios in which the aggregate (usually arithmetic mean) estimate of a group outperforms even the most accurate estimate of a single member of the group. The wisdom of crowds is a rather straightforward mathematical implication of error cancellation. By way of illustration, let us consider the case of two media figures Dick Morris (a Republican but former adviser to President Bill Clinton) and Jim Cramer, host of CNBC's television show "Mad Money." In November 2012, these two men each predicted the electoral college outcome of the upcoming 2012 US Presidential election (Figure 4):

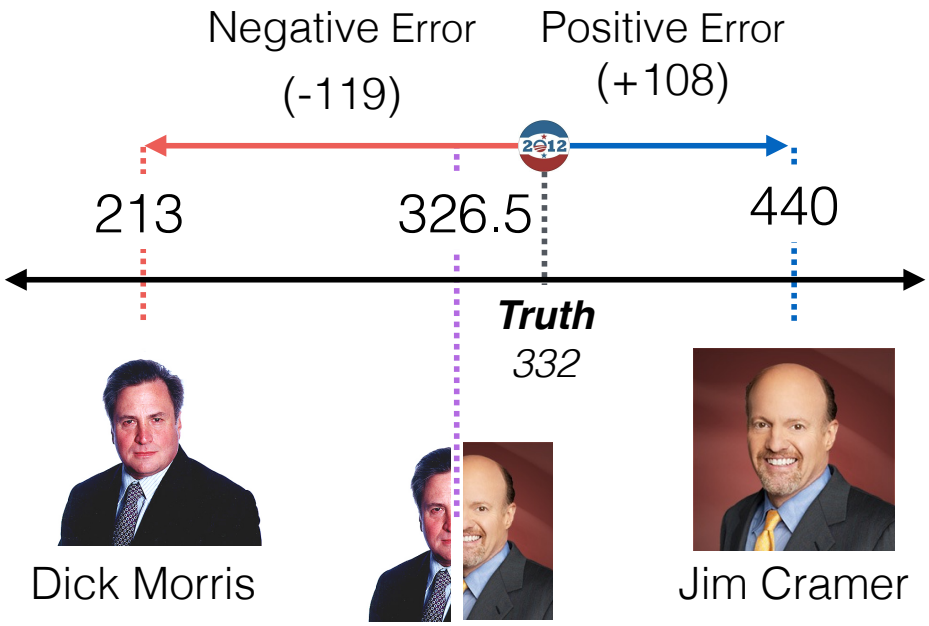


Figure 4: Example of error cancellations in aggregated judgments. The average estimate of the number of votes for Obama is much closer to the truth than is either judge's original estimate.

Dick Morris challenged existing poll results and projected a landslide victory for Romney, predicting that Obama would win only 213 of a necessary 270 votes (Morris, 2012). Jim Cramer made an equally bold prediction but in favor of Obama, projecting a 440-vote victory for the incumbent (Fitzgerald, 2012).

On the evening of November 6, Obama won the Presidential election with 332 electoral college votes. Looking back at Morris's and Cramer's predictions, both were way off the mark: Morris underestimated Obama's votes by 119 (|213 – 332|) and Cramer overestimated by 127 (|440 – 332|). On average, the two pollsters were off by an astounding 123 votes. However, consider the accuracy of the *average* of their two predictions: their average estimate of 326.5 deviated from the true value by only 5.5 votes! How is it possible that the average estimate is so

much more accurate than the prediction of either individual pollster? The reason is error cancellation – because the two estimates had large opposing errors (one positive and one negative), their average was much closer to the true answer than either individual estimate.

The inner-crowd. While extensive psychological research has been conducted on the wisdom of crowds since the 1970s (e.g., Hogarth, 1978), researchers have only recently applied the phenomenon to an individual judge's *inner-crowd* (Herzog & Hertwig, 2009; Vul & Pashler, 2008). In Figure 5, I present a conceptual overview of the processes underlying the inner-crowd.

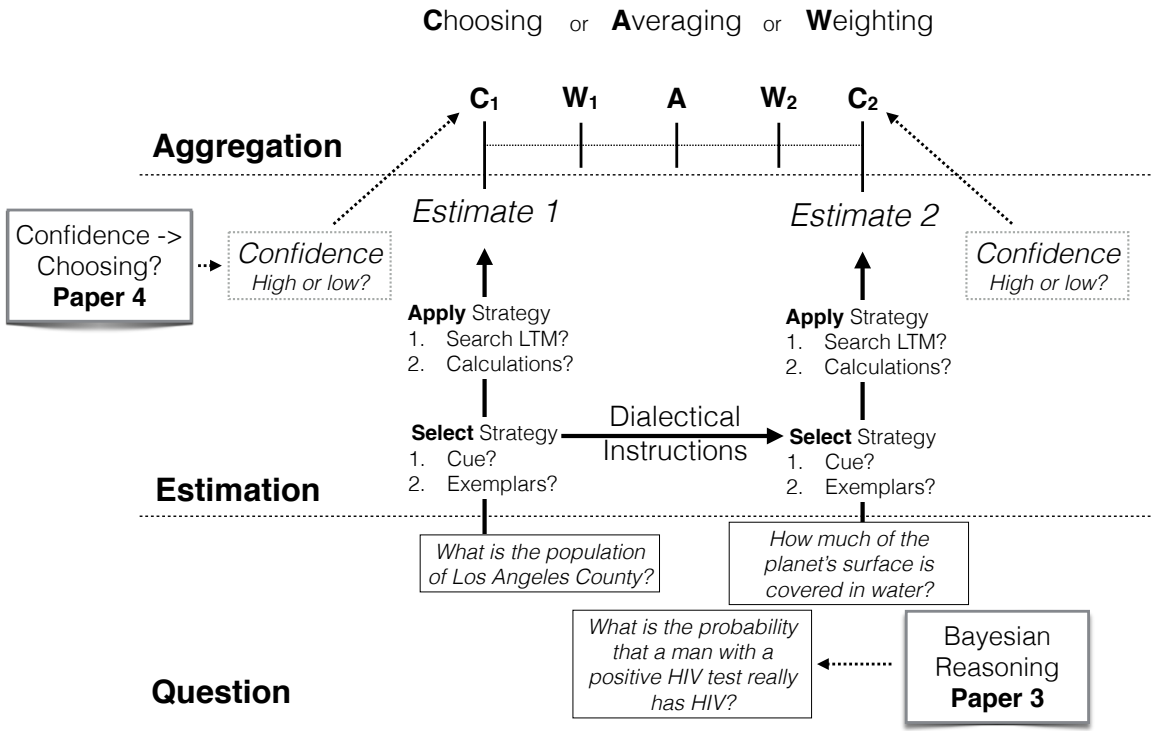


Figure 5: Conceptual model of processes underlying the inner-crowd. In paper 3, we explore the benefits of the inner-crowd on Bayesian reasoning judgments. In paper 4, we test how people can outperform the simple average of their inner-crowd by using confidence-based choice.

The inner-crowd works as follows: After viewing a question, judges follow three phases. In phase 1 (left side of Figure 5), they form an initial estimate. This estimation phase contains two parts: strategy selection and strategy application. In the strategy selection part, the judge decides both what kind of estimation strategy to use and which information to use within the chosen strategy (e.g., which cues to use, which exemplars to search for). Previous inner-crowd research has provided only a general statistical description of the estimation process by describing strategies as a combination of truth, bias, and random error (Herzog & Hertwig, 2009). In papers 3 and 4, we take a theory-based approach to understand the strategies that people use. In paper 3, where judges are faced with a probabilistic estimation task, we assume that judges select one of many simple intuitive strategies. In paper 4, where judges estimate county populations based on binary cues, we assume that judges follow an exemplar-based model of estimation (specifically, Juslin et al.'s, 2007, Naïve Sampling Model) and select a probe cue to search long-term memory.

In the second, *application* part, judges apply their strategy to derive final estimates. For the probability estimation task described in paper 3, this would mean following the arithmetic calculations dictated by the selected strategy (i.e., averaging the base-rate and hit-rate). For the county population estimation task in paper 4, this could mean searching long-term memory for examples of counties that are similar to the target specified in the question (e.g., what counties do I know that are similar to Los Angeles county?), and then using the populations of those examples to estimate the population of the unknown target.

Dialectical bootstrapping. After deriving their first estimate, judges either immediately begin a second estimation phase or undergo an intervention (such as a time delay; Vul & Pashler, 2008) designed to increase their estimate diversity. In our research, we focus on *dialectical*

bootstrapping, a method of increasing the diversity of the inner-crowd (Herzog & Hertwig, 2009, 2014a, 2014b). In dialectical bootstrapping, judges are explicitly encouraged to generate second, *dialectical* estimates that have differently signed errors than their original estimates. To this end, dialectical estimates should rely on different knowledge, assumptions, or strategies than the initial estimates do. Going back to the U.S. Presidential election example, if a person's first estimate looks like Dick Morris's, her dialectical estimate should look like Jim Cramer's. While dialectical bootstrapping does not specify a single dialectical intervention, Herzog and Hertwig have typically used Lord, Lepper, and Preston's (1984) *consider-the-opposite* technique, which encourages judges to think of reasons why their first estimates may have been wrong, and to derive new estimates on the basis of those considerations. In several studies, Herzog and Hertwig (2009, 2014a, 2014b) have found that consider the opposite instructions do indeed decrease signed error correlations, and increase subsequent averaging gains, between first and second estimates relative to control groups. However, the specific process underlying this change in signed errors has not been studied. In papers 3 and 4, we explicitly model the effects of dialectical instructions on strategy change and averaging gains in two new estimation paradigms. In paper 3, we apply dialectical bootstrapping to a Bayesian reasoning paradigm and explore how dialectical instructions change strategy use and subsequent averaging gains in different statistical environments. We propose that people can use dialectical bootstrapping to combine multiple, non-Bayesian algorithms to become more Bayesian without any knowledge of Bayes theorem.

Paper 3: How the inner-crowd can help non-Bayesians become more Bayesian

Phillips, N. D., Herzog, S., & Hertwig, R. (in prep). How the inner-crowd can help non-Bayesians become more Bayesian.

Are people intuitive Bayesians? Bayes theorem states that people should update their probabilistic beliefs by integrating base-rates (prior probabilities) with new information in the form of a hit-rate (the likelihood of an event given a hypothesis) and a false-alarm rate (the likelihood of an event given an alternative hypothesis). In the 1960s, Ward Edwards claimed that people rely too much on base-rate information and do not sufficiently update their beliefs in the presence of new information (hit-rates and false-alarm rates). However, they are close enough to the Bayesian norm to be labeled "conservative Bayesians" (Edwards, 1968). Fast-forward to 1972 at the start of the heuristics-and-biases movement, and it seems readers have been bamboozled. According to Kahneman and Tversky (1972), not only are people not "conservative Bayesians," they are "not Bayesian at all." In particular, Kahneman and Tverksy suggested that people routinely violate normative rules by using a "representativeness heuristic" that completely ignores base-rate information.

Since the 1970s, evidence has been accumulating that people cannot be simply categorized as "conservative Bayesians" or "not Bayesian at all." Instead, people apply a variety of simple strategies that use and combine statistics in different ways (McKenzie, 1994). Moreover, depending on the statistical environment to which they are applied, strategies have different biases relevant to Bayes theorem: some tend to have positive biases (over-estimate), others to have negative biases (under-estimate). Against this background, can people improve their Bayesian reasoning judgments by harnessing an inner-crowd of non-Bayesian strategies? In other words, can judges become more Bayesian by being both "conservative" (i.e., relying too

much on base-rates) and "not Bayesian at all" (i.e., ignoring base-rates) in one mind? Moreover, can *dialectical bootstrapping* increase strategy diversity and subsequent averaging gains?

Dialectical bootstrapping is a method of boosting the benefits of the inner-crowd (Herzog & Hertwig, 2009). The procedure works by having a judge generate both an initial estimate to a problem and a second, *dialectical* estimate based on different considerations and assumptions. Judges then combine their estimates by averaging them into a single, final estimate. Previous research suggests that dialectical bootstrapping increases estimate diversity and subsequent averaging gains (Herzog & Hertwig, 2009, 2014a, 2014b). However, no study has directly modeled discrete strategy change in the inner-crowd framework. Thus, it is unclear to what extent dialectical bootstrapping produces qualitatively different strategy use in one mind. In this paper, we directly measure strategy use and strategy change in the Bayesian reasoning task in a simulation study followed by two experiments.

In a simulation study, we took simple strategies proposed in the literature (Gigerenzer & Hoffrage, 1995; McKenzie, 1994) and simulated their solitary performance relative to Bayes in two different environments. In the "Valid Cue" (VC) environment, cue values ranged across the entire probability space, with the added restriction that false-alarm rates were smaller than hit-rates. This is a highly uncertain environment in which it is difficult to predict cue values in advance. In the "Rare Event plus Valid Cue" (RE+) environment, base-rates were small, hit-rates were large, and false-alarm rates were small. This environment represents important real-world environments in which a judge predicts the probability of a rare event (e.g., a rare disease) based on a cue with a large hit-rate and a small false-alarm rate (e.g., a medical test). In addition to being of interest for domain-specific reasons (e.g., medical reasoning), this environment is known to produce large opposing biases in simple strategies that ignore one or more cues

(McKenzie, 1994). If a person combines strategies with different biases in this domain, they could potentially reap large averaging gains.

The key results from our simulation were as follows: First, strategies had much larger biases in the RE+ than the VC environment. Second, consistent with our predictions, averaging gains were consistently higher in the RE+ than the VC environment. Finally, averaging two strategies led to larger averaging gains when the two strategies have different base-rate usage (i.e., one uses base-rates and one ignores base-rates) than when they have the same base-rate usage. These results suggest that if people construct multiple strategies and combine them, they stand to improve their accuracy – especially in RE+ environments.

We tested our simulation results in two online experiments. In both experiments, participants were first presented with several Bayesian reasoning problems that required them to estimate the probability of an event given information on the base-rate, hit-rate, and false-alarm rate. In study 1, the problems were vignettes taken from Gigerenzer and Hoffrage (1995). In study 2, the problems were given in a standardized "boxes and balls" format. All participants then gave a second set of estimates to each problem. In control conditions, participants were asked to estimate again as if they were seeing the problem for the first time. In dialectical conditions, participants read "consider-the-opposite" instructions that directed them to actively construct a new strategy.

We used statistical modeling techniques (Lewandowsky & Farrell, 2010) to classify the strategy each participant used in each estimate phase. Using these classifications, we could determine whether or not a participant used the same strategy in both phases (possibly with different application errors) or used a categorically new strategy.

Our key results were as follows: First, participants given dialectical instructions were substantially more likely to change strategies than were control participants. This is the first analysis to demonstrate this effect. Second, participants who switched strategies between first and second estimates reaped larger potential averaging gains than did participants who maintained the same strategy with random error. The effect was even stronger when participants switched between a strategy that did not use base-rates and one that did use base-rates. These findings demonstrate the importance of strategy diversity in averaging gains. Finally, consistent with our simulation results, the effect of strategy switching on averaging gains was larger in RE+ environments than in VC environments. This suggests that domains with extremely rare events, where people traditionally do poorly relative to Bayes, are especially conducive to averaging.

In summary, we find that people can indeed improve their estimates in Bayesian reasoning tasks by using dialectical bootstrapping. Importantly, they can do this without any training in Bayesian reasoning and without change in stimuli formats (e.g., Gigerenzer & Hoffrage, 1995). While we do not suggest that dialectical bootstrapping is the best solution to "fixing" errors in a Bayesian reasoning task, our results do show that dialectical bootstrapping is an effective method for people to improve their estimates in tasks where the optimal rule is unknown.

Confidence in the inner-crowd: Can choosing outperform the average?

Most prior research on both the wisdom of crowds and the inner-crowd has tested the accuracy of the average (i.e., simple mean) of the crowd relative to individual judgments (e.g., Herzog & Hertwig, 2009, 2014a, 2014b; Vul & Pashler, 2008). However, both advice-taking (Soll & Larrick, 2009) and wisdom-of-crowds (Surowiecki, 2004) research has identified cases where the average can be beaten by strategies that favor one estimate over another. Specifically,

when estimates are highly correlated (bracketing rates are low) and one estimate is much more accurate than the other, it can be better to choose the more accurate estimate than to take the mean (Soll & Larrick, 2009). Are there cases in the inner-crowd where these conditions hold and people can outperform averaging by choosing a single estimate, and could confidence be the key?

Confidence has gotten a bad rap in cognitive psychology. A long history of confidence research with both laypeople and experts (notably physicians; Christensen-Szalanski & Bushyhead, 1981) suggests that people are much more confident in their judgments than is empirically warranted.[3] However, despite reliably finding data indicative of overconfidence, researchers have also consistently found that confidence is a valid cue to accuracy (e.g., Winkler, 1971; Yaniv & Foster, 1997; Yates, 1990). In other words, although people are generally overconfident, the *more* confident they are in their estimates, the more accurate their estimates will be. No previous research has tested the accuracy of confidence judgments in the inner-crowd context. If high-confidence estimates tend to be more accurate than low-confidence estimates, then *choosing* high-confidence estimates could potentially outperform simple averaging. However, if confidence is weakly related to accuracy, then choosing could backfire and lead to worse performance than simple averaging. In paper 4, we propose that people can use confidence ratings to boost the effects of a choosing strategy in their inner-crowd.

Paper 4: Confidence and Dialectical Bootstrapping Facilitates Choosing in The Inner-Crowd

[3] But see the debate on whether overconfidence is a true phenomenon or the result of improper measurement (e.g.; Erev, Wallsten & Budescu, 1994).

Phillips, N. D., Herzog, S., Kämmer, J., & Hertwig, R. (in prep.). Confidence and Dialectical Bootstrapping Facilitates Choosing in The Inner-Crowd.

A growing body of research has shown that people can improve their judgments by harnessing a diverse inner-crowd (Herzog & Hertwig, 2009, 2014a, 2014b; Vul & Pashler, 2008). In the same way as groups of diverse individuals can produce an average judgment that outperforms that of even the best individual member, a single person can benefit from combining multiple estimates drawn from a diverse pool of internal information and strategies.

Previous studies on the inner-crowd have focused on the accuracy of the average (arithmetic mean) of an individual's inner-crowd. There is good reason for this; the arithmetic mean is an elegant combination rule that benefits from error cancelation amongst diverse estimates with opposing errors. Advice-taking research has shown that trying to "chase the expert" by choosing one estimate and ignoring the rest outperforms taking the arithmetic mean only when three strict criteria are met: (a) Errors must be highly correlated (i.e., bracketing rates must be low), (b) the accuracy of one source (i.e., advisor) must be substantially higher than that of the other, and (c) the most accurate source must be easily detected. If any of these conditions fail, chasing the expert will lead to poorer performance than taking the simple mean. If all conditions hold, however, chasing the expert can be the better strategy. Can these conditions be satisfied in the inner-crowd? If so, is confidence the key?

On the one hand, relying on confidence judgments to improve estimation seems like a fool's errand. While people trust high-confidence advisers more (Sniezek & Van Swol, 2001) and give more weight to advice from advisers with high confidence than those with low confidence (Yaniv, 2004), many researchers argue that confidence judgments are notoriously

inaccurate (Glaser, Langer, & Weber, 2013; Soll & Klayman, 2004) and the result of a biased information processing system (e.g., Klayman, Soll, González-Vallejo, & Barlas, 1999). However, a counter-movement suggests that this picture of the systematically biased judge is wrong. Rather than being biased information processers, humans might be "naïve intuitive statisticians" (Fiedler & Juslin, 2006; Juslin et al., 2007) who are unbiased in their assessment of sample information, but naïve with respect to potential sampling biases that might make their sample unrepresentative of the respective population. Consistent with the naïve intuitive statistician idea, there is evidence that confidence judgments do contain veridical information about estimates: although people are generally overconfident, confidence *is* a reliable predictor for accuracy (e.g., Winkler, 1971; Yaniv & Foster, 1997; Yates, 1990). In other words, people's (overconfident) high-confidence estimates tend to be more accurate than their (still overconfident) low-confidence estimates. Accordingly, advice-taking research has found that, in some cases, the optimal way to aggregate information from two judges is to use a *maximum confidence slating* heuristic, where the advice from the most confident judge is taken and the advice from the least confident judge is ignored (Koriat, 2012). But will the benefits of high-confidence choosing carry over to the inner-crowd? Are multiple confidence judgments from the same mind sufficiently correlated with accuracy to allow choosing to outperform averaging? Finally, do people actually rely on their confidence judgments when deciding how to aggregate multiple estimates from their inner-crowd?

To measure the descriptive and normative role of confidence in the inner-crowd, we created a cue-based estimation study where judges gave repeated estimates and confidence judgments. In each of 16 questions, judges (i.e., participants in an empirical study and agents in a simulation) estimated the population of a real, but unnamed U.S. county based on four binary

cues. In addition to giving their population estimates for each county, judges gave 90% confidence intervals. After giving initial (phase 1) estimates, judges were assigned to one of three conditions: one control, and two dialectical. Judges in the control condition were told to give a second set of estimates as if they were seeing the questions for the first time. Those in the dialectical "consider-the-opposite" (D-CTO) condition read Lord et al.'s (1984) "consider-the-opposite" instructions. These instructions have been used in previous dialectical bootstrapping research to increase estimate diversity. Finally, those in the dialectical "consider-other-exemplars" (D-COE) condition read a novel set of instructions that we explicitly designed to increase estimate diversity for people using an exemplar-based model of estimation (e.g., Juslin et al.'s, 2007, Naïve Sampling Model, NSM). In phase 2, judges gave a second set of estimates and confidence intervals for each county. Finally, in phase 3, we removed the county cue values and asked judges to make their best estimates based solely on their previous responses from phases 1 and 2.

In order to make predictions for how confidence should be related to accuracy in this task, we conducted an agent-based simulation where agents gave NSM-based estimates for the same stimuli given to our experimental participants. Among other variables, we assigned each agent a long-term memory storage composed of exemplars of US counties, a short-term memory capacity, and an estimation strategy consistent with the three conditions in the study. We had three key simulation results: First, confidence was highly correlated with estimate accuracy. Second, for the majority of agents, choosing high-confidence estimates outperformed averaging. Finally, (simulated) dialectical instructions increased the benefits of high-confidence choosing.

Our empirical results largely coincided with our simulation. While participants were generally overconfident, high-confidence estimates were consistently more accurate than low-

confidence estimates for a majority of participants. Thus, confidence was indeed a valid cue for accuracy. When we compared the estimate accuracy of taking the simple average relative to choosing high-confidence estimates for each participant, we found that high-confidence choosing outperformed both initial estimates and average estimates for a majority of participants. However, for those participants where confidence was unrelated to accuracy, averaging outperformed choosing. Additionally, dialectical instructions reliably boosted the potential benefits of high-confidence choosing. Finally, modeling results suggested that most participants did *not* use confidence-based strategies and instead either used a simple average strategy or chose their first or second estimates. Thus, most participants could have improved their phase 3 estimates by relying on their confidence estimates more than they actually did.

In conclusion, this paper provides new insights into how people do, and should, benefit from their inner-crowd. While previous research on the inner-crowd has emphasized gains from averaging estimates, we find that people can use their confidence judgments, a much-derided measure, to extract better gains from their inner-crowd. Moreover, just as dialectical bootstrapping increases gains from averaging, it can also increase gains from high-confidence choosing.

General Discussion

In 1955, Simon struck down *homo economicus*. In its place, he left the outline of a boundedly rational decision maker who makes "good-enough" decisions given informational, physiological, and psychological constraints that prevent her from applying rational, normative models of decision making. In this dissertation, I have attempted to fill in a small portion of this outline by showing how people can make good decisions in competitive environments (papers 1 and 2) and apply the wisdom of their inner-crowd to Bayesian reasoning (paper 3) and cue-based estimation tasks (paper 4).

Bridging Research Gaps

In this dissertation, I have sought to connect research areas that are typically kept separate. In the process, I show how each area can provide critical insights into the others.

Game theory and psychology. In our work on the CSG and how competition affects decisions under uncertainty, we connect behavioral game theory research (specifically economic games of timing; Dutta & Rustichini, 1993) to psychological models of information search in decisions under uncertainty (Hertwig et al., 2014). In so doing, we highlight shortcomings in each individual approach. While behavioral game theory provides a rich account of how people should (and, to some extent, do) make decisions in competitive tasks (Camerer, 2003), it typically ignores uncertainty at the level of choice options and instead assumes that decision makers have complete knowledge of all options. While psychology and behavioral ecology have produced models explaining how organisms manage uncertainty in real-world decision making through information search, updating, and stopping rules, previous research has failed to address how competition affects decision making under uncertainty. Given the promising initial results

derived from the CSG in this dissertation, I hope that the CSG will become a new tool for behavioral game theorists and psychologists to collaborate on their common questions.

Estimation and group aggregation. In our work applying the inner-crowd to Bayesian reasoning and cue-based estimation, we seek to connect disparate lines of research on estimation, on the one hand, and group aggregation (wisdom of crowds and advice-taking), on the other. Previous work on estimation has explored how people use cue- and/or exemplar-based strategies to derive best estimates and confidence intervals. However, in this research context, the questions stop once solitary estimates have been made. In Bayesian reasoning research, different groups have claimed that people are either "conservative Bayesians" or "not Bayesian at all." Yet, this research ignores that people, from groups to individuals, can use a diverse set of strategies that can then be harnessed to improve estimates relative to individual biased judgments (paper 3). In confidence research, researchers place perhaps too much focus on the accuracy of individual confidence judgments and miss the fact that even biased individual confidence judgments can be used to improve the judgment of a group (paper 4).

We also show how advice-taking and wisdom-of-crowds research can stand to benefit from estimation research. Specifically, we used simulations to establish the ecological rationality of different aggregation methods (i.e., averaging versus choosing). In the Bayesian reasoning domain, we modeled the estimation accuracy of both individual and averaging strategies to reveal specific statistical environments that benefit averaging (i.e., "RE+" environments that pair a rare event with a diagnostic cue). Additionally, we found characteristics of strategies – specifically, use of base-rate information – that predict when averaging will benefit accuracy. In the cue-based estimation task presented in paper 4, we used Juslin et al.'s (2007) Naïve Sampling Model to predict when people should use the average of their inner-crowd and when they should

instead *choose* their high-confidence estimates. Results of an agent-based simulation using psychologically grounded assumptions (including limited working memory spans, finite exemplars stored in long-term memory, and noisy retrieval processes) indicated that most people should have a high correlation between confidence and accuracy in the task and thus benefit from confidence-based estimation. Accordingly, in our study, we found that most participants stood to gain much more from choosing their high-confidence estimates than from taking their average.

Conclusion

This dissertation demonstrates two ways in which people depart from the mythical *homo economicus*. By adapting information search to the presence of competition and tapping into the wisdom of crowds within one mind through dialectical bootstrapping, flesh-and-blood organisms can improve their decisions under uncertainty.

References

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology, 32*(2), 91-134. doi:10.1016/0022-2496(88)90042-9

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. New York, NY: Princeton University Press.

Christensen-Szalanski, J., & Bushyhead, J. (1981) Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7(*4), 928-935. doi:10.1037//0096-1523.7.4.928

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review, 112*(4), 951-978. doi:10.1037/0033-295X.112.4.951

Dutta, P. K., & Rustichini, A. (1993). A theory of stopping time games with applications to product innovations and asset sales. *Economic Theory, 3*(4), 743-763. doi:xxx

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17-52). New York, NY: Wiley.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519-527. doi:10.1037/0033-295X.101.3.519

Fiedler, K., & Juslin, P. (Eds.). (2006). *Information sampling and adaptive cognition*. New York, NY: Cambridge University Press.

Fitzgerald, S. (2012, November 4). CNBC's Jim Cramer: Obama by a landslide. Retrieved from http://www.newsmax.com/US/Jim-Cramer-election-prediction/2012/11/04/id/462697/

Francis Galton. (n.d.). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Galton

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: Take the best and its relatives. In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 75-95). New York, NY: Oxford University Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684. doi:10.1037/0033-295X.102.4.684

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart.* New York, NY: Oxford University Press.

Glaser, M., Langer, T., & Weber, M. (2013). True overconfidence in interval estimates: Evidence based on a new measure of miscalibration. *Journal of Behavioral Decision Making*, *26*(5), 405-417. doi:10.1002/bdm.1773

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*(1), 75-90. doi:10.1037//0033-295X.109.1.75

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*(4), 591-635. doi:10.1016/S0364-0213(03)00031-4

Hertwig, R. (in press). Decisions from experience. In G. Keren & G. Wu (Eds.), *Blackwell handbook of decision making.* Oxford, UK: Blackwell.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534-539. doi:10.1111/j.0956-7976.2004.00715.x

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition, 115*(2), 225-237. doi:10.1016/j.cognition.2009.12.009

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*(2), 231-237. doi: doi:10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2014a). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*. doi:10.1016/j.tics.2014.06.009

Herzog, S. M., & Hertwig, R. (2014b). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 218-232. doi:10.1037/a0034054

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*(1), 40-46. doi:10.1016/0030-5073(78)90037-5

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114*(3), 678-703. doi:10.1037/0033-295X.114.3.678

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454. doi:10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-291. doi:10.2307/1914185

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216-247. doi:10.1006/obhd.1999.2847

Koriat, A. (2012). When are two heads better than one and why? *Science*, *336*(6079), 360-362. doi:10.1126/science.1216549

Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition, 124*(3), 334-342. doi:10.1016/j.cognition.2012.06.002

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*(6), 1231-1243. doi:10.1037//0022-3514.47.6.1231

McKenzie, C. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology, 26*(3), 209-239. doi:10.1006/cogp.1994.1007

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature, 377*(6551), 725-728. doi: 10.1038/377725a0

Morris, D. (2012, November 6). Prediction: Romney 325, Obama 213. Retrieved from http://thehill.com/opinion/columnists/dick-morris/266027-prediction-romney-325-obama-213-

Pachur, T., Hertwig, R., & Wolkewitz, R. (2014). The affect gap in risky choice: Affect-rich outcomes attenuate attention to probability information. *Decision*, *1*(1), 64-78. doi:10.1037/dec0000006

Phillips, N. D., Hertwig, R., Kareev, Y., & Avrahami, J. (2014). Rivals in the dark: How competition influences search in decisions under uncertainty. *Cognition*, *133*(1), 104-119. doi: 10.1016/j.cognition.2014.06.006

Pirolli, P. L. (2007). Information foraging theory: Adaptive interaction with information. New York, NY: Oxford University Press.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*(2), 207-236. doi:10.1037/0096-3445.135.2.207

Rofecoxib. (n.d.). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Rofecoxib

Rotjan, R. D., Chabot, J. R., & Lewis, S. M. (2010). Social context of shell acquisition in *Coenobita clypeatus* hermit crabs. *Behavioral Ecology*, *21*(3), 639-646. doi:10.1093/beheco/arg027

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science*, *12*(3), 185-190. doi:10.1111/1467-9280.00334

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99-118. doi:10.2307/1884852

Sniezek, J., & Van Swol, L. (2001). Trust, confidence, and expertise in a judge–advisor system. *Organizational Behavior and Human Decision Processes, 84*(2), 288-307. doi:10.1006/obhd.2000.2926

Soll, J., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 299-314. doi:10.1037/0278-7393.30.2.299

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 780-805. doi:10.1037/a0015145

Stephens, D. W., Brown, J. S., & Ydenberg, R. C. (Eds.). (2007). *Foraging: behavior and ecology.* Chicago, IL: University of Chicago Press.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* Garden City, NY: Doubleday.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*(7), 645-647. doi:10.1111/j.1467-9280.2008.02136.x

Winkler, R. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association, 66*(336), 675-685. doi:10.2307/2284212

Wulff, D., & Hertwig, R. (2014). *The description–experience gap in the sampling paradigm: A meta-analytic review.* Manuscript in preparation.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93*, 1-13. doi:10.1016/j.obhdp.2003.08.002

Yaniv, I., & Foster, D. (1997). Precision and accuracy of judgmental aggregation. *Journal of Behavioral Decision Making, 10(*1), 21-32. doi:10.1002/(SICI)1099-0771(199703)10:1<21::AID-BDM243>3.0.CO;2-G

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall.

Cognition 133 (2014) 104–119

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Rivals in the dark: How competition influences search in decisions under uncertainty

Nathaniel D. Phillips [a,*], Ralph Hertwig [a], Yaakov Kareev [b], Judith Avrahami [b]

[a] Max Planck Institute for Human Development, Berlin, Germany
[b] The Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem, Israel

ARTICLE INFO

Article history:
Received 24 September 2013
Revised 8 June 2014
Accepted 10 June 2014

Keywords:
Decisions under uncertainty
Competition
Information search
Decisions from experience

ABSTRACT

In choices between uncertain options, information search can increase the chances of distinguishing good from bad options. However, many choices are made in the presence of other choosers who may seize the better option while one is still engaged in search. How long do (and should) people search before choosing between uncertain options in the presence of such competition? To address this question, we introduce a new experimental paradigm called the competitive sampling game. We use both simulation and empirical data to compare search and choice between competitive and solitary environments. Simulation results show that minimal search is adaptive when one expects competitors to choose quickly or is uncertain about how long competitors will search. Descriptively, we observe that competition drastically reduces information search prior to choice.

1. Introduction

Whether the question is what to eat, where to live, or with whom to mate, decisions are often made under competitive conditions. This holds for species ranging from humans to hermit crabs. Arguably choosier than humans are about their housing, hermit crabs are always on the look-out for new and better shells. Because the abdomen of a hermit crab is extremely vulnerable, hermit crabs need find suitable seashells to protect their vital organs in order to pass their genes on to the next generation. When a solitary crab encounters an empty shell, it thoroughly inspects the potential new home. The crab will meticulously explore the outer surface of the shell looking for holes and weak points. It will then insert its vulnerable abdomen into the shell opening to see whether the potential new home is a good fit. If the shell passes this thorough inspection, the crab may decide to discard its current shell and exchange it for the new one. However, when a *group* of crabs simultaneously encounters an empty shell, each individual crabs' search process is dramatically truncated. In this competitive situation, the crab nearest to the shell will make a split-second decision on whether or not to take it based on a brief visual inspection alone (Rotjan, Chabot, & Lewis, 2010).

Swap the hermit crab for a human and the shell for a television on a clearance rack, and intuition suggests that human behavior may be similar to that of hermit crabs'. On a slow shopping day, the leisurely shopper can take his time deciding whether or not to buy the television. He can thoroughly examine the television's attributes, look up expert reviews on his smartphone, or take advantage of the wisdom of crowds by soliciting advice from friends on a social networking site. However, on a frantic shopping day like Black Friday, the same shopper is likely to behave very differently. Surrounded by dozens of other eager

* Corresponding author. Address: Max Planck Institute for Human Development, Center for Adaptive Rationality (ARC), Lentzeallee 94, 14195 Berlin, Germany. Tel.: +49 30 824 06 475.
E-mail address: phillips@mpib-berlin.mpg.de (N.D. Phillips).

http://dx.doi.org/10.1016/j.cognition.2014.06.006
0010-0277/

shoppers, he might spend only a few moments looking at the television before deciding to grab it before someone else does. Why might competition reduce pre-decisional search so dramatically? What costs and benefits do organisms reap by reducing their search efforts in the presence of competition? What factors in choice options and the social environment affect good search rules? In this paper, we seek to provide initial answers to these questions using a new experimental paradigm that we call the *competitive sampling game*.

Organisms rarely have complete and certain information about options before making even the most consequential choices; instead, they must make choices in the darkness of uncertainty. To shed light on the available options, they must learn about those options' possible outcomes and their associated probabilities through an exploratory search process (Real, 1991). Most people go on dates before proposing marriage, vacationers research and compare hotels before deciding where to stay, and hermit crabs inspect new shells before making a move. After a period of exploration, organisms *exploit* an option by making a long-term consequential choice. Exploration and exploitation represent two diametric goals associated with choice, namely, gathering information about options (exploration) versus consuming an option (exploitation) based on current information (Cohen, McClure, & Yu, 2007). Although exploration provides organisms with more information, it can come at costs in the form of money, time, or lost opportunities. There is thus a tradeoff between exploration and exploitation: If you search too little, you might struggle to distinguish good from bad options. If you search too much, you may suffer from excessive search costs.

In solitary choice situations, the exploration–exploitation tradeoff has been extensively studied both theoretically (Brezzi & Lai, 2002; Gittins, 1979; Gittins, 1989) and empirically (Gans, Knox, & Croson, 2007; Groß et al., 2008), mostly in "multi-armed bandit" problems in which individuals attempt to maximize their payoffs from multiple gambles with initially unknown reward distributions. However, previous research on the exploration–exploitation tradeoff has largely ignored a real-world search cost that dramatically changes how organisms behave: the impact of competition during search. Although search affords more information about available options, it also increases the risk that good option(s) will be taken by competitors.

In this article, we research how competition affects pre-decisional exploration from a descriptive as well as a normative perspective. The essence of what we study concerns supply and demand. In a solitary environment, the "supply," that is, the number of options available to choose from, is stable. It cannot be affected by the actions of others. Hence, a solitary decision maker can engage in extensive exploration, allowing her to carefully separate good from bad options at leisure before making a consequential choice. In contrast, in a competitive environment, "demand" increases and the danger lurks that competitors will claim desirable options, leaving the thoroughly exploring decision maker with an inferior option set to choose from. With the increased tension between exploration and exploitation driven by competition, decision makers might be best advised to choose as soon as they detect an option that is likely to be good enough. But when does that moment come? Does search under competition indeed become as truncated as the crab's shell search and the shopper's television search suggest and, if so, how good or bad are the resulting choices? To address these questions, we take advantage of an experimental tool that has recently been used to study the process of search in a range of solitary choice situations (Erev & Barron, 2005; Hertwig, Barron, Weber, & Erev, 2004; Weber, Shafir, & Blais, 2004): the sampling paradigm from research on decisions from experience (Hertwig & Erev, 2009). In this paradigm, participants explore options with a priori unknown underlying probability distributions before deciding between them (exploration before exploitation). In the present research, we pit a solitary variant of this paradigm against a novel competitive variant that we call the *competitive sampling game*.

1.1. Decisions from experience

In the sampling paradigm, a solitary player learns about (i.e. explores) options with a priori unknown payoff distributions that differ in value by sampling outcomes for as long as she wishes, without financial cost. When ready, she chooses (i.e. exploits) her preferred option on the basis of her sampling experience. This final choice then results in a real financial consequence, such as a random payment drawn from the option's payoff distribution. Since the information decision-makers gain through sampling reduces uncertainty about options and increases the likelihood of choosing good over bad options, a key measure in the sampling paradigm is how long people search for information before making a choice. Given that sampling has no cost other than time, one might expect solitary choosers to sample extensively, but previous research shows that protracted search is not the norm. Across studies, participants have generally been found to take between 11 and 19 draws, or about 7 ± 2 samples per option before making a final choice between two gambles (for a review, see Hertwig, in press). Researchers have proposed several reasons why people do not search extensively in solitary choice: small sample statistics can be quite accurate where differences are large enough to matter (Johnson, Budescu, & Wallsten, 2001), frugal search reduces choice difficulty (Hertwig & Pleskac, 2010), short-term maximization goals prompt limited search (Wulff, Hills, & Hertwig, 2014), short-term memory constrains information use, and opportunity costs mount as search continues (Hertwig, in press).

1.2. The Competitive Sampling Game (CSG)

In this paper we introduce a competitive variant of the sampling paradigm called the competitive sampling game. In the game, players choose between two options realized as urns on the computer screen. Each urn contains 100 virtual balls, with each ball bearing a number. The distribution of numbers in an urn dictates its value. Before making a final consequential choice, players have the

opportunity to learn about the distribution of numbers in each urn by drawing random balls (with replacement), one at a time, from either urn as often as and in any order they wish at no financial cost. When a player decides to stop sampling and chooses an urn, she receives the expected value of the distribution of numbers in her chosen urn. In the *solitary* condition, players play alone, as in the sampling paradigm (see Hertwig et al., 2004). In the *competition* condition, they play in pairs. Each player samples independently but at the same rate, meaning that all players see the same number of samples. As long as both players wish to continue sampling, they both do so. As soon as one or more players decide to stop searching and choose an option, all sampling stops and the choosing phase begins. Players receive the option of their choice following the rule of first come, first served. If only one player, the "chooser," decides to stop sampling and make a final choice, that player obtains her chosen option. This forces the other player, the "receiver," to accept the remaining option. If both players simultaneously stop sampling then one of two outcomes can occur: If players want different options, each player gets the option of his or her choice. If both players want the same option, the options are randomly assigned to each player.

The competitive sampling game is akin to "games of timing" (Dutta & Rustichini, 1993), in which two players independently decide when to stop a game and seize a reward while the reward either increases (preemption games) or decreases (war-of-attrition games) over time. An example of a preemption game is the "grab-the-dollar" game, in which two players have the option of either grabbing the money on a table or waiting for an additional period, during which the pot increases by one unit (Park & Smith, 2008). The players' dilemma is that they want both to wait for a larger pot and to be the one claiming the money. The competitive sampling game has the nature of a preemptive race; here, the value of the options becomes clearer over time, but the first person to terminate sampling can decide which option to exploit. Nonetheless, it differs fundamentally from previous games of timing in that players face uncertainty not only about the other's behavior but also about what is at stake—that is, the distribution of each option's outcomes. In other words, the competitive sampling game is a competitive social game (Hertwig, Hoffrage, & the ABC Research Group, 2013), representing situations in which organisms need to trade off exploration of the quality of options for earlier exploitation in order to reduce the risk of the best option being snatched away by a competitor. In the following sections, we address the normative question of how much search is optimal in different variants of the competitive sampling game, and then describe the results of an experimental study.

2. How should accuracy and opportunity be traded off: a simulation study

How should decision makers adjust exploration efforts between solitary and competitive environments? To answer this question, we began by making the following assumptions about the choice ecology, sampling rules, decision rules, and social environment, respectively. Let us emphasize here that our conclusions regarding good sampling sizes in the game will be contingent on these assumptions. In the discussion, we turn to alternative, more elaborate assumptions and more complex environments.

2.1. Choice ecology

Each game presents players with two options with two-outcome payoff distributions. Each distribution has a positive and a negative outcome that occur with complementary probabilities. Positive (O^+) and negative (O^-) outcomes are drawn from uniform distributions ranging from 0 to +100 and −100 to 0, respectively. The probability of the positive outcome $p(O^+)$ is drawn from a uniform distribution with support [0, 1], while the probability of the negative outcome $p(O^-)$ is set to $1 - p(O^+)$. We define the option in the pair with the higher expected value as the *H* option, and we define the performance of a strategy as its likelihood of obtaining the *H* option.[1] For our analyses, we generated 10,000 pairs of payoff distributions and averaged expected strategy performance across all pairs.

2.2. Sampling rules

All players use a "fixed-*N*" sampling rule, where *N* represents the player's planned sampling size. A player with a fixed planned sampling size *N* will elect to continue sampling until the $N + 1$ sampling round, at which point he will stop search and choose. Players distribute their samples equally,[2] between the two options except where *N* is odd, in which case the player will allocate one additional sample to a randomly chosen option. Strategies with small *N* values dictate little exploration prior to exploitation, while those with large *N* values mandate extensive exploration. We calculated expected performance for fixed-*N* strategies with planned sampling sizes ranging from 1 to 50.

2.3. Decision rules

Given a pair of players, the player with the smaller planned sampling size is the *chooser* in the game and gets to decide which option to take. The player with the larger planned sampling size is the *receiver* and automatically receives the remaining option that was not chosen by their competitor. Choosers choose the option that has the highest observed sample mean (i.e., highest mean reward). This rule has been proposed in the context of *n*-armed bandit

[1] We discuss other reasonable performance measures in Section 5. For now, we note that for the two-gamble case, the probability of obtaining option *H* is similar, if not identical, to other performance measures such as the probability of outperforming one's competitor. Additionally, assuming that distributions are not heavily skewed, the probability of obtaining option *H* in most cases should be very similar to the average expected reward.

[2] This assumption is made for simplicity. Although, on average, people draw roughly equal samples from both options in solitary decisions from experience (see Hertwig et al., 2004 Fig. 1), there is also evidence that sampling effort is impacted by factors such as the variability of outcomes encountered during search (Lejarraga, Hertwig, & Gonzalez, 2012).

problems as a computationally simple method for estimating the values of actions such as the play of one of a slot machine's levers, and for using the estimates to select an action (Sutton & Barto, 1998). Unlike in *n*-armed bandit problems, in the sampling paradigm as studied here, the outcomes in the sampling stage inform the players about the value of an option but do not yet represent actual rewards (that is, sampling is exogenous; Denrell, 2007).

In the case that players sample only once, and thus observe an outcome from one option only, they use the following decision rule: If the sample is positive, *choose that option*, if the sample is negative, *choose the other option*. This rule is in the spirit of the win-stay, lose-shift strategy that has been shown to be effective in repeated games environments (e.g., Nowak & Sigmund, 1993). We refer to our variant of this strategy as the *take-good-enough, otherwise-shift strategy*. It specifies that a player will take an observed option if the sample was satisfying (any positive value in our simulation), otherwise he will reject the observed option and take the alternative, unobserved option.[3]

2.4. Social environments

The cost of sampling in competitive environments is likely to depend on the probability that a desirable option will be scooped up by a competitor. For this reason, we expect that the performance of sampling strategies will depend on the specific social environment an organism is in. To measure how competitors' decision speed affects the performance of different levels of search, we simulated choice performance in four different social environments. Mathematically, we defined social environments in terms of the probability distribution of opponents' sampling sizes. We generated four social environments: a *slow* environment in which competitors tend to have large sampling sizes and thus require extensive information before making a decision; a *fast* environment in which competitors tend to have small sampling sizes and are primarily motivated to not let good options slip away; an *uncertain* environment where competitors vary equally between small, medium, and large sampling sizes; and an *as-if solitary* environment that consisted of searchers taken from the original Hertwig et al. (2004) study on solitary decisions from experience. This environment is called as-if solitary because competitors behave as if they are in a solitary environment. To the extent that competition will likely reduce search, the as-if solitary environment represents one possible distributional upper bound on how long individuals will search under competition.

2.5. Simulation results

We show mean performance results across 10,000 randomly generated pairs of payoff distributions drawn from the aforementioned choice ecology (see Appendix A for details). Again, the benchmark used to assess the performance of a sampling rule is the probability that an agent obtains the option with the higher expected value in a gamble pair (the *H* option). We present the simulation results in three sections. First, we contrast expected outcomes for agents who are choosers versus agents who are receivers in a game as a function of the number of sampling rounds in that game. Second, we demonstrate an imbalance in the costs of oversampling versus undersampling. Finally, taking into account this imbalance in costs, we derive the best search length for each social environment.

To what extent does being the chooser (i.e., the one whose planned sampling size is smallest) increase the likelihood of obtaining the *H* option? Is it always good to be the chooser in a game or is it sometimes better to be the receiver and allow a competitor to choose? To answer this question, we calculated the probability that an agent obtains the *H* option given that he ends the game as the chooser across sampling rounds 1–50.[4] In other words, assuming the game lasts for x sampling rounds, what is the probability that an agent obtains the *H* option if he or she is the chooser in the game? Fig. 1 shows the expected outcomes for choosers compared to receivers across rounds 1–50. Recall that, as our implementation of the competitive sampling game requires that the receiver take the option not chosen, the probability that the receiver obtains the *H* option is simply the complement of the probability that the chooser chose it.

We draw two main conclusions from the data presented in Fig. 1. First, across all sampling rounds, choosers are always expected to obtain the *H* option with probability greater than .50. Because receivers receive the *H* option with the compliment of the chooser's probability, receivers always obtain the *H* option with a probability less than .50. No matter how few samples one takes, the expected outcome of being a chooser is always better than the expected outcome of being a receiver. Second, the probability that a chooser obtains the *H* option increases monotonically with additional sampling rounds, but with marginally decreasing gains. In other words, the gain in information a chooser gets from an additional sample in early sampling rounds is larger than the gain in later sampling rounds. From these two findings, it follows that it is always better to have more sampling rounds *as long as* one ends the game as the chooser. In Fig. 1, this means that an agent should try to get as far to the right on the choosing line as possible without being "scooped" and dropping down to the increasingly negative receiving line.

These findings also allow us to construct an optimal sampling rule for an omnipotent player who knows how long her competitor plans to sample. If the omnipotent player knows that his competitor has a fixed and known planned sampling size of n_c, then his best sampling size is $n_c - 1$ (or 1 when $n_c = 1$). Of course, most people are not omnipotent and do not have perfect knowledge of their

[3] Importantly, this rule assumes no uncertainty aversion (Ellsberg, 1961), in that players do not hesitate to take a completely unobserved option when an observed option is found to contain negative outcomes. In the empirical study, we test this assumption (and find evidence against it). However, for the purposes of simplicity we maintain the assumption in determining good sampling sizes in the simulation.

[4] Hertwig and Pleskac (2010) conducted a very similar simulation that paralleled ours (in the case where a player is always the chooser). Our results are virtually identical.

opponent's planned sampling size. Instead, we suspect that people will base their sampling decisions on their expectations of their competitors' behavior, where expectations are defined as a probability distribution over sampling sizes. In other words, players could ask themselves: "How likely is my competitor going to stop search on each sampling round? Once a player has these expectations, she then needs to take into account the costs of over- versus undersampling. If the costs of undersampling (deciding too quickly) are larger than the costs of oversampling (deciding too slowly), then a player should err on the side of sampling too much relative to her expectations of her opponent. On the other hand, if the costs of undersampling are smaller than the costs of oversampling, then the player should shorten her search relative to her expectations of her opponent.

To determine the relative costs of over- versus undersampling in the competitive sampling game, we calculated the expected choice performance of an agent given all combinations of planned sampling sizes from 1 to 10 that an agent and his or her competitor might implement. These data are presented in Fig. 2, where the *x*-axis represents an agent's planned sampling size and the *y*-axis represents a competitor's planned sampling size.

Consistent with our previous conclusion, Fig. 2 shows that for any competitor's planned sampling size n_c, an agent's best planned sampling size, n_a, equals $n_c - 1$ (or 1 when $n_c = 1$). For example, if a competitor's planned sampling size is 8, the best planned sampling size for the agent is 7, with an 83% chance of obtaining option *H*. However, consider the cost of over- versus undersampling against this competitor. If the agent undersamples by 2, with a planned sampling size of 5, he will still be the chooser in the game and have an 80% chance of obtaining option *H*—a drop of only 3 percentage points relative to the best possible outcome. If, on the other hand, the agent oversamples by 2, with a planned sampling size of 9, he will be the receiver in the game and will have only a 16% chance of obtaining option *H*—a plunge of 67 percentage points.

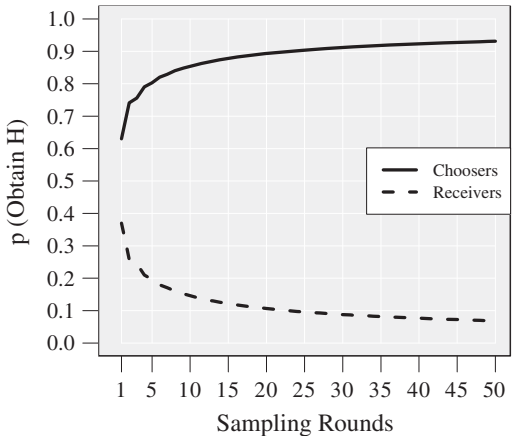


Fig. 1. Mean probability that an agent obtains the *H* option given that the game stops at a specified sampling round across 10,000 gambles. The choosing line shows the probability for the chooser, the player with the larger planned sampling size, while the receiving line shows the probability for the receiver, the player with the smaller planned sampling size.

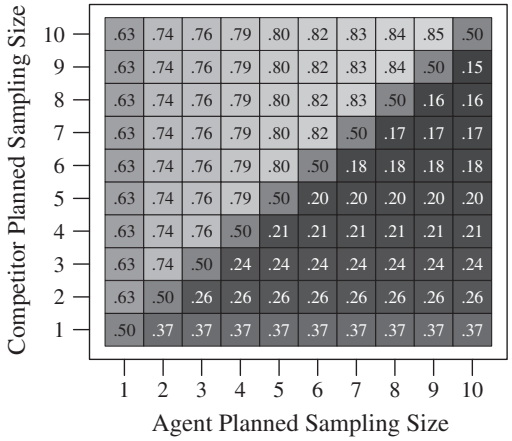


Fig. 2. Probability that an agent obtains the *H* option given the agent's planned sampling size and the competitor's planned sampling size. The figure shows mean values over 10,000 randomly generated two-outcome gambles.

These results show that the cost of oversampling is much larger than the cost of undersampling: it is always better to undersample (by any amount) and keep the chooser advantage, than to oversample (by even one sample) and suffer the receiver disadvantage.

Given that it is better to err on the side of undersampling versus oversampling, how should an individual behave in different social environments? In other words, how little should one sample before making a decision given certain expectations about the behavior of competitors? To answer this, we calculated a player's expected probability of obtaining the *H* option given his or her planned sampling size within each of the four social environments (see Appendix B for details). In the slow environment, competitors had relatively large planned sampling sizes (mean of 30), following a bounded, discretized normal distribution with a standard deviation of 5. In the fast environment, competitors had relatively small planned sampling sizes (mean of 3.33) following a right-skewed distribution. In the uncertain environment, competitors had—with equal probability—any planned sampling size from 1 to 50 (mean of 25.5). Finally, in the as-if solitary environment, competitors had a right-skewed distribution of sampling sizes (mean of 18). Fig. 3 (left panel) shows the probability mass functions for each of these social environments.[5]

The right panel of Fig. 3 shows the expected probability than an agent obtains the *H* option given the planned sampling size within a specific social environment. In a slow environment, agents with a planned sampling size of 18 did best, with an 88.2% chance of obtaining *H*. In contrast, in an uncertain environment, the best planned sampling size was 6, with a 75.8% chance of obtaining *H*. In a fast

[5] For the purposes of the prescriptive analyses, these distributions can represent either the variability in the behavior of one's opponent from one game to another, or the variability in the behavior of an entire population of individual competitors. Assuming that an opponent's sampling rule in each game is an independent, random sample from its parent distribution, the mathematics are the same whether we attribute variability to inter- or intra-individual causes.

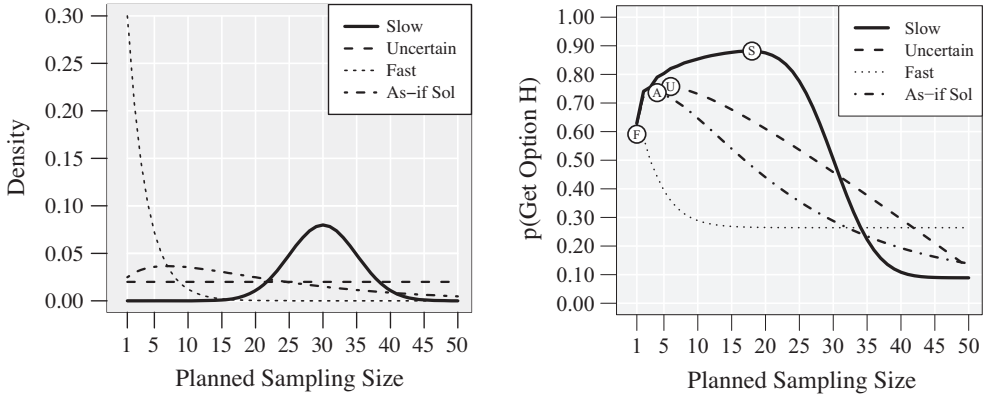


Fig. 3. Left panel: Probability of an agent encountering a competitor with a given planned sampling size in the slow, fast, uncertain, and as-if solitary environments (see text). Right panel: Results from the agent-based simulation averaged across 10,000 randomly generated decision problems. The *x*-axis indicates the planned sampling size of an agent, and the *y*-axis shows the expected probability of obtaining the higher expected value option (*H*) as a function of the planned sampling size in the four social environments. While lines are continuous, the underlying data is discrete.

environment, agents with a planned sampling size of just 1 did best, obtaining the *H* option 59% of the time. In the as-if solitary environment, a planned sampling size of 4 proved best, with a 73.8% chance of obtaining *H*. Finally the only sampling size that ensures that one will obtain option *H* with probability no less than .50, regardless of the behavior of one's competitor, is a sample size of just 1 using the take-good-enough, otherwise-shift strategy.

These results show that players should dramatically reduce exploration in a competitive context when they lack clear information of their competitors' intentions. Consider for illustration the uncertain social environment, in which a competitor is equally likely to stop anywhere between 1 and 50 samples (see Fig. 3, left panel). If an omnipotent player knew exactly how long her competitor planned to sample, then the best strategy would simply be to take one fewer samples than her opponent. However, in this uncertain social environment, the exact planned sampling size of the opponent is unknown. What happens if a player plans to sample one less than the *expected* planned sampling size of her competitor? In this uncertain social environment where the expected planned sampling size of a competitor is 25.5, this rule would dictate a planned sampling size of 25. Reference to Fig. 3 (right panel) shows that this planned sampling size constitutes dramatic oversampling, as the best planned sampling size for this social environment is only 6—less than one-fourth of the competitor's expected sampling size. In this example, planning to sample one round less than the expected sampling size of one's uncertain competitor leads to oversampling by over 19 rounds. The reason behind this dramatic effect of uncertainty about the other agent's actions on the best planned sampling size is the disproportionate costs of over- versus undersampling.

To conclude, adaptive sampling in the competitive sampling game depends on expectations of one's competitors. If competitors value accuracy highly and consequently represent a slow social environment, decision makers can afford to gather more information. Yet there is considerable asymmetry in the costs of over- versus underestimating competitors' need for accuracy. In our choice ecology, underestimating one's competitor's sampling size, no matter by what degree, will always ensure that one will be the chooser and thus more likely than not to obtain option *H*. On the other hand, overestimating one's competitor's sampling size, no matter by what degree, will always ensure that one will be the receiver and thus more likely than not to obtain the short end of the stick (i.e., the lower expected value option, *L*). For these reasons, we find that it is better to err on the side of underestimating the competitor's sampling size and minimizing the risk of being scooped.

These simulations show that competition presents a substantial additional cost of search. Consequently, we expect that real people will search much less in a competitive compared to a solitary context. But how much more restricted will it be? Will real people competing with others decrease their search in a magnitude prescribed by our simulations? Or will people be reluctant to decrease predecisional search so dramatically? To answer this question, we conducted an empirical study on the competitive sampling game and compare the search behavior of people participating in solitary to competitive games.

3. An empirical investigation of the competitive sampling game

3.1. Method

A total of 180 students from the University of Basel participated in the study.[6] They received a flat fee of CHF 7.50 (approximately $8.12 at the time) for their participation, as well as a bonus contingent on their winnings in the game. The mean bonus across both experimental conditions was CHF 1.18 (approximately $1.26) with a standard deviation of CHF 1.19. Participants completed the study in groups of four, each on a separate computer. They received no information about the choice ecology prior to beginning the task. All players began by playing three practice games without financial consequences to familiarize themselves with the experimental interface (see Appendix C for practice game parameters). They were then presented with five decision tasks. Each decision task contained two, two-outcome

[6] Gender data were not recorded due to a programming error.

Table 1
Choice ecology.

	Task 1		Task 2		Task 3		Task 4		Task 5	
Gamble set	H	L	H	L	H	L	H	L	H	L
1	(37, 0.44, −17): 6.76	(25, 0.22, −11): −3.08	(29, 0.36, −13): 2.12	(43, 0.35, −20): 2.05	(33, 0.46, −15): 7.08	(49, 0.28, −23): −2.84	(55, 0.41, −26): 7.21	(55, 0.28, −26): −3.32	(37, 0.44, −17): 6.76	(37, 0.26, −17): −2.96
2	(43, 0.43, −20): 7.09	(29, 0.24, −13): −2.92	(49, 0.35, −23): 2.20	(33, 0.35, −15): 1.80	(25, 0.5, −11): 7.00	(37, 0.26, −17): −2.96	(55, 0.41, −26): 7.21	(55, 0.28, −26): −3.32	(37, 0.44, −17): 6.76	(37, 0.26, −17): −2.96
3	(55, 0.41, −26): 7.21	(55, 0.28, −26): −3.32	(37, 0.44, −17): 6.76	(37, 0.26, −17): −2.96	(29, 0.48, −13): 7.16	(43, 0.27, −20): −2.99	(49, 0.42, −23): 7.24	(33, 0.25, −15): −3.00	(25, 0.36, −11): 1.96	(37, 0.35, −17): 1.90
4	(37, 0.44, −17): 6.76	(25, 0.22, −11): −3.08	(29, 0.36, −13): 2.12	(43, 0.35, −20): 2.05	(37, 0.44, −17): 6.76	(55, 0.28, −26): −3.32	(49, 0.42, −23): 7.24	(49, 0.28, −23): −2.84	(33, 0.46, −15): 7.08	(33, 0.25, −15): −3.00
5	(33, 0.46, −15): 7.08	(33, 0.25, −15): −3.00	(49, 0.42, −23): 7.24	(49, 0.28, −23): −2.84	(55, 0.35, −26): 2.35	(37, 0.35, −17): 1.90	(43, 0.43, −20): 7.09	(29, 0.24, −13): −2.92	(25, 0.5, −11): 7.00	(37, 0.26, −17): −2.96
6	(55, 0.41, −26): 7.21	(37, 0.26, −17): −2.96	(25, 0.36, −11): 1.96	(37, 0.35, −17): 1.90	(29, 0.48, −13): 7.16	(43, 0.27, −20): −2.99	(49, 0.42, −23): 7.24	(49, 0.28, −23): −2.84	(33, 0.46, −15): 7.08	(33, 0.25, −15): −3.00
7	(43, 0.43, −20): 7.09	(43, 0.27, −20): −2.99	(33, 0.46, −15): 7.08	(49, 0.28, −23): −2.84	(29, 0.48, −13): 7.16	(29, 0.24, −13): −2.92	(55, 0.35, −26): 2.35	(37, 0.35, −17): 1.90	(37, 0.44, −17): 6.76	(25, 0.22, −11): −3.08
8	(55, 0.41, −26): 7.21	(37, 0.26, −17): −2.96	(49, 0.35, −23): 2.20	(33, 0.35, −15): 1.80	(25, 0.5, −11): 7.00	(37, 0.26, −17): −2.96	(43, 0.43, −20): 7.09	(43, 0.27, −20): −2.99	(29, 0.48, −13): 7.16	(29, 0.24, −13): −2.92
9	(25, 0.36, −11): 1.96	(37, 0.35, −17): 1.90	(49, 0.42, −23): 7.24	(33, 0.25, −15): −3.00	(29, 0.48, −13): 7.16	(29, 0.24, −13): −2.92	(43, 0.43, −20): 7.09	(43, 0.27, −20): −2.99	(37, 0.44, −17): 6.76	(55, 0.28, −26): −3.32
10	(55, 0.41, −26): 7.21	(37, 0.26, −17): −2.96	(29, 0.36, −13): 2.12	(43, 0.35, −20): 2.05	(33, 0.46, −15): 7.08	(49, 0.28, −23): −2.84	(37, 0.44, −17): 6.76	(37, 0.26, −17): −2.96	(25, 0.5, −11): 7.00	(25, 0.22, −11): −3.08
11	(49, 0.35, −23): 2.20	(33, 0.35, −15): 1.80	(43, 0.43, −20): 7.09	(29, 0.24, −13): −2.92	(37, 0.44, −17): 6.76	(37, 0.26, −17): −2.96	(25, 0.5, −11): 7.00	(25, 0.22, −11): −3.08	(37, 0.44, −17): 6.76	(55, 0.28, −26): −3.32
12	(49, 0.42, −23): 7.24	(33, 0.25, −15): −3.00	(55, 0.35, −26): 2.35	(37, 0.35, −17): 1.90	(29, 0.48, −13): 7.16	(43, 0.27, −20): −2.99	(37, 0.44, −17): 6.76	(37, 0.26, −17): −2.96	(25, 0.5, −11): 7.00	(25, 0.22, −11): −3.08

Note: Gamble sets used in both the solitary and the competitive conditions. Rows correspond to the 12 different combinations of decision tasks. *H* represents the higher expected value option, and *L* represents the lower expected value option within each decision task. Each option is a discrete, two−outcome random variable with one positive and one negative outcome that occur with complementary probabilities. Values in parentheses are the value and the probability of the positive outcome, and the value of the negative outcome, for each gamble. The value outside the parentheses is the expected value of the gamble.

gambles, each with one positive and one negative outcome occurring with complementary probabilities. The gamble sets were constructed such that in certain pairs the options differed in expected value and in others they did not; likewise, in certain pairs the options differed in range and in others they did not (see Appendix C for a full description of how gamble parameters were selected). Three different orders of each of the 12 gamble sets were created, resulting in 36 unique experimental sessions (see Table 1). Location of the urns on the screen was randomly determined for each decision task and on each run.

At the outset of each decision task, participants saw two options represented visually as opaque urns. They were told that each urn contained 100 virtual balls, each of which was worth a (not necessarily unique) number of points. Participants were informed that they would be rewarded with one-tenth of the average value of all the balls in the urn they chose (or were allocated). Each of the participants ($n = 36$) assigned to the solitary condition completed one of the 36 unique experimental sessions. These participants could sample from the urns as many times as they wished before making a final choice. Having made a final choice of an urn in a decision task, they moved onto the next task. The other 144 participants played each decision task in the *competition condition*. At the beginning of each task, they were paired randomly with one of the other three participants. This pairing was done independently between tasks. Players were not told which person (of the three) they were playing against in each decision task.

Every decision task, in both the solitary and competition conditions, began with one mandatory sampling round. On every subsequent sampling round, each player indicated whether he or she wanted to sample from an urn or to make a final choice. These decisions were made privately and were only revealed to both players after both had made a sampling or choice decision. If both wanted to take a sample, they were asked to click on an urn and viewed a randomly sampled outcome from that urn. Players could see which urn the other player sampled from, but could not see the outcome the other player observed. If, after observing a sample, both players wanted another sample then another sampling round began. If one player decided to make a final choice (becoming the "chooser"[7]), she then selected the urn she wanted and her choice was recorded. Subsequent to the chooser's choice, the other player (the "receiver") was informed that her competitor had made a choice and that he must take the remaining non-chosen urn. If both players made a choice on the same sampling round, one of two outcomes was possible: If the two players chose different urns, they each received the urn of their choice. If both players chose the same urn, the two urns were randomly assigned to the players. After final choices were made and players learned which urn they received, they were randomly paired again and the next decision task began. The random pairing was done independently of prior rounds, so a player could play the same opponent on sequential games. Participants did not receive immediate task-by-task feedback on how much money they won from their chosen urns. At the end of the session, participants were informed how much money they had earned across the five decision tasks and were paid accordingly.

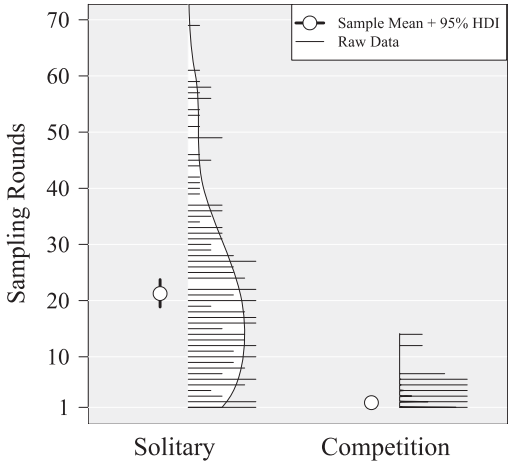


Fig. 4. Distribution of sampling rounds across all decision tasks and individuals, separately for the solitary and the competition conditions. One sampling round value of 100 in the solitary condition is not displayed. The 95% HDI interval for the solitary condition is plotted 'behind' the solitary sample mean.

4. Results

We used Bayesian graphical modeling for all inferential statistics. Bayesian posterior densities were calculated using the R2Jags package in R. Posterior densities were calculated with uninformative uniform priors, 10,000 iterations, a burn-in value of 1000, and no thinning. We conducted Bayesian hypothesis tests using Bayes Factors calculated using the Savage–Dickey method for nested model comparisons. We used the conventions developed by Jeffreys (1961) to determine the categorical degrees of strength indicated by Bayes Factors (BF). All raw data and complete code are available in our online supplementary materials. In the following sections, we first report on search and then on choice, comparing both behaviors in the competition condition relative to the solitary condition.

4.1. How drastically do people restrict explorative behavior under competition?

We measure exploration efforts by the number of sampling rounds tasks lasted prior to a choice. For solitary games, this is simply the number of samples the player took. For competitive games, this is the number of sampling rounds that occurred prior to the first choice. Fig. 4 presents the distribution of sampling rounds across all decision tasks in the solitary and competition conditions. In the solitary condition, the median number of sampling rounds was 18 (mean of 21.05; 95% highest density interval [HDI]: 18.78, 23.67).[8] In the competition condition, in

[7] Players were not explicitly given the "chooser" and "receiver" labels in the experiment.

[8] The mean for the distribution of sampling sizes in the solitary condition was calculated from the 95% HDIs for the p and r parameters in the negative binomial distribution.

contrast, the median number of sampling rounds was 1 (mean of 1.82; 95% HDI: 1.61, 2.08).[9] The difference in sample means was 19.27 (95% HDI: 17.00, 21.92) and provides extremely strong evidence against the hypothesis that the means of the two distributions were the same (BF > 100). Thus, the amount of sampling by participants who were competing for resources was dramatically lower than that of participants who were searching alone.[10]

4.2. How much does very restricted search compromise decision quality?

To see how the restricted search in the competition condition affected decision quality, we calculated how often choosers in the competition condition chose the *H* option relative to receivers and to participants in the solitary condition. These results are presented in Fig. 5. In the solitary condition, players chose the *H* option in 71% (95% HDI: 64%, 78%) of decision tasks. This result constitutes extremely strong evidence for the hypothesis that participants in the solitary condition were more likely than chance to choose the *H* option (BF > 100). Next we analyze the outcomes for choosers and receivers in the competition condition. In cases where players chose the same option simultaneously, one player was randomly assigned to be the chooser (and obtained the option both chose) and the remaining player was assigned to be the receiver (and obtained the alternative option). Choosers obtained the *H* option in 58% of decision tasks (95% HDI: 53%, 63%)—fewer than in the solitary condition (BF = 12.36, strong evidence), but more than would be expected by chance alone (BF = 8.90, moderate evidence).

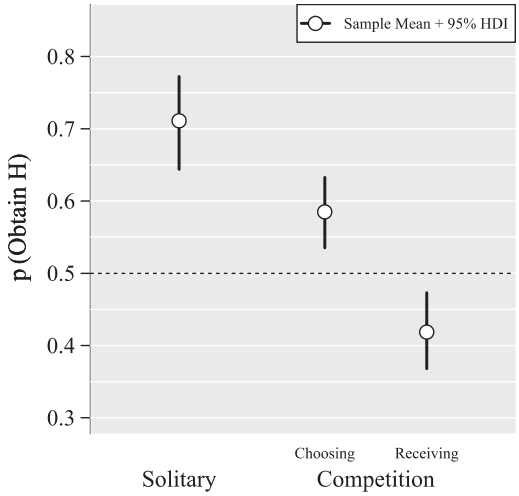


Fig. 5. Proportions of tasks where players obtained the higher expected value (*H*) option in the solitary and the competition conditions. Error bars represent 95% highest density intervals for the population probability.

Thus, the reduced information available to fast choosers in the competitive condition indeed reduced their choice performance relative to solitary choosers. Nevertheless, as choosers took the better option at above chance level, receivers obtained option *H* in only 42% (95% HDI: .37, .47) of cases[11] This is consistent with our simulation analysis (Fig. 3, right panel) showing that fast choosing is advantageous under competitive conditions.

4.3. How did players make choices based on minimal information?

The following analyses focus on the competition condition only. To further analyze the specific decisions that produced the distribution of sampling rounds in the competition condition, we looked at how quickly players terminated sampling. For each of the 720 decision tasks (144 participants × 5 decision tasks each), we recorded the total number of sampling rounds that occurred in the task, and whether players were choosers or receivers (i.e., two choosers or one chooser and one receiver). We found that in 32% (227 of 720) of all cases, participants were choosers who decided to choose immediately after the first sampling round. Of these choices, 88% (200 of 227) were consistent with the take good enough, otherwise-shift heuristic. The remaining 12% (27 out of 227) either chose an option with an observed negative value, or did not choose an option with an observed positive value. In 22% (159 of 720) of all cases, participants were receivers after the first sampling round, because they opted to continue sampling while their competitors decided to choose. Participants made it to the second round in only 46% (334 of 720) of cases. Of these 334 participants, 34% (113 of 334) were

[9] The mean for the distribution of sampling sizes in the competition condition was calculated from the 95% HDIs for the *p* parameter in the geometric distribution.

[10] The fact that sampling rounds decreased in competition relative to solitary conditions, is necessary, but not sufficient evidence that competition reduced individual sampling decisions. The reason for this lies in how sampling rounds are defined. Under competition, sampling rounds are defined at the level of pairs of participants rather than individual participants. Because sampling rounds are restricted by the behavior of the fastest player in a pair, we would expect a decrease in sampling rounds in the competitive task relative to the solitary task even if players did not change their sampling rules. For example, if two players employ fixed-*N* rules of 5 and 10, respectively, across solitary and competitive games, the average number of sampling rounds in the solitary games would be 7.5, while the average number of sampling rounds under competition would (always) be 5. To test whether or not this shrinkage effect could explain the different distributions of sampling rounds between solitary and competition conditions, we generated all possible pairs of sampling rounds from the solitary game and calculated the minimum sampling round number from each pair. This represented the expected distribution of sampling rounds in the competition condition if behavior was the same as in the solitary condition. The median number of sampling rounds in this distribution was 11 (mean of 12.61, 95% HDI: 12.45, 12.76). The difference in the mean sampling rounds between this distribution and the observed distribution for the competition condition was 10.76 (95% HDI: 10.49, 11.06), thus offering extremely strong evidence against the hypothesis that the means of the two distributions are the same (BF > 100). We conclude that the difference between the mean number of sampling rounds in the competition condition and the solitary condition was due not only to the rules of the competitive game (i.e., the fastest player determines sample size) but also to the fact that competition per se shifted the balance from exploration to exploitation.

[11] Receivers did not receive option *H* at a percentage exactly equal to one minus the percentage that choosers wanted option *H* (which was 57%). This is due to the effects of simultaneous choosing by choosers. In games where both players simultaneously chose option *H*, both wanted option *H*, but only one player got it.

Table 2
Distribution of participant-level proportion of decision tasks ending with a choice.

0/5	1/5	2/5	3/5	4/5	5/5
2 (1.4%)	16 (11.1%)	34 (23.6%)	45 (31.3%)	30 (20.8%)	17 (11.8%)

choosers who decided to stop sampling, 25% (82 of 334) were receivers, and 41% (139 of 334) "survived" to the third sampling round. This analysis demonstrates that, while not all choosing decisions were made after one round (as prescribed by the analysis of the fast environment; Fig. 3), people indeed drastically confined their information search, and to a greater extent than prescribed by the uncertain and solitary search environments.

What makes people decide to choose after just one sample versus to continue sampling? Our data suggest the valence of the first sample influences this decision. Of the 261 cases in which a player experienced a positive outcome in the first sample, he or she stopped sampling and chose immediately in 130 cases (50%; 95% HDI: 44%, 56%). Of the 459 cases in which a player experienced a negative outcome in the first sample, he or she stopped sampling in only 97 cases (21%; 95% HDI: 18%, 25%). Thus, players were more willing to immediately choose an option with a known positive outcome than they were to reject an option with a known negative outcome and take a completely uncertain option (BF > 100, extreme evidence)—perhaps a manifestation of aversion to ambiguity (Ellsberg, 1961).

4.4. How closely did players pursue a single strategy?

Next, we examined how consistent individual participants were in their behavior across tasks. In the competitive sampling game, some players could always try be the chooser, while others might require so much pre-decisional information that they always defer the choice to their competitor. To see if people had stable choosing versus receiving outcomes, we calculated the percentage of all 5 games that each participant was a chooser. Table 2 reports the distribution of these percentages across individuals. We found little evidence that people's search strategies resulted in stable outcomes across decision tasks. Only 19 individuals (13.2%) were either always choosers (17) or always receivers (2), while the remaining 86.8% ended some tasks as the chooser and other tasks as the receiver.

Next, we determined whether and how individuals changed their behavior across the tasks. For example, if a player ended up as a receiver in one task, did she decrease her sampling in order to increase the chance of being a chooser in the subsequent task? In contrast, did the chooser take the liberty of sampling a little more in the next task? To answer this question, we calculated the marginal probability that a player was the chooser each task in addition to the probability of being the chooser conditional on his or her status in the previous task (i.e., chooser versus receiver). If behavior in a game changes as a function of the outcome in the previous game, we would expect differences in the conditional probability of choosing when a player was a chooser in the previous game compared to being a receiver in the previous game. Table 3 reports the results.

We begin by looking at the marginal probability of choosing across decision tasks. If players adjust their explorative efforts downward with each round, the probability that they end games by choosing (rather than receiving) should increase and converge toward 1. We did not find substantial evidence for this, as the probability of choosing oscillated between around 57% and 61% across tasks. Next, we looked at the probability of choosing, conditioned on the outcome of the previous task. We found no evidence that being a receiver in one round prompted less search and a higher probability of being the chooser in the next task. In other words, we do not find clear evidence that players changed their behavior across tasks based on experience. One possible explanation for this finding is that, although players were told whether they were the chooser or the receiver in any given round, they did not receive immediate feedback on the direct monetary consequences of their behavior. Without immediate consequential feedback, players may not have had sufficient information to adjust their sampling strategies.

5. General discussion

We designed the competitive sampling game to extend the sampling paradigm used in research on decision from experience (see Hertwig & Erev, 2009) to competitive decision tasks under uncertainty. The task enables researchers to investigate how people adapt their exploration efforts in response to the simultaneous presence of uncertainty about nature (i.e., the parameters of the payoff distribution) and uncertainty about the social environment of competitors (i.e., their desire for accuracy versus their desire to beat their competitors to the punch). In our initial research using the game, we found that competition dramatically affects the exploration–exploitation tradeoff. In a simulation analysis, we found that sampling sizes as low as 1 or 2 can be best in certain environments when people compete with others for advantageous options. Empirically, our participants showed dramatically reduced search in competitive task compared to solitary one.

5.1. Varying the number of players and options

In our experiment and simulations, we contrasted a solitary task with one player and two options, with a competitive task with two players and two options. One could wonder whether it is merely the presence of competition, or the ratio between competitors and the number of available options that drives the need for speedy decisions.[12] A moment's reflection makes it clear that good sampling rules

[12] We thank Jonathan Nelson for pointing this out.

Table 3
Choosing behavior across sequential decision tasks.

Rounds/ decision task	*p*(choosing)	*p*(choosing|previous choosing)	*p*(choosing|previous receiving)
1	82 (56.9%)		
2	83 (57.6%)	48 (58.5%)	35 (56.4%)
3	87 (60.4%)	51 (61.4%)	36 (59.0%)
4	83 (57.6%)	54 (62.1%)	29 (50.9%)
5	89 (61.8%)	55 (66.3%)	34 (55.7%)

are likely to depend both on the number of options available, and the number of competitors: If there are many options available relative to players, then one can take more time for sampling, knowing that even if all other players make quick decisions, there will likely be many good options remaining. To find out the relationship between the number of competitors and number of options on good decision speed, we ran a supplementary set of simulations in which we systematically manipulated three factors: number of players (from 2 to 5), number of options (from 2 to 5) and the speed of competition (fast, uncertain, and slow). For each factor combination, we simulated the performance of an agent using a fixed sampling size of 1 through 15. In contrast to our previous simulations, we now define performance as the average expected reward the agent obtained across all simulations.[13] Details of the simulation are in Appendix E.

We present the main results of the simulation in Fig. 6. Each plot shows an agent's planned sampling size on the horizontal axis, and the agent's expected reward on the vertical axis. Each line corresponds to a different social environment, mirroring three of the four (excluding the as-if solitary condition) from our earlier simulations (see left panel of Fig. 3). Plots in each column refer to different numbers of players, from 2 to 5, while plots in each row refer to different numbers of options, from 2 to 5. The top left graph (2 players and 2 options) replicates the same social and environmental structure as our initial simulation. Within each plot, the sampling size that maximizes an agent's expected rewards for each social environment is highlighted with an enlarged point.

We briefly summarize 4 key results from Fig. 6. First, holding the number of players constant, as the number of options increases the best sampling sizes tend to increase. Additionally, the expected reward given the best sampling size increases as well. This means that the more options there are, the longer one should search, and the better one's expected outcomes will be. Second, holding the number of options constant, as the number of players increases the best sampling sizes tend to decrease. Both of these results support our prior intuitions: the degree to which people should reduce sampling in the presence of competition depends on the number of options available and the number of other players.

[13] We use expected reward instead of the probability of obtaining the highest expected value option for two reasons. First, organisms frequently want to obtain good options, not necessarily the best option. Second, as the number options increases, the probability that any player will discover and take the highest expected value option will necessarily decrease. This makes it more difficult to compare performance between option number conditions.

Next, we look at the effect of the player/option ratio on performance. If the ratio of players to options remains the same, does the absolute number of players and options matter? We find that indeed, there is a substantial effect. Consider games where the player: option is 1:1. Here, we find that as the absolute number of players and options increases, the risk involved with taking large samples increases in a fast environment, but decreases in a slow one. To see this, compare the expected rewards of having a large (15) sampling size in the 2:2 game (top left panel) compared to the 5:5 game (bottom right panel). In the 2:2 game, extensive search against slow competition leads to an expected reward of around 20, while the same level of search against fast opponents leads to an expected loss of around −10. Here, the difference in expected rewards against slow and fast opponents is 30. Now consider the 5:5 game. Here, the expected reward against slow opponents *increases* to around 35, while the expected loss against fast opponents *decreases* to around −15. Now, the difference in expected rewards between competitors is 50, an increase of 66% in the range of potential outcomes compared to the smaller 2:2 game. This means that when the absolute number of players and options increases, while keeping the player to options ratio 1:1, both the potential benefits one can gain using extensive search against slow competition increases while the potential losses on can suffer against fast competition increases. In other words, the more players and options are in the game, the more risk one runs (with 'risk' defined as the difference between the expected reward with the best sample size and with the largest sample size) by extensive search, in fast and uncertain environments. Independently of this effect, the main result from our previous analyses still hold—the faster you expect your opponents to decide, the faster you should decide, regardless of how many options and how many players are in the game.

Next, we consider games where there are more players than options. These games are akin to real-world problems such as house-hunting and mate-search where there may be more 'buyers' than 'sellers.' For example,[14] in Beijing, men outnumber women, and thus (heterosexual) men find themselves in a competitive game with more players than options (Jacobs, 2011). Assuming that options cannot be shared among players, these games necessitate that some players will leave without an option of their own. In our simulation, we assumed that these players receive neither rewards nor losses from leaving empty-handed. However, one can easily imagine real world decisions where this assumption does not hold. If you are competing with others for one of three open positions at a company, it might be much worse to get no job at all than to get a random (or even the worst) job of the three. Similarly, in mate-selection, leaving empty-handed could very well be the worst possible outcome from an evolutionary perspective. To incorporate this cost, one could assign a fixed negative loss for players that leave the game empty-handed, with larger losses representing domains where it is especially bad to leave with no option (e.g.; mate-search). While we do not run these

[14] We thank Jonathon Nelson for providing this example.

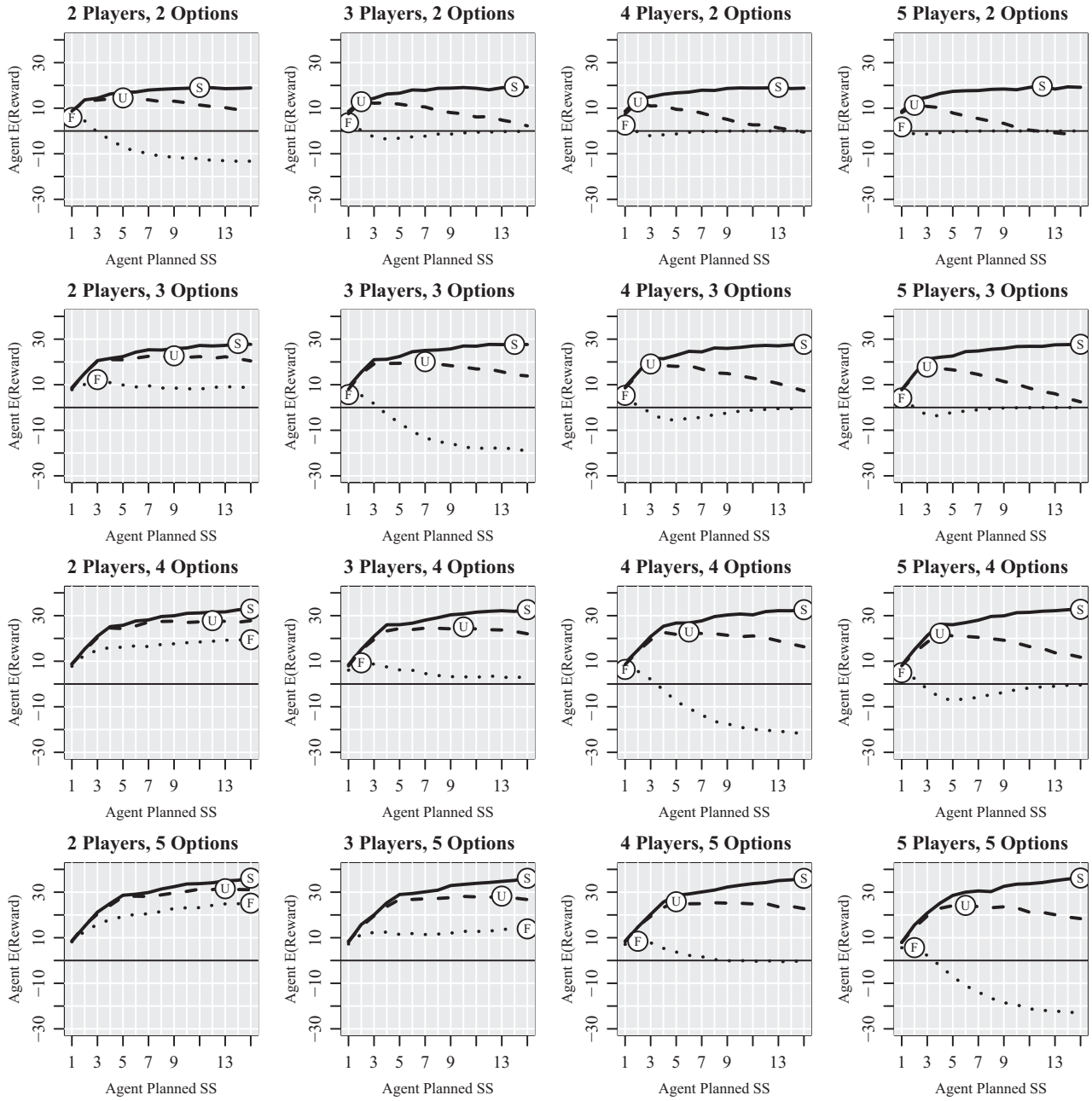


Fig. 6. Simulation results depicting the expected reward (vertical axes) of an agent in the competitive sampling game given a specified sampling size (horizontal axes). Separate plots in each column correspond to different numbers of players in the game, while separate plots in each row correspond to different numbers of options in the game. Solid, dashed, and dotted lines refer to the slow, uncertain, and fast social environments respectively.

simulations, our prediction for the effect is clear: as the cost of leaving empty-handed increases, sampling sizes should decrease.

5.2. Alternative search and decision rules

In our simulations, we assumed that players used a "fixed-*N*" sampling rule. That is, players were assumed to have a fixed sampling size threshold that they had to reach before making a decision. Additionally, we assumed that players distribute their samples equally between options. We chose to limit our analyses to this simple class of search rules as a starting point for exploring the effects of different exploration efforts on performance. Of course, the fixed-*N* sampling rule plus equal allocation constitutes just one of many possible search rules people are likely to use. For example, one promising, more complex, class of search rules are those that compare sample statistics with an information threshold in order to decide whether to stop or to continue sampling. These rules have been found to be promising both normatively and descriptively (e.g., Busemeyer & Rapoport, 1988). While we cannot claim that fixed-N sampling rules represent either the best approach to the competitive sampling game or that most

participants employed them, we do predict that fixed-N rules will mimic the behavior of more complex search rules. For example, a fixed-1 rule should behave very similarly to a rule that says "Always try to choose before your opponent" or "Choose after the difference in sample means is greater than ε (where ε is a small threshold)." Similarly, a Fixed-30 rule will behave similarly to a conservative rule such as: "Choose when the expected probability of choosing the best option is greater than p (where p is a large probability)." Here is our point: to the extent that small fixed-N values mimic search rules that make do with minimal information before making a decision, and large fixed-N rules mimic those that require extensive information, our conclusion that competition should and does reduce the amount of information people require before making a decision should hold.

Finally, we did not deal with agents' expectations concerning their competitor's search behavior Rather, we took the competitor sampling size distributions as given, and determined which sampling size best responds to them. We do suspect that real people in competitive tasks (in our empirical study and in the real world) try to predict the sampling size of their opponents and, through some iterated process of strategic thinking (e.g., Ho, Camerer, & Weigelt, 1998) settle on a sampling size. Future research should model the processes by which a person generates expectations of other player's behavior, translates those expectations into a decision rule, and after gaining experiences, updates their expectations.

5.3. When will very frugal search fail?

Our simulation results are based on aggregation across specific distributions of gambles and are only valid within those distributions. It is clear that other gamble distributions can lead to very different results. One important factor is how much the gamble distributions are favorable for decisions based on small samples; if options are 'unfriendly' to small samples, then our previous conclusions will not hold. The gamble distributions in our stimuli and empirical study, were indeed small-sample-friendly. When averaged over 10,000 simulated pairs of gambles, we found that a sample size of 1 results in an expected rate of obtaining the H option of greater than .50 (see Fig. 1). Thus, we created an environment where a sample of size 1 using the take-good-enough, otherwise-shift rule was, on average, sufficient. However, this success rate does not generalize to any gamble environment. In Appendix D, we show that only gamble pairs where the sum of the probability of obtaining a positive outcome from option H and the probability of obtaining a negative outcome from option L is greater than 1 guarantees an expected probability of choosing the H option that is greater than .50 (see proof in Appendix D). We label gamble pairs that satisfy this condition as "one-sample favorable." In our simulations, the proportion of gamble pairs that were one-sample favorable was .787; in these gambles, the probability of choosing option H using one sample was .688. The remaining portion of gamble pairs that were not one-sample favorable was thus .213; in these gambles, the probability of choosing option H using one sample .414. This result highlights the fact that the accuracy of decisions based on very small samples will depend on the specific distributions encountered by agents.

One of the most important findings from early research on decisions from experience was that, in experience-based decisions, low-probability (rare) events appear to receive less impact than they deserve in light of their objective probability (Hertwig et al., 2004). This effect is, among other factors (Hertwig & Erev, 2009), caused by the fact that people do not search long enough to experience rare events often enough or at all during search. Formally, in environments where the ranks of *most* samples from each option diverge greatly from the true rank of options' long-term average values, choosing based on a small sampling size can lead to a small probability of obtaining the H option. In other words, in gambles where small samples (e.g., a single date with a potential mate, a glance at a TV on sale) produce data that are inconsistent with an option's long-term value (e.g., a disastrous first date with Mr. or Ms. Right, a paid celebrity endorsement of a low-quality product), frugal predecision sampling can lead people to choose poor options. For example, consider a payoff distribution that delivers +1 with probability .9 and −100 with probability .1 and thus has an expected value of −9.1. Small samples are unlikely to reveal the rare but large negative outcome of −100, making the distribution look advantageous to most agents that inspect it only briefly. This suggests that fast choosers in competitive environments involving rare events run the risk of choosing options that appear beneficial in the short term but have detrimental long-term consequences resulting from rare but impactful negative events (e.g., "black swans"; Taleb, 2007). Indeed, in such environments, a player could even benefit from competing against others who are "tricked" into grabbing options with apparent short-term gains, but actual long-term losses.

6. Conclusion

Our findings suggest that competition shifts the balance between exploration and exploitation in an uncertain choice environment: Faced with the threat of being outpaced in the process of making a decision, people dramatically reduce search. As our results show, this is a smart thing to do in ecologies in which competitors can be expected to choose quickly, and modal samples are good indicators of an option's value. Although exploitation means forgoing the benefits of exploration that can be enjoyed in solitary situations, those who seize the first-mover advantage do better than those who do not in many (but, of course, not all) competitive situations. It is a proverbial truth that you should "look before you leap" (see also Savage, 1954/1972, p. 16). In our competitive environment, it emerged that a quick peek before leaping was very helpful—but that more extensive looking permitted the competitor to leap first and gain an edge.

Acknowledgments

This research was supported by the Swiss National Science Foundation project number 100014_129572 and by

the Israel Science Foundation grant 121/11. We thank the members of the Center for Adaptive Rationality (ARC), Arend Hintze, Julian Marewski, and Jonathan Nelson for many helpful comments. We also are grateful to Valerie M. Chase and Susannah Goss for editing the manuscript. We thank Daniel Lowengrub of the Hebrew University of Jerusalem for programming the competitive sampling game. The first author thanks D.S.P, J.D.T, and V.C.H. for inspiring him throughout the preparation of this manuscript.

Appendix A

A.1. Calculating the expected probability of obtaining option H given a choice ecology and social environment

We calculated a player's expected probability of obtaining the option with the higher expected value (H) given its planned sampling size using Eq. (1):

$$p(H|n_i) = p(H|n_c > n_i) \cdot p(n_c > n_i) + p(H|n_c < n_i) \cdot p(n_c < n_i) + p(H|n_c = n_i) \cdot p(n_c = n_i) \quad (1)$$

Eq. (1) represents the weighted sum of three possible scenarios that differ with respect to the relationship between the player's planned sampling size (n_i) and the planned sampling size of the competitor (n_c), that is, whether the player faces an opponent with a smaller, larger, or the same planned sampling size.

The first half of the first term in Eq. (1) corresponds to the probability of obtaining the H option given that the competitor will sample longer than the player, and, consequently, the probability that the player will obtain H equals that of choosing the H option given n_i samples. Because choices are based on sample means, this equals the probability that the order of the sample means from the two options matches the order of the population means. In this case, the option with the higher sample mean will also be the option with the higher population mean and the player will choose the better option. Formally:

$$P(H|n_c > n_i) = p(\bar{x}_H > \bar{x}_L|n_i) \quad (2)$$

Note that this calculation is specific to the choice ecology under consideration. For two outcome payoff distributions such as those used here, this can be calculated directly by comparing the results of two binomial distributions. The second half of the first term is the weight given to this outcome, defined as the probability of encountering an opponent with a larger sample size than the player's sample size.

The remaining two terms in Eq. (1) follow the same logic. When the competitor samples less than the player, the probability that the player obtains the H option is the probability that the competitor will *not* choose the H option. This equals the probability that the opponent observes sample means whose order is not equal to the true order of population means and can be calculated as follows:

$$p(H|n_c < n_i) = p(\bar{x}_H < \bar{x}_L|n_c) \quad (3)$$

Finally, the third term in Eq. (1) represents the expected outcome when both players have the same sampling size. This is set to .5 and is independent of the sampling distributions of payoff distributions H and L (=lower expected value distribution):

$$p(H|n_c = n_i) = .5 \quad (4)$$

Appendix B

B.1. Distributions of planned sampling sizes for competitive social environments

In the *fast* environment F, the probability that a randomly sampled agent has a planned sampling size n_k is given by a geometric distribution with $p = .3$, ranging from 1 to 50 and normalized to sum to 1:

$$f(F = n_k) = \frac{(1-.3)^{n_j-1} \cdot .3}{\sum_{i=1}^{50}(1-.3)^{i-1} \cdot .3} \quad n_k = 1, 2, \ldots, 50$$

In the *slow* environment S, the probability that a randomly sampled agent has a planned sampling size n_k is a reflected version of F around the point $n_k = 25.5$:

$$f(S = n_k) = \frac{(1-.3)^{50-n_k-1} \cdot .3}{\sum_{i=1}^{50}(1-.3)^{i-1} \cdot 3} \quad n_k = 1, 2, \ldots, 50$$

In the *uncertain* environment U, the probability that a randomly sampled agent has a planned sampling size n_k is given by the discrete, uniform distribution with bounds at 1 and 50:

$$f(U = n_k) = \frac{1}{50} \quad n_k = 1, 2, \ldots, 50$$

In the *as-if solitary* environment A, the probability that a randomly sampled agent has a planned sampling size n_k is given by a negative binomial distribution with $p = .071$ and $r = 1.59$. The distribution is bounded from 1 to 50:

$$f(A = n_k) = \frac{\binom{n_k + 1.59 - 1}{n_j}(1-.071)^{1.59} \cdot .017^{n_k}}{\sum_{i=1}^{50}\binom{1 + 1.59 - 1}{i}(1-.071)^{1.59} \cdot .071^{i}} \quad n_k = 1, 2, \ldots, 50$$

Appendix C

C.1. Properties of the practice games

Practice 1		Practice 2		Practice 3	
H	L	H	L	H	L
(32, .458, −13)	(39, .352, −18)	(47, .417, −25)	(42, .349, −21)	53, .407, −24)	(35, .458, −17)

C.2. How gamble parameters were selected

We started out with four binary-valued options, labeled A1–A4. Their values were (−17, 37), (−20, 43), (−23, 49), and (−26, 55), for options A1 through A4, respectively. Each of these options had three versions, differing in expected value; the expected values were high (EV = 7), medium (EV = 2), or low (EV = −3). The three EVs were obtained by modifying the probabilities of the two outcomes. Additionally, each A option had a corresponding B option, for which the two values spanned a smaller range (hence, smaller variance); the range of the B option was about 2/3 that of the corresponding A option. We then constructed 12 gamble sets, each comprising five decision tasks (displayed in Table 1). One involved a choice between an A option with an EV of 7 and a B option with an EV of −3 (e.g., A1High and B1Low); a second involved a choice between another A option and its corresponding B option, in which the A option had the low EV and the B option the high EV (e.g., A3Low and B3High); a third involved a choice between another A and B pair in which the two were both of the medium value, with EV = 2 (e.g., A2Medium and B2Medium); a fourth involved a choice between two (large variance) A options belonging to the same set but differing in value (e.g., A4High and A4Low); a fifth involved a choice between two (small variance) B options belonging to the same set but differing in value (e.g., B4High and B4Low). There were 12 such sets, and we used them all.

Appendix D

D.1. Conditions that make decision tasks one-sample favorable

Consider an environment containing two gambles (options) H and L, where $E(H) > E(L)$. Assume a player selects a gamble at random and draws a random sample. Let the random variable S represent the selected option where $S \in \{H, L\}$. Let the random variable $X \in R$ be outcome drawn be the outcome drawn from the selected gamble. Finally, let the random variable $C \in \{H, L\}$ be the chosen option.

Consider a player using a one-sample search and decision rule: (1) Select an option at random and draw one sample. (2) If the sample value is positive, choose the selected option. If the sample value is negative, chose the unselected option. From the law of total probability, the probability that player will choose option H can be written as the sum of the probabilities of two disjoint events:

$$p(C = H) = p(S = H)p(X > 0|S = H) + p(S = L)p(X < 0|S = L)$$

Because options are selected at random, $p(S = H) = p(S = L) = .50$:

$$p(C = H) = .5p(X > 0|S = H) + .5p(X < 0|S = L)$$

Moving terms around

$$p(C = H) = .5(p(X > 0|S = H)) + p(X < 0|S = L)$$

It follows that for $p(C = H)$ to be greater than .50, $p(X > 0|S = H) + p(X > 0|S = L)$ must be greater than 1.0.

Appendix E

E.1. Second simulation procedure

We simulated the performance of agents with varying fixed sampling sizes playing the competitive sampling game against varying numbers of competitors, number of options, and competitor speed. The key parameters we varied were: N.Players (2, 3, 4, 5, 6): the number of competitors in the game. N.Options (2, 3, 4, 5, 6): the number of options (gambles) in the game. Competition.Speed (Slow, Uniform, Fast): the decision speed of competitors. This created 72 simulation classes. For each simulation class, we simulated the decision performance of 15 agents playing the competitive sampling game, each using a fixed sample size of 1–15. We aggregated each agent's performance over 5000 stochastic factors: (1) the outcome distributions within each of the (N.Options) options and (2) the specific stopping rules of its (N.Competitors) competitors. Each option represented a discrete, two-outcome gamble with one positive and one negative outcome, each occurring with complementary probabilities. For each of the options, we drew a positive outcome from Unif(0, 100) and a random negative outcome from Unif(−100, 0). We then drew the probability of the positive outcome ($p+$) from Unif(0, 1) and set the probability of the negative outcome ($p-$) to $1 - p+$. We constructed the probability mass function for each option independently of other options. For each of the competitors, we drew a sample size from its corresponding decision speed distribution (Slow, Uniform, or Fast). These distributions corresponded to those in Appendix B

Each game proceeded as follows: agents sampled equally from options until the first player reached its sampling size. That agent then choose the option with the highest observed sample mean (with ties broken at random). In the case where two agents stopped at the same time, one of two outcomes could occur: If they wanted different options, they each got their desired option. If they wanted the same option, then the desired option was randomly given to one agent, and the remaining agent then attempted to take is next most desired option. After all agents who stopped on that round received an option, the game continued with the remaining players and options. At the end of each game, each agent got the expected value of its chosen option. In the games where there were more players than options, if a player ends the game with no option (because all options were taken by other players), then it received a reward of 0.

Appendix F. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2014.06.006.

References

Brezzi, M., & Lai, T. L. (2002). Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control, 2*(1), 87–108.

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology, 32*(2), 91–134.
Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B, 362*(1481), 933–942.
Denrell, J. (2007). Adaptive learning and risk taking. *Psychological Review, 114*(1), 177–187.
Dutta, P. K., & Rustichini, A. (1993). A theory of stopping time games with applications to product innovations and asset sales. *Economic Theory, 3*(4), 743–763.
Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics, 75*, 643–669.
Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112*(4), 912–931.
Gans, N., Knox, G., & Croson, R. (2007). Simple models of discrete choice and their performance in bandit experiments. *Manufacturing and Service Operations Management, 9*(4), 383–408.
Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B, 41*, 148–177.
Gittins, J. C. (1989). *Multi-armed bandit allocation indices*. New York, NY: Wiley.
Groß, R., Houston, A. I., Collins, E. J., McNamara, J. M., Dechaume-Moncharmont, F. X., & Franks, N. R. (2008). Simple learning rules to cope with changing environments. *Journal of the Royal Society Interface, 5*(27), 1193–1202.
Hertwig, R. (in press). Decision from experience. In G. Keren & G. Wu (Eds.), *Blackwell handbook of judgment and decision making*. Oxford, UK: Blackwell.
Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534–539.
Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences, 13*(12), 517–523.
Hertwig, R., Hoffrage, U., & the ABC Research Group (2013). *Simple heuristics in a social world*. New York, NY: Oxford University Press.
Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition, 115*(2), 225–237.
Ho, T. H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best response in experimental "p-beauty contests". *The American Economic Review, 88*(4), 947–969.
Jacobs, Andrew (2011, April 11). *For many Chinese men, no deed means no dates.* <http://www.nytimes.com/2011/04/15/world/asia/15bachelors.html>.
Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.
Johnson, T. R., Budescu, D. V., & Wallsten, T. S. (2001). Averaging probability judgments: Monte Carlo analyses of asymptotic diagnostic value. *Journal of Behavioral Decision Making, 14*(2), 123–140.
Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition, 124*(3), 334–342.
Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature, 364*(6432), 56–58.
Park, A., & Smith, L. (2008). Caller number five and related timing games. *Theoretical Economics, 3*(2), 231–256.
Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science, 253*(5023), 980–986.
Rotjan, R. D., Chabot, J. R., & Lewis, S. M. (2010). Social context of shell acquisition in *Coenobita clypeatus* hermit crabs. *Behavioral Ecology, 21*(3), 639–646.
Savage, L. J. (1954). *The foundations of statistics* (2nd rev. ed.). New York, NY: Dover.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York, NY: Random House.
Weber, E. U., Shafir, S., & Blais, A. R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review, 111*(2), 430–445.
Wulff, D. U., Hills, T. T., & Hertwig, R. (2014). The impact of long- and short-run frames on search and choice in decisions from experience. Submitted for publication.

The Janus Face of Darwinian Competition

Arend Hintze [*], **Nathaniel Phillips** [†], **Chris Adami?** [‡], **Ralph Hertwig** [†]

[*]MSU, [†]MPIB, and [‡]

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Without competition organisms would not evolve any meaningful physical or cognitive abilities, and as such competition can be understood as the driving force behind Darwinian evolution. But can this trend be extrapolated, and more competitive environments necessarily evolve organisms with more sophisticated cognitive abilities than those in less competitive environments? Answering this question will tell us if there is a breaking point at which competition does more harm than good, or if ultimately the most competitive environments will produce the most sophisticated organisms. We evolve decision strategies of virtual agents that have to perform a repetitive sampling task in three different environments. These environments differ in the degree in which the actions of a competitor can affect the fitness of the sampling agent, and in the variance of the sample. We find that under weak competition agents evolve decision strategies that sample often and make accurate decisions, which not only improve their own fitness, but are also good for the entire population. However, under extreme competition the Janus face reveals it's dark side and forces agents to sacrifice accuracy for speed, and also prevents agents to sample more often than higher variance in the environment would require it. Modest competition is therefore a good driver for evolving cognitive abilities and the population as a whole, whereas too much competition is devastating.

term | term | term

Introduction

Competion is the basic principle of Darwinian evolution. Over time it brings out better adapted organisms, and weeds out weaker physical and cognitive designs. But does this mean that more competition will always result in more adaptive and sophisticated cognition? There are many biological examples of competition driving the evolution of cognitive abilities [West-Eberhard(1979), Whiten and Byrne(1997), Flinn et al.(2004)Flinn, Geary, and Ward, Arbilly et al.(2014)Arbilly, Weissman, Feldman, and Grodzinski], and we successfully use competition in genetic algorithms [Stanley et al.(2005)Stanley, Bryant, and Miikkulainen, Edlund et al.(2011)Edlund, Chaumont, Hintze, Koch, Tononi, and Adami, Marstaller et al.(2013)Marstaller, Hintze, and Adami], yet there is also evidence that find that the level of accuracy achieved in human decision making can be lower under competition [Phillips et al.(2014)Phillips, Hertwig, Kareev, and Avrahami] than without. In decision making, more competition between agents typically forces them to answer faster. However faster responses exact a cost: they rely on less information and in the most extreme on little to no information. Less information often - yet not always [Gigerenzer et al.(2011)Gigerenzer, Hertwig, and Pachur]- means a lower level of inferential accuracy. How does evolution tradeoff accuracy and speed? We will show that in less competitive environments accuracy wins over speed, whereas more competition necessitates quicker and less accurate decisions. Ultimately, in extremely competitive situations agents may rely on minimal information to prevent competitors from claiming desirable options, possibly leaving the thoroughly exploring agent with an inferior option set to choose from. The downside of this strategy however is risking to end up with an inferior option due to minimal sampling.

In nature we find a good example of this exploration/exploitation tradeoff [Gittins et al.(2011)Gittins, Glazebrook, and Weber]. Hermit crabs outgrow the shells they live in, and from time to time, have to find better ones. In order to improve their housing situation, they need to find a new shell,and investigate (sample) their potential new home [Rotjan et al.(2010)Rotjan, Chabot, and Lewis, Phillips et al.(2014)Phillips, Hertwig, Kareev, and Avrahami]. Once they find that the new shell is better than their own (outside reference) they move into the new shell. Obviously, hermit crabs have been evolved to sample sufficiently well in order to improve their housing situation while they grow. If a lonely hermit crab finds a new shell it will sample sufficienty often to make an accurate assessment. If the hermit crab, on the other hand, finds the shell in the presence of a conspecific the situation is radically different. Being slower in arriving at a decision than a competitor can mean that the competitor claims the better shell. However, a hermit crab changing into another shell does not destroy its old one, and therefor a population of directly competing hermit crabs sorts out the inferior shells for the benefit of everyone - a process similar to *crowd sourcing*. This situation was conducive for the evolution of a cognitive strategy that samples sufficiently often. Contrary to that, imagine a situation of a different organism where leaving a shell destroys it, or more generally a resource perishes. Here immediate decisions under direct competition are potentially more adaptive than thorough sampling, and thus could prevent organisms from evolving the cognitive ability to explore (sample) thoroughly. It seems as if the type of competition but also the nature of the resource competed about [Tilman(1982)] matters.

To show how varying degrees of competition can lead to the evolution of the ability to sample thoroughly as well as prevent its evolution, we computationally evolve decision making agents in environments featuring indirect or direct competition. In the following we will describe these different environments: The game we use is a competitive variant of a sampling paradigm [Hertwig et al.(2004)Hertwig, Barron, Weber, and Erev, Weber et al.(2004)Weber, Shafir, and Blais, Phillips et al.(2014)Phillips, Hertwig, Kareev, and Avrahami], where players have to draw random numbers from an urn, and decide whether the urn they are sampling from has a higher mean value than an outside reference. This is a typical example for experienced-based decision making [Hertwig and Erev(2009)]. The outside reference has a payoff known to the agent, whereas for the urn, only sampling can reveal its value. The outside reference could be understood as another urn that has already been exhaustively sampled from. Dependent of the agents decision, it will either receive the value

Reserved for Publication Footnotes

of the outside reference or the mean value of the urn sampled from. This game is designed in such a way that sampling from the unknown urn more often, while integrating the sampled data, will increase the ability of the agent to accurately assess its value (see law of large numbers [Bernoulli(1713)]). The accuracy of an assessment given a certain number of samples depends on the underlying distribution. The performance of agents competing in an evolving population will determine their fitness, and ultimately those players who sample more often will make more accurate decisions than those who sample little. This only holds as long as decisions of competing agents can not remove the more desirable urn. This is the case when competition is indirect and happens only on the population level (See Figure:1 A). We call this the *indirect competition environment*.

In order to allow for agents to directly compete in our model, two agents sample from the same urn and try to establish whether the urn has a higher or lower mean than each agent's respective reference urn. Once an agent decides to either stay with its reference urn or to pick the urn sampled from, the game ends, and the other agent receives its reference urn's value as payoff. If both agents decide simultaneously to pick the same urn, it will be given randomly to either agent, while the other agent will receive the value of the reference urn as payoff (See Figure:1 B). This is the case when competition is direct. That is, one agent's decision directly affects the other agent's choice environment, and, by extension, its fitness. We will call this the *direct competitive environment*.

In the third environment we amplify the effect of direct competition even more. Both agents are in possession of an urn, and they know the payoff of their respective urns (outside reference). They can sample from the other agent's urn and decide to claim it, leaving the competitor with the abandoned urn. Agents can stop the game by deciding to keep their urn or by choosing the other agent's urn. Now agents do not only compete over an outside resource, but can actively decrease their opponents fitness (See Figure:1 C). We call this the*extreme competitive environment*. We now test how these different environments affect the evolution of sampling strategies and their effect on agents' payoff.

Methods

Agents sample from urns that upon sampling return a value drawn from normal distribution, with a mean of 1, 2, 3, or 4, and with a variance of 0.1, 1.0, 3.0, or 5.0 respectively. We distinguish three evolutionary environments, involving indirect competition, direct competition, and extreme competition. The first and simplest condition forces a player to choose between two urns. The first urn's mean payoff is known to the player and thus more sampling is not necessary [Busemeyer(1985)]. This urn is called the reference. The second urn's mean is unknown and the agent can sample in order to decide whether to keep the reference, or to claim the other urn. Once an agent decides upon an urn it receives it's value as a payoff. Agents choose between three alternative actions: *stay*, *continue*, or *select*. An agent who chooses to *stay* thus decides to claim the urn with the known payoff. An agent who *continues* will draw one more sample from the unknown urn. The agent that *selects* claims the sampled urn. Each agent can have a range of possible mappings between sampled experiences and actions (stay, continue, or pick). The agents experiences are represented by two different parameters. The first parameter (m) is the difference between the average of all samples taken so far and the reference urn. The second parameter is the number of samples already taken (n). Here, m describes how different both urns are, but this estimate depends on the number of samples taken. A large difference after one sample might be misleading, whereas a large difference after many samples is more likely to be a valid indicator. The parameter n allows the agents to take the number of samples into account. To avoid infinitely many samples, we set the maximum number of samples to 100. We start the game always with the continue action so that an agent samples at least once. We encode the decision strategy of an agent as two probabilities that determine the actions it chooses. The first probability defines whether or not an agent stays. If an agent decides against staying, the second probability defines the likelihood of an agent to continue or to select. In order to make the probabilities dependent on the difference (m) and the number of samples (n), we use an exponential function that incorporates these two parameters according to the following equation:

$$\begin{aligned} p(m,n) = g_1 m^3 + g_2 n^3 + g_3 m^2 + g_4 n^2 + g_5 m \\ + g_6 n + g_7 m^2 n + g_8 m^2 n + g_9 mn + g_{10}. \end{aligned} \quad [1]$$

We use this equations to encode the strategy of each agent, once to determine the probability of a player to stays or not, and another time to decide whether an agent continues or picks. This equation is not limited to be between 0.0 and 1.0. Therefore, a negative value or $p(m,n)$ is defined to be a probability of 0.0, whereas a value of $p(m,n) > 1.0$ is defined to be a probability of 1.0. To allow for agents to evolve a wide variety of different probabilities to stay, continue and pick, two parameter vectors g are required. The first is used to determine the probability of the initial choice to either stay or to choose one of the two other actions. The second vector consequently determines the probability to continue or to pick. The values of the parameter vectors g can be understood as the genome of the agent and determines its decision strategy. We use a well mixed population of 1024 agents, and at each update play each agent against four randomly selected neighbors (when necessary). After each update 1% of the agents are replaced proportional to the payoff they accumulated over the last updates (Moran death birth process using roulette wheel selection [Moran(1962)]). Each of the components of the vectors g have a 1% chance to mutate, once an agent was selected to make offspring. A mutation adds a uniform random number from the interval $[-0.5, 0.5]$ to the mutating component, with no upper or lower limit thereafter. The simulation is run for 500,000 updates. A random organism from the population at the end of the simulation is selected, and the line of decent for this organism is reconstructed. The population usually converges fast to a most recent common ancestor (15,000 updates, data not shown); therefore we choose the agent at update 450,000 as the representative result of that simulation run. Running the simulation longer does not change the results, because agents reached the fitness optimum, or can not find ways to further improve their strategies.

Each of the three competitive environments is used in 100 replicate experimental runs. In the indirect competitive environment the agent's performance solely depends on its strategy; in the other two environments the performance of an agent also depends on its opponent. Consequently we can not measure the performance of an agent by itself. Pitting an agent against a very un-evolved or poorly choosing opponent would overestimate its performance. Pitting it against an extremely well choosing opponent would underestimate it. Therefore, we measure an agents performance by pitting it against itself. Thus, the representative agent at the end of the simulation competed against itself $1,000,000$ times. Three outcome criteria are measured: the number of samples taken

before selecting an urn, how well sampling was tuned to the variance in the environment, and how often each of the urns was taken by that agent. However, we distinguish between the individual's and the population payoff. The results for 100 representative agents at the end of the simulation are averaged.

Results

How does the degree of competition affect what decision strategy evolves, measured in terms of the outcome criteria? Before we turn to the number of samples taken, we first show the evolved probabilities to stay, continue, or pick in Figure 2. All agents evolved a high probability to stay with their reference urn when the difference between sampled urn and reference urn is negative. This negative difference indicates a situation in which the reference urn has a higher payoff and thus should be chosen. The maximum of this probability changes only dependent on the intensity of competition. The more competitive the environment the higher the likelihood to choose the reference urn. Consequently, agents evolved a high probability to select the sampled urn when the difference between sampled urn and reference is positive. And again, the more competitive the environment, the earlier we find the maximum probability of picking the sampled urn. Figure 3 shows the number of samples taken before an urn was selected. This number generally decreases with competitiveness of the environment, and only increases with increasing variance among the samples taken from an urn. This is expected, since a wider distribution requires more samples to assess the true mean. Interestingly, in the extremely competitive environment the number of samples for either variance is merely 1. In this environment agents base their decision only on the minimal sample of 1. Here strategies evolved to choose extremely fast rather than to gauge the urn's mean value by drawing more samples.

The different types of environments not only have an effect on the strategy evolved, but also on the payoff. If every agent in the indirect competitive environment (see Figure:1 A) plays optimally the population is expected to have an optimal gain. An agent in the indirect competition environment can at best always choose the better of the two urns, thus resulting in a maximum average payoff of $3.\bar{3}$. Randomly choosing would result in a payoff of 2.5 (See Supplementary Information). Evolved strategies in this environment come close to this optimal payoff with 3.12 on average (See Figure:4 left). Here the individual's payoff is identical to the average payoff in the population. Furthermore, competition causes agents to evolve strategies that sample dependent on the variance of the environment (See Figure 3).

In the direct competitive environment (See Figure:1 B), two players are confronted with three urns in total. Ideally the best agent will be able to always choose the best out of three urns, with a payoff of 3.75. The competing agent, if not getting the highest possible payoff, but the remaining second highest, receives 2.5 (See Supplementary Information for the outcome of choices 5).

A perfect strategy playing itself will win in 50% of the cases, and consequently loose in 50%. Therefor the average maximum payoff is 3.125 ($\frac{3.75}{2} + \frac{2.5}{2}$). The average payoff is 3.01 and is thus close to the expected maximum payoff (See Figure:4 middle). Again the individual's and average population payoff is higher than randomly choosing. However, competition causes agents to only evolve a moderate ability to sample more often in higher variance environments (See Figure 3).

In the extreme competitive environment(See Figure:1 C) both agents compete over two urns at the same time, and the loosing agents always receives the urn the player who picks first left behind. Here the best possible strategy optimally receives $3.\bar{3}$, whereas the worst possible strategy would only get $1.\bar{6}$, while choosing randomly gives 2.5. Strategies evolved in the extreme competitive environment on average receive 2.49, that is the payoff of a randomly choosing agent (See Figure:4 right). Here competition is so extreme that the risk that a competitor beats one to the punch is larger than a benefit gained from one more sample. This prevents the evolution of repeated sampling as well as the ability ot adjust sampling as a function of environmental variance (See Figure 3)

Discussion

The starting point of our investigation was the question whether more competition will always result in better adapted cognition, or if overly competitive environments may even hamper the evolution of adaptive decision strategies? Organisms evolved in environments of varying competitive pressures, and these conditions shaped their behaviors in the present. We designed three different environments in which agents had to evolve a decision strategy under indirect, direct, and extreme competition. Indirect competition drove the evolution of repeated sampling and sampling that was rsponsive to the variance in the environment(See Figure 3). Direct competition let to a similar result except that sampling was only moderately responsive to the variance in the environment. Extreme competition, in contrast, forced agents to make an immediate decision (based on one sample), and evolved no sensitivity to the variance in the environment. To avoid misunderstanding, agents evolved optimal decision strategy for the environment they faced. Yet, agents in the extremely competitive environment evolved the least sophisticated decision strategy, measured in terms of number of samples and sampling being responsive to environmental variance. A related consequence is the agents inability to evolve decision strategies that not only improve the agent's fitness but also the fitness of every member in the population. In less competitive environments, however, agents evolved a strategy that not only maximized their own payoff, but also allowed the average payoff in the population to increase. Our results show that competition can optimize decision strategies not only for the benefit of the individual, but also for the benefit of the others. But there comes a point at which competition becomes too much of a good thing. Under extreme competition, the agents evolved behavior that exclusively bets on speed over accuracy. Excessive competition reveals the dark side of the Janus face, by inhibiting the evolution of decision strategies, that can tradeoff speed for accuracy.

ACKNOWLEDGMENTS. – text of acknowledgments here, including grant info –

Arbilly et al.(2014)Arbilly, Weissman, Feldman, and Grodzinski. Michal Arbilly, Daniel B Weissman, Marcus W Feldman, and Uri Grodzinski. An arms race between producers and scroungers can drive the evolution of social cognition. Behavioral Ecology, 25(3): 487–495, May 2014.

Bernoulli(1713). J Bernoulli. Jacobi Bernoulli,... Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola Gallice scripta De ludo pilae reticularis. 1713.

Busemeyer(1985). J R JR Busemeyer. Decision making under uncertainty: a comparison of simple scalability, fixed-sample, and sequential-sampling models. Journal of experimental psychology Learning, memory, and cognition, 11(3):538–564, June 1985.

Edlund et al.(2011)Edlund, Chaumont, Hintze, Koch, Tononi, and Adami. Jeffrey A JA Edlund, Nicolas N Chaumont, Arend A Hintze, Christof C Koch, Giulio G Tononi, and Christoph C Adami. Integrated information increases with fitness in the evolution of animats. PLoS computational biology, 7(10):e1002236–e1002236, October 2011.

Flinn et al.(2004)Flinn, Geary, and Ward. Mark V Flinn, David C Geary, and Carol V Ward. Ecological dominance, social competition, and coalitionary arms races:Why humans evolved extraordinary intelligence. Evolution and Human Behavior, 26(1): 10–46, December 2004.

Gigerenzer et al.(2011)Gigerenzer, Hertwig, and Pachur. Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur. Heuristics. The Foundations of Adaptive Behavior. Oxford University Press, May 2011.

Gittins et al.(2011)Gittins, Glazebrook, and Weber. John Gittins, Kevin Glazebrook, and Richard Weber. Multi-armed Bandit Allocation Indices. John Wiley & Sons, February 2011.

Hertwig and Erev(2009). Ralph Hertwig and Ido Erev. The description–experience gap in risky choice. Trends in Cognitive Sciences, 13(12):517–523, November 2009.

Hertwig et al.(2004)Hertwig, Barron, Weber, and Erev. Ralph R Hertwig, Greg G Barron, Elke U EU Weber, and Ido I Erev. Decisions from experience and the effect of rare events in risky choice. Psychological science : a journal of the American Psychological Society / APS, 15(8):534–539, August 2004.

Marstaller et al.(2013)Marstaller, Hintze, and Adami. Lars Marstaller, Arend Hintze, and Christoph Adami. The evolution of representation in simple cognitive networks. Neural Computation, 25(8):2079–2107, August 2013.

Moran(1962). Patrick Alfred Pierce Moran. The statistical processes of evolutionary theory, 1962.

Phillips et al.(2014)Phillips, Hertwig, Kareev, and Avrahami. N D Phillips, Ralph Hertwig, Y Kareev, and J Avrahami. Rivals in the dark: How competition affects information search and choices. Cognition, August 2014.

Rotjan et al.(2010)Rotjan, Chabot, and Lewis. Randi D Rotjan, Jeffrey R Chabot, and Sara M Lewis. Social context of shell acquisition in Coenobita clypeatus hermit crabs. Behavioral Ecology, 21(3):639–646, May 2010.

Stanley et al.(2005)Stanley, Bryant, and Miikkulainen. K O Stanley, B D Bryant, and R Miikkulainen. Evolving neural network agents in the NERO video game. In Proceedings of the IEEE, 2005.

Tilman(1982). David Tilman. Resource Competition and Community Structure. Princeton University Press, 1982.

Weber et al.(2004)Weber, Shafir, and Blais. Elke U EU Weber, Sharoni S Shafir, and Ann-Renee AR Blais. Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. Psychological Review, 111(2):430–445, April 2004.

West-Eberhard(1979). M J West-Eberhard. JSTOR: Proceedings of the American Philosophical Society, Vol. 123, No. 4 (Aug. 30, 1979), pp. 222-234. In Proceedings of the American Philosophical Society, 1979.

Whiten and Byrne(1997). Andrew Whiten and Richard W Byrne. Machiavellian Intelligence II. Extensions and Evaluations. Cambridge University Press, September 1997.

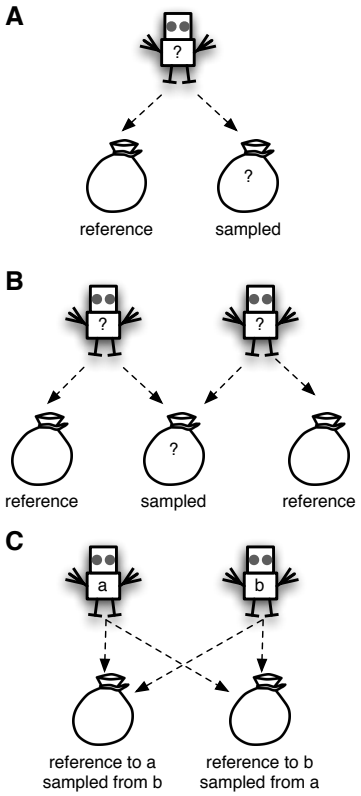


Fig. 1. Different Competitive Situations

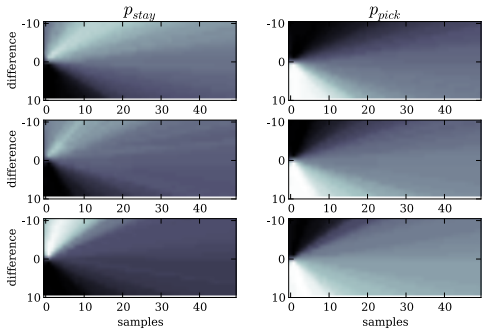


Fig. 2. Probabilities of actions: The probabilities to stay (p_{stay}) on the left, and the probability to pick (p_{pick}) on the right, mapped over the range of differences $[-10, 10]$ between urns (y axis) and number of samples drawn $[0, 50]$ as gray scales. The probability to continue is implicit since it is $1 - p_{pick}$. White represents high probabilities, black low. At the **top** results for indirect competition (See Figure:1A), in the **middle** results for direct competition (See Figure:1B),and at the **bottom** results for extreme competition (See Figure:1C).

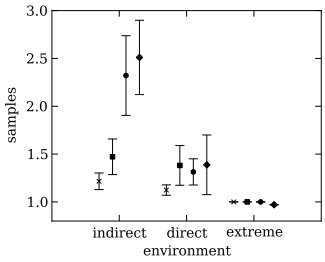


Fig. 3. Sampling Time: Average number of samples the representative agents took, for each of the three different environments: indirect (left), direct (middle), and extreme (right). Each environment was tested using four different variances on the distribution of the urns. 'X' indicate a variance of 0.1, squares indicate a variance of 1.0, circles for a variance of 3.0, and diamonds represent a variance of 5.0. The error bars indicate two standard errors.

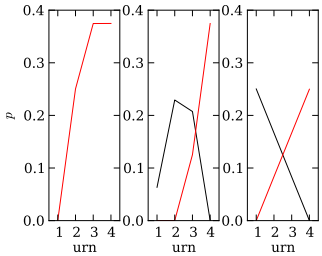


Fig. 4. Probability that each evolved strategy ended up choosing an urn. Left most figure for indirect competition. Here each player samples alone, and the probability (p) that an evolved strategy decides to pick an urn of a given size (x-axis) in red. The middle figure shows the probability of a player to pick an urn in the direct competition environment, assuming that player always wins playing against an identical player in red, and in black assuming the same for a player that always looses against an identical competitor. On the right the same for evolved players in the extreme competitive game.

Fig. 5. Supplementary Information about the outcome of choices. In the indirect competitive environment that agent is randomly presented with two different urns, and thus can experience the following six possible scenarios: 1-2, 1-3, 1-4, 2-3, 2-4, 3-4. Choosing the best in all cases sums to 20, and taking into account that each scenario has the same probability to appear we find a mean payoff for optimal choosing to be $3.\bar{3}$. In case of choosing always the worst, the sum is 10, and consequently the expected least payoff on average is $1.\bar{6}$. In the direct competitive environment, the two agents are presented with four possible scenarios of payoff in the three urns: 1-2-3, 1-2-4, 1-3-4, 2-3-4. The agent choosing the optimum will receive 15 in total, and because each of the four scenarios occurs with the same probability 3.75 is on average the expected maximum payoff. The second agent could now choose the remaining second highest urn which amounts to 10. The expected best mean payoff for the second agent choosing optimally is therefor 2.5. In the extreme competitive environment the best choosing agent, similar to the indirect environment can choose the best of two option resulting in $3.\bar{3}$ as the expected average maximal payoff. In turn the opponent would then be left with the lower urn, resulting in $1.\bar{6}$ as the expected payoff. Randomly choosing would result in 2.5 since any of the four options would be chosen with equal probability.

How The Inner Crowd Can Help Non-Bayesians Become More Bayesian

Nathaniel D. Phillips, Stefan M. Herzog, and Ralph Hertwig

Max Planck Institute for Human Development, Berlin, Germany

Author Note

Nathaniel D. Phillips, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development; Stefan M. Herzog, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development; Juliane Kämmer, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development; Ralph Hertwig, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development. Nathaniel D. Phillips is now at Social Psychology and Decision Sciences, University of Konstanz.

This research was supported in part by grants form the Swiss National Science Foundation 100014_129572/1 to Stefan M. Herzog and Ralph Hertwig.

Correspondence concerning this article should be addressed to Nathaniel Phillips, Department of Social Psychology and Decision Sciences, University of Konstanz, 78457 Konstanz, Germany. Email: nathaniel.phillips@uni.konstanz.de

Abstract

Can people become more Bayesian by contradicting themselves with dialectical bootstrapping (i.e., simulating the wisdom of a diverse crowd within their minds)? In a simulation study, we found that if people were to average non-Bayesian strategies when making diagnostic judgments, their answers would come closer to the Bayesian answer. Accuracy improved most in problems with a rare target event (i.e., p[hypothesis] < 0.2) and when combining strategies that did and did not use base rates, respectively (e.g., averaging the base rate, p[hypothesis], with the likelihood of the data under the target hypothesis, p[data|hypothesis]). In two empirical studies, we found that contradicting oneself with dialectical bootstrapping (Herzog & Hertwig, 2009) increased the diversity of strategies used (as assessed by formal strategy classification analyses). As diversity increased, averaging gains increased as well. This research suggests that people can use their eclectic inner crowd to become more Bayesian even without any explicit knowledge of Bayes' rule.

Keywords: Bayesian reasoning, probabilistic inference, dialectical bootstrapping, inner crowd, crowd within, wisdom of crowds

How The Inner Crowd Can Help Non-Bayesians Become More Bayesian

In order to make good decisions, organisms must find ways to successfully navigate a world of both risk and uncertainty, where information about their internal and external environment is incomplete and outcomes are impossible to predict with perfect precision. One way that organisms can make better decisions in unpredictable worlds is by relying on information that is probabilistically related to important outcomes, states, or hypotheses (H, Brunswik, 1943). For example, consider a mouse that is deliberating whether or not to move to a new foraging patch. To help its decision, the mouse would benefit from knowing if there are predators lurking in the shadows of the new patch. If the 'predator hypothesis' were false, then the mouse could benefit from the move; however, if the predator hypothesis were true, the mouse should avoid the patch. To help make its decision, the mouse could rely on a social cue in the environment; for example, "Are there other mice foraging in the new patch?" If the answer is 'no' then the mouse may infer that the probability of a predator is more likely than if the answer was 'yes,' and use this information to inform its decision.

While this social cue should be useful, it is only probabilistically related to the predator hypothesis and cannot give a definitive answer. How should the mouse, or any organism, make good inferences about important hypotheses given noisy, probabilistic cues? Statistically, the normative solution to this class of estimation problems is Bayes theorem. Experimental psychologists have compared human probabilistic inference to Bayes theorem since the 1960s. The result has been an ongoing debate regarding whether or not people act as "intuitive Bayesians," with some researchers suggesting that peoples' judgments are correlated with Bayes but are too "conservative" (Edwards, 1968) and

others claiming that people are "not Bayesian at all" (Kahneman & Tversky, 1972, p. 450). In the current paper, we argue that both camps are (figuratively speaking) arguing for black or white answers to gray problems; rather than using a single strategy, we show that people use a variety of strategies (both across people and within one mind) that differ in their accuracy relative to Bayes theorem. We then suggest that people can become more Bayesian by generating and combining this diverse set of non-Bayesian strategies within one mind. At the same time, we use the Bayesian reasoning paradigm to model how *dialectical bootstrapping* (Herzog & Hertwig, 2009) produces strategy change within one mind.

The paper proceeds as follows: First, we review research on how people solve Bayesian reasoning tasks and find that people are not generally "conservative Bayesians," nor are they "not Bayesian at all." (Kahneman & Tversky, 1972, p. 450) Rather, people use a variety of 'intuitive' strategies (Gigerenzer & Hoffrage, 1995; McKenzie, 1994) that vary in their complexity and errors relative to Bayes theorem. We suggest that, in the same way groups of error-prone individuals can produce surprisingly accurate judgments (Larrick, Mannes, & Soll, 2012; Surowiecki, 2004), individuals may be able to harness an inner crowd (Herzog & Hertwig, 2014a) of intuitive strategies to improve their accuracy in Bayesian reasoning tasks. We then review recent research on the wisdom of crowds within one mind, and suggest that *dialectical bootstrapping* (Herzog & Hertwig, 2009), a method of increasing the benefits of the inner-crowd, could increase strategy diversity in probability estimation tasks. To make predictions for when people should benefit from the inner-crowd, we conduct a simulation where we average non-Bayesian strategies and compare their errors in different environments. From this, we find both environmental

and strategy characteristics that are especially conducive to averaging. Finally, we present results from two experiments that test the extent to which participants can, and do, benefit from averaging non-Bayesian strategies. We conclude that people indeed can become more Bayesian by harnessing their inner crowd.

Bayesian Reasoning

In a typical Bayesian reasoning task, participants are asked to estimate the probability of a hypothesis (or event) given three pieces of information: base-rate (i.e., the prior probability of the hypothesis; BR), hit-rate (i.e., the likelihood of data given the critical hypothesis; HR), and false-alarm rate (i.e., the likelihood of data given an alternative hypothesis; FAR). For example, consider a doctor whose patient has non-specific symptoms. The doctor could have the *critical* hypothesis "This patient has malaria" and an *alternative* hypothesis "This patient does *not* have malaria." In a Bayesian framework, judges always begin with an *a priori* belief in the critical hypothesis before observing any additional information. This *a priori* belief is captured by the critical hypothesis' base-rate. In this example, the base-rate could be the proportion of people in the doctor's patient population with malaria. The judge can then observe some data, such as a diagnostic medical test, that probabilistically differentiates between the critical hypothesis and the alternative hypothesis. For example, in testing for malaria, a doctor can examine a patient's blood under a microscope and look for evidence of the *Plasmodium falciparum* parasite. If the data is more likely given the critical hypothesis than the alternative hypothesis, then the judge's posterior probability estimate should increase relative to the base rate. This likelihood of getting a signal consistent with the critical hypothesis is given by the hit-rate: for example, the probability of

obtaining a positive test for malaria given that the patient really has the disease (i.e., the malaria hypothesis is true). Finally, the false-alarm rate indicates the likelihood of getting a signal consistent with the critical hypothesis given the critical hypothesis is actually false. For example, the probability of obtaining a positive test result given that the patient really does *not* have malaria. Once a judge has these three pieces of information, Bayes theorem specifies that a judge should combine them via Bayes rule:

$$Bayesian\ Posterior\ Probability = \frac{BR \cdot HR}{BR \cdot HR + (1 - BR) \cdot FAR} \qquad \text{Equation 1.}$$

(Equation 1) to calculate a posterior probability (but see Birnbaum, 1983; Gigerenzer, 1996).

"Conservative Bayesians" vs. "not Bayesian at all". Psychologists have compared human judgments to Bayes theorem since the 1960s. Early work by Edwards and colleagues (Edwards, 1968; Phillips & Edwards, 1966) measured peoples' probability estimates using a "poker chips and bags" paradigm. Participants imagined that random chips were drawn from one of several bags containing different combinations of red and blue poker chips and were asked to estimate the probability that each type of bag had been chosen as more randomly drawn chips were revealed. From this work, Edwards and found concluded that participants act as 'conservative Bayesians', in that they overweighed base-rate information and did not adjust their estimates sufficiently in light of new information (i.e., hit rates and false-alarm rates). In a review of related literature, Peterson and Beach (1967) concluded that, despite conservatism and other deviations from normative statistical rules, Bayes theorem and other normative rules provided a useful initial descriptive model of human statistical inference.

The claim that normative statistical rules such as Bayes theorem provided adequate descriptions for human inference was quickly challenged. Kahneman and Tversky (1972, 1973) compared human judgment to normative models and claimed to have found striking discrepancies between the two. For example, while Bayes theorem requires attention to base-rate information, Kahneman and Tversky (1972) claimed that people judged as if they ignored base-rates. They concluded that, instead of being intuitive statisticians or conservative Bayesians, people are users of biased heuristics such as the representativeness heuristic (Kahneman & Tversky, 1972), which violate normative rules by ignoring normatively relevant information such as base rates (i.e., exhibiting "base-rate neglect") and sample sizes. In sharp contrast to Peterson and Beach (1967), Kahneman and Tversky (1972) concluded that "In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all" (p. 450).

Debate over the "Bayesian man" continued to rage over the next decades, with particular attention to whether or not people use base-rate information (e.g., Barbey & Sloman, 2007; Koehler, 1996). While earlier research claimed to routinely find base-rate neglect (e.g., Lyon & Slovic, 1976), more recent research found that people routinely use base-rates (Christensen-Szalanski & Bushyhead, 1981; Stanovich & West, 1998; Pennycook & Thompson, 2012; for a summary, see Koehler, 1996). The current body of research seems to suggest that, rather than categorically ignoring base-rates, people seem to selectively use base-rates based on factors such as the causal nature of base-rate information (Ajzen, 1977), the perceived relevance of base-rates (Bar-Hillel, 1980), the verbiage of vignettes (Macchi, 1995), and the reasoning system people use (Barbey & Sloman, 2007).

Strategy variability in Bayesian reasoning. So who between the rival camps of Edwards vs. Kahneman and Tversky were right? Are people conservative Bayesians who rely too much on base-rate information? Or do they not Bayesian at all because they neglect base rates entirely? Rather than relying on a single strategy with a consistent bias, we argue that there is strategy variability both between and within people. To demonstrate this, the following example vignette from Bar-Hillel (1980, p. 228):

> Two cab companies operate in a given city, the Blue and the Green (according to the color of cab they run). Eighty-five percent of the cabs in the city are Blue, and 15% are Green. A cab was involved in a hit-and-run accident at night, in which a pedestrian was run down. The wounded pedestrian later testified that though he did not see the color of the cab due to the bad visibility conditions that night, he remembers hearing the sound of an intercom coming through the cab window. The police investigation discovered that intercoms are installed in 80% of the Green cabs, and in 20% of the Blue cabs. What do you think are the chances that the errant cab was Green?

A Bayesian reasoner would answer this problem by extracting the base-rate (15%), hit-rate (80%) and false-alarm rate (20%) information, and combining it using Bayes theorem to arrive at a posterior probability estimate of 41.38% (rounded to two decimal places). In Figure 1, we present the distribution of responses to this Bayesian reasoning problem from Bar-Hillel (1980). Responses between 15% and 41.38% represent conservatism (over-weighting of base-rates), while responses between 41.38% and 80% represent anti-conservatism (under-weighting of base-rates). A response of 80% (i.e., the hit rate) could be considered the quintessential base-rate-neglect answer.

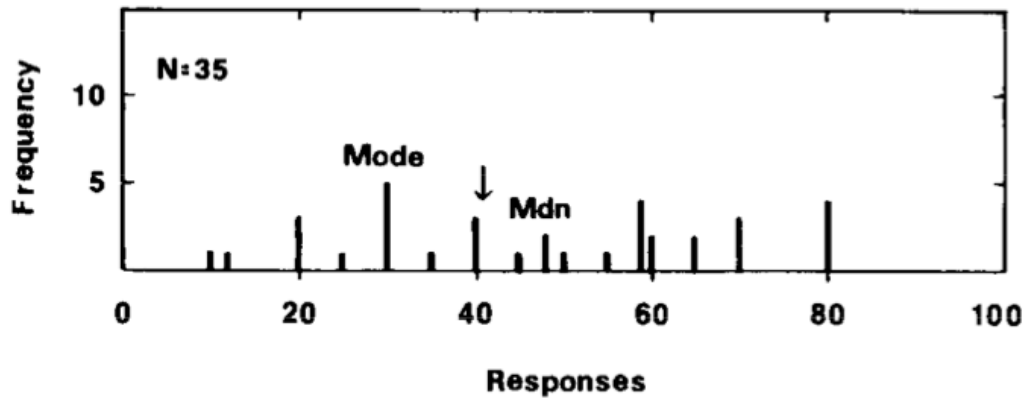


Figure 1: Distribution of responses to Bar-Hillel's (1980) problem 7 (Figure 4 in Bar-Hillel, 1980). The downward facing arrow at a response of 41.38% indicates the Bayesian posterior probability. The median-participant response is indicated by the letters "Mdn" at 48%.

The diversity of responses in Figure 1 shows that people are not conservative Bayesians, nor are they not Bayesian at all—they act as if they use a variety of strategies[1]. We are not the first to observe this. Upon observing the estimate variability in Figure 1, Bar-Hillel remarked that "[T]here is no prevailing strategy or integration favored by a large proportion of the [participants]" (1980, p. 228). Thus, rather than simply being conservative Bayesians or habitual base-rate neglectors, it seems that people use different strategies that can either result in errors of different sign and magnitude. This leads us to our key research question: If people use a variety of simple strategies for Bayesian

[1] We do not mean to argue that every psychological theory that does not explicitly account for response variability is immediately falsified by said variability. Rather, we argue that the wide, systematic response variability in Bayesian reasoning tasks suggests that Kahneman and Tversky's (1972) and Edward's (1968) conclusions are inadequate at best and misleading at worst.

estimation, could they benefit from averaging those strategies within their own mind (cf. Herzog & Hertwig, 2009)? Indeed, we can find some suggestive evidence from Bar-Hillel's (1980) existing data that indeed they can. Bar-Hillel's (1980) data in Figure 1 show a *wisdom of crowds* effect (Surowiecki, 2004): while the Bayesian response (indicated by a downward facing arrow located at a response of 41%) is not a common response, the median participant estimate of 48% is quite close to Bayes. To what extent is this effect representative of other Bayesian reasoning tasks and how could an individual replicate these gains? We address this question in the context of research on the inner-crowd (Herzog & Hertwig, 2014a).

The Wisdom of The Inner Crowd

People can improve accuracy of their quantitative judgments by harnessing a wisdom of crowds within one mind (Herzog & Hertwig, 2014a) in the same way that groups of separate people gain crowd benefits (Armstrong, 2001; Hogarth, 1978; Larrick et al, 2012; Surowiecki, 2004). The process works as follows: an individual generates an initial estimate (x_1) to a problem with an unknown answer T. The person then, possibly after an intervention such as a time delay (Vul & Pashler, 2008) or dialectical processing instructions (Herzog & Hertwig, 2009; Herzog & Hertwig, 2014b; Phillips et al., 2014), generates a second estimate (x_2) to the same problem. Finally, the individual estimates x_1 and x_2 are combined into an average estimate x_{avg12} – usually the arithmetic mean of x_1 and x_2. When the error of the average estimate x_{avg12} is less than the error of the original estimates (typically their initial estimate x_1), the person benefits from their inner crowd (Herzog & Hertwig, 2009).

All crowds, regardless if they derive from one person or many, derive their benefits from error cancellation (Herzog & Hertwig, 2014a). When two estimates x_1 and x_2 fall on different sides of ('bracket') the truth T, the average estimate x_{avg12} will always have a smaller error than the average error of the original estimates (Larrick & Soll, 2006). In cases where the initial estimates have large opposing errors, the accuracy of x_{avg12} can even be higher than the accuracy of the *best* original estimate. For this reason, a crowd can be 'wiser' than even its most accurate individual member.

Dialectical bootstrapping. Because error cancellation drives the wisdom of crowds effect, researchers have tried different methods to get people to generate a diverse set of estimates, including increasing the time delay between first and second estimates (Vul & Pashler, 2008; but see also Steegen, Dewitte, Tuerlinckx, & Vanpaemel, in press). In the current paper, we focus on *dialectical bootstrapping* (Herzog & Hertwig, 2009), which attempts to increase inner-crowd benefits by having participants generate estimates using different knowledge that in turn should increase estimate diversity. By using a dialectical instructions inspired by the consider-the-opposite technique (Lord, Ross & Lepper, 1984), dialectical bootstrapping has been found to increases estimate diversity and improves averaging gains in the inner-crowd in both general knowledge tasks (Herzog & Hertwig, 2009; 2014b), and multiple-cue judgment tasks (Phillips, Herzog, Kämmer & Hertwig, 2014).

While dialectical bootstrapping has shown promise in these tasks, the specific estimation procedures people use, and the way people change their estimates from one phase to another, has been less clear. Judgment research has proposed a wide variety of estimation models, from rule-based (e.g.; Hertwig, Hoffrage & Martignon, 1999; von

Helversen & Rieskamp, 2008) to exemplar models (Juslin, Winman & Hansson, 2007; Lindskog, Winman & Juslin, 2013; Stewart, Chater & Brown, 2006). However, no research on the inner crowd has attempted to model the specific estimation strategies people use, and how people change strategies as a result of dialectical instructions. This is a substantial gap because without knowing which strategies people use, and how people switch from one strategy to another, we cannot predict the conditions (i.e., judgment tasks) under which people will benefit from the inner-crowd. In the current paper, we seek to fill this gap by modeling strategy use in the inner-crowd in Bayesian reasoning tasks while simultaneously measuring how people can use dialectical bootstrapping to improve their judgments in Bayesian reasoning tasks.

Ecological rationality of averaging the inner crowd. Using the simple average (i.e., arithmetic mean) of a group is often a wise aggregation strategy (Davis-Stober, Budescu, Dana, & Broomell, 2014); however, the accuracy of averaging compared to alternative strategies, such as weighted-averaging or choosing, depends on several statistical properties of the judges and the estimation environment. The Probability, Accuracy and Redundancy (PAR; Soll & Larrick, 2009) model specifies three of these statistics where one must decide how to aggregate advice from two judges: the *probability* of detecting the most accurate judge, the relative *accuracy* of one judge to the other, and the degree of error *redundancy* (i.e., error correlation) between judges. As the first two statistics (probability and accuracy) increase and the third statistic (redundancy) decreases, the benefits of the simple average relative to other strategies decrease. Unless one judge is much more accurate than the other, one can easily identify the more accurate

judge and there is a low correlation between the judges' errors, then the simple average will be more accurate than choosing the estimates of the judge deemed more accurate.

Thus, the accuracy of the average depends on the statistical *environment* to which it is applied. In applying environmental considerations to Bayesian reasoning tasks, we define an environment as a distribution of *cue profiles* that satisfy certain criteria, where a cue profile is a combination of the three statistics BR, HR and FAR. For example, a reasoning task with a base rate of 20%, a hit rate of 70% and a false-alarm rate of 10% corresponds to the cue profile of [20%, 70%, 10%]. Conceptually, we view environments as different 'worlds' of cue profiles that strategies can be applied to.

McKenzie (1994) accessed the accuracy of individual non-Bayesian strategies in Bayesian reasoning tasks, and found that while many ignore one of the three key statistics, some nonetheless performed reasonably well relative to Bayes rule. In particular, the strategy that averaged base-rate and hit-rate information performed extremely well[2]. However, McKenzie (1994) also found that the accuracy of strategies could vary dramatically depending on the statistical environment they are applied to. Generally, strategies that ignore base-rate information tend to do reasonably well in domains with moderate base-rates (e.g.; between 40% and 60%), but quite poorly in domains with extreme base-rates (e.g.; less than 10% or greater than 90%).

In our analyses, we expand on McKenzie's (1994) results by calculating the accuracy of the average of pairs of intuitive strategies in two different statistic environments. We call the first stimuli environment "*Valid Cue* (VC)." In the VC environment, base rates range from 0 to 100%, hit rates range from 0% to 100%, and

[2] Indeed, Mckenzie's (1994) finding that an averaging strategy performed quite well was one of the inspirations for the current paper.

false-alarm rates are strictly less than hit-rates[3]. We call this environment "Valid Cue" because it contains a cue whose hit-rate is larger than its false-alarm rate, and thus should normatively be used to change beliefs. This stimuli environment has a relatively high level of uncertainty, because it is difficult *a priori* to predict what the cue-profile will be for any specific question in the environment.

Many interesting diagnostic problems involved the detection of a rare event (such as a rare medical condition) based on an imperfect, but informative cue (such as a medical test). To model the effects of intuitive strategy combination in these tasks, we generated a "*Rare Event plus Valid Cue* (RE+)". In this environment, base-rates range from 0% to 20%, hit-rates range from 80% to 100%, and false-alarm-rates range from 0% to 20%[4]. We are particularly interested in the effectiveness of averaging in RE+ environments for two reasons: First, because cue profiles in the RE+ environment have extreme base-rates, they lead to large errors when people ignore base-rates (Kahneman & Tversky, 1972). Second, because statistic values in RE+ environments are so extreme, we expect strategies to have large biases in RE+ environments. To the extent that different strategies have large opposing biases, this could lead to substantial averaging gains.

Applying the Inner Crowd to Bayesian Reasoning Tasks

In the next section, we report results from a simulation where we test the benefits of averaging non-Bayesian estimation algorithms in the two different statistical environments. We seek to answer four key questions in the simulation: First, to what

[3] In cases where hit-rates are smaller than false-alarm rates (e.g., a faulty smoke alarm that is more likely to activate when smoke is *not* present than when it is present), we assume that judges would reverse the interpretation of the cues, forcing hit rates to be larger than false-alarm rates.

[4] The 20% and 80% cut-off values we used are somewhat arbitrary. We chose them because they separated our stimuli in study 1 into two relatively equal sets. In a similar simulation, McKenzie (1994) used cut-offs ranging from 10% and 90%. Our general conclusions also hold for the cut-off values used by McKenzie.

extent does averaging non-Bayesian strategies decrease error? Second, do all strategies equally benefit from averaging or are some strategies more conducive to averaging than others? Third, are there specific environments that favor averaging? Finally, does the use, or non-use, of base-rate information predict the extent to which averaging benefits strategies?

Simulation Study

We began by constructing a list of simple, non-Bayesian strategies that have been proposed in two previous papers (Gigerenzer and Hoffrage, 1995; McKenzie, 1994). We extracted seven non-Bayesian, non-averaging[5] strategies from these papers and present them in Table 1.

Strategy Number	Long Name	Short Name	Formula	Uses Base Rate?
1	Likelihood	Hit.Rate	HR	No
2	False Alarm Complement	cFAR	1 – FAR	No
3	Relative Likelihood	Rel.Lik	HR / (HR + FAR)	No
4	Likelihood Subtraction	Lik.Sub	HR – FAR	No
5	Base Rate	BR	BR	Yes

[5] Two averaging strategies (i.e.; those that compute the average of cue values or average of strategies), called "Likelihood Average" and "Relative Likelihood Average" have been proposed in the literature. For our simulation, because we are interested in seeing how the accuracy of individual, non-averaging, strategies changes as a result of averaging, we did not include these two averaging strategies in the simulation. However, we do include them in our behavioral classification analyses in studies 1 and 2.

6	Joint Occurrence	Joint	BR * HR	Yes
7	Hit Rate Minus Base Rate	HRmBR	HR – BR	No*

Table 1: Description of seven strategies used in our simulation. Strategy 7 (Hit rate minus base rate) technically uses Base rate information but in a non-Bayesian manner as it *decreases* the posterior probability of the hypothesis as the base-rate of the hypothesis increases.

We would like to briefly emphasize the psychological relevance of three of these strategies. Strategy 1 (Likelihood), which uses only hit rate information, is one instantiation[6] of the representativeness heuristic (Kahneman & Tversky, 1972). Strategy 2 (False Alarm Complement) amounts to what many people do when they confuse null-hypothesis p-values and the Bayesian posterior probability of the null hypothesis (Gigerenzer, 2004). Finally, Strategy 5 (Base Rate) amounts to an extreme version of conservatism (Edwards, 1968).

The seven strategies in Table 1 depart from Bayes theorem in two fundamental ways. First, they differ in which cues they use, with different strategies using different combinations of the three statistics. Importantly, these intuitive strategies are frugal as none of them use all three statistics as Bayes theorem requires. Second, the strategies differ in how they combine the statistics. Two of the strategies (BR only and HR only) use only single statistics, one strategy (FAR complement) takes the complement of a

[6] Though as Gigerenzer (1996) points out, the representativeness heuristic is difficult to precisely define.

statistic, and two take the difference between two statistics (HR minus BR and Likelihood subtraction). Only two of the strategies include multiplicative operations involving the base rate (as Bayes theorem requires): The Joint Occurrence strategy multiplies two statistics, while the Relative Likelihood strategy combines addition and division operations.

Simulation Procedure and Results

The simulation procedure was modeled after that of McKenzie (1994) who also simulated the performance of non-Bayesian strategies[7]. However, in contrast to McKenzie (1994), who defined accuracy using correlation measures, we define strategy performance using absolute deviation (and mean absolute deviation, MAD)[8].

First, we generated a matrix containing all combination of the three (BR, HR and FAR) cues' values from .01 to .99 in steps of .01, leading to 1,000,000 cue profiles. Next, for each cue profile, we calculated the estimate for each of the 7 strategies listed in Table 1. We then calculated both the signed error (bias) and absolute error (accuracy) between each strategy's estimate and the corresponding Bayesian posterior probability. Note that because we applied each strategy without error, there was no random sampling involved in any of our calculations.

[7] Our simulation departs from that of McKenzie (1994) in one key respect. McKenzie defined each cue profile in the form of a 2 x 2 frequency contingency table. He then created a set of cue profiles by creating all combinations of cell frequencies, where each cell contained a number from 1 to 50. As we show in Appendix D, this procedure leads to a non-uniform distribution of base-rates, hit-rates, and false-alarm rates across cue profiles, with a peak in at 0.50. Additionally, it allows for hit-rates to fall below false-alarm rates, which we expect will not hold for most real-world tasks. To correct for these issues we manipulated statistics directly instead of using McKenzie's frequency contingency table approach.

[8] While correlation measures are informative, they can mask large absolute differences between a strategy and Bayes and do not reveal strategy biases. To illustrate, a strategy can be perfectly correlated with Bayes while simultaneously having a large bias.

Bias of non-Bayesian strategies . We begin by exploring the accuracy of individual strategies. Signed error distributions for each strategy separated by the two statistic environments are displayed in Figure 2:

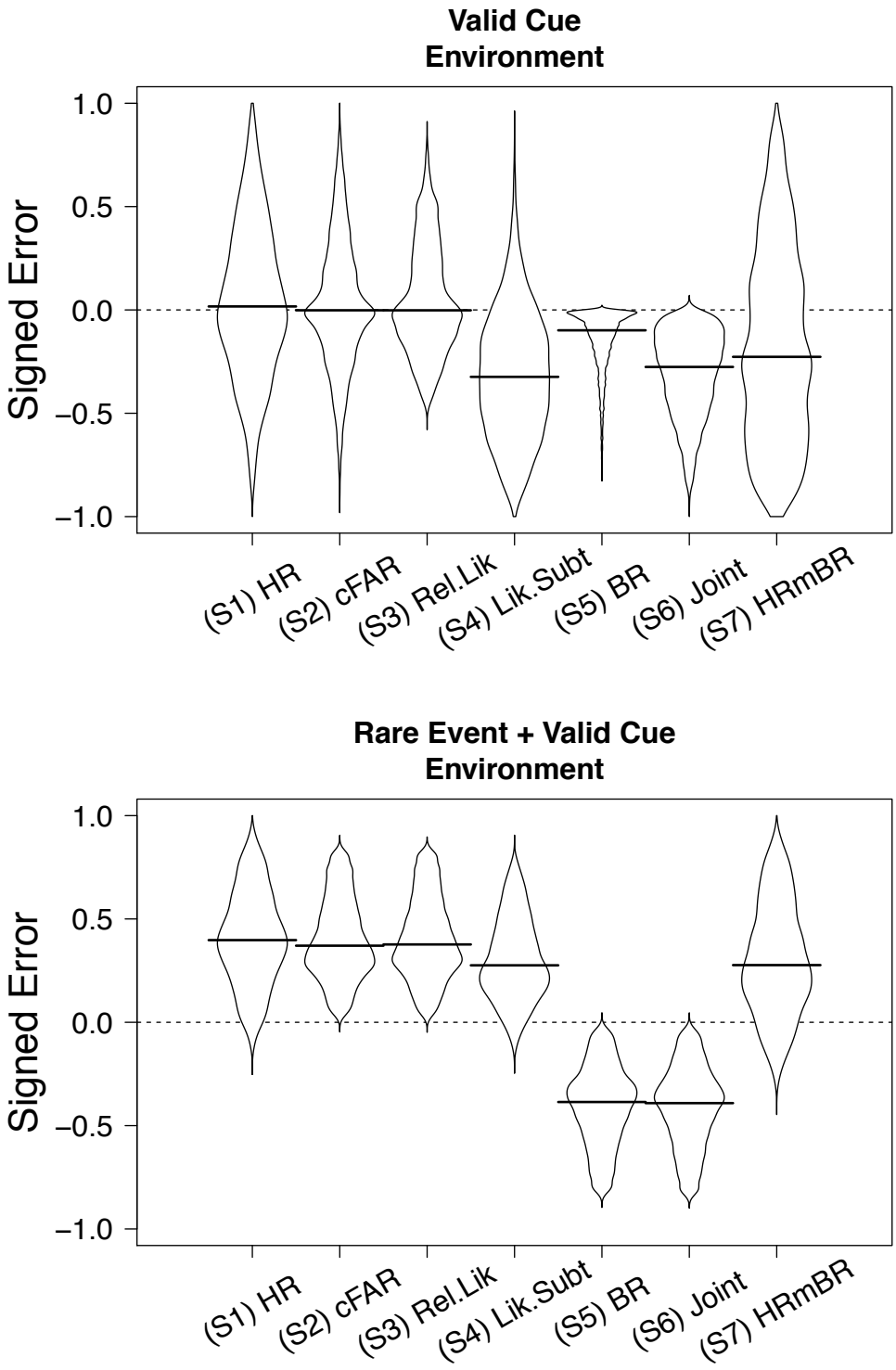


Figure 2: Distribution of signed errors for individual strategies in the VC (top panel) and RE+ (bottom panel) statistic environments respectively. Each figure in

the plot is a "bean": a smoothed and vertically mirrored density (Kampstra, 2008). Horizontal lines represent medians.

We begin by examining error distributions in the VC environment (Figure 2, top panel). Here, some strategies (Joint (S6), Likelihood Subtraction (S4), and Hit Rate minus Base Rate (S7)) tended to have moderately large negative biases, meaning that they tended to give estimates lower than Bayes, while the remaining strategies had small biases. Moreover, most strategies had large error ranges, with some cue profiles leading to large negative errors and others leading to large positive errors. The one exception was the Base Rate (S5) strategy, which tended to give small, negative errors across most stimuli profiles. Clearly, in this environment, the Base Rate (S5) strategy appears to be both frugal and moderately accurate relative to other simple strategies.

A different picture emerged in the RE+ environment. In this environment, we see clearer biases in strategies. All but two strategies (Joint (S6) and Base Rate (S5)), systematically overestimated the posterior probabilities (i.e.; had a positive bias). In contrast, Joint (S6) and Base Rate (S5) overwhelmingly estimates that were too (i.e.; had a negative bias). Compared to VC environments, the Base Rate (S5) strategy does substantially poorer in RE+ environments.

These results allow us to make some predictions as to which environments favor averaging, and which pairs of strategies will benefit from averaging in which environment. Because strategies are more likely to have large and opposing biases in the RE+ environment, we expect larger averaging gains here than in the VC environment. Second, in the RE+ domain we expect that combining strategies with negative biases

(such as Joint (S6) and Base Rate (S5)) with strategies with positive biases will lead to greater averaging gains than combining strategies with similar biases. In the next section, we compare the accuracy of individual strategies with averaging strategies.

Accuracy of averaging strategies. Next, we compared the accuracy of the individual strategies in Table 1 to the averages of pairs of strategies. We label averaging strategies using the convention S_{iaj}, where *i* and *j* are the two strategies being averaged. For example, the strategy S_{1a2} takes the average of Likelihood (S_1) and False Alarm Complement (S_2). We generated estimates for all 21 (7 choose 2) averaging strategies by taking the average estimate of each pair of strategies for each stimuli profile. Next, we calculated the absolute difference between the estimates of the averaging strategies and the Bayes posterior probability for all cue patterns in both stimuli environments. We present full absolute error distributions of each strategy in Figure 3:

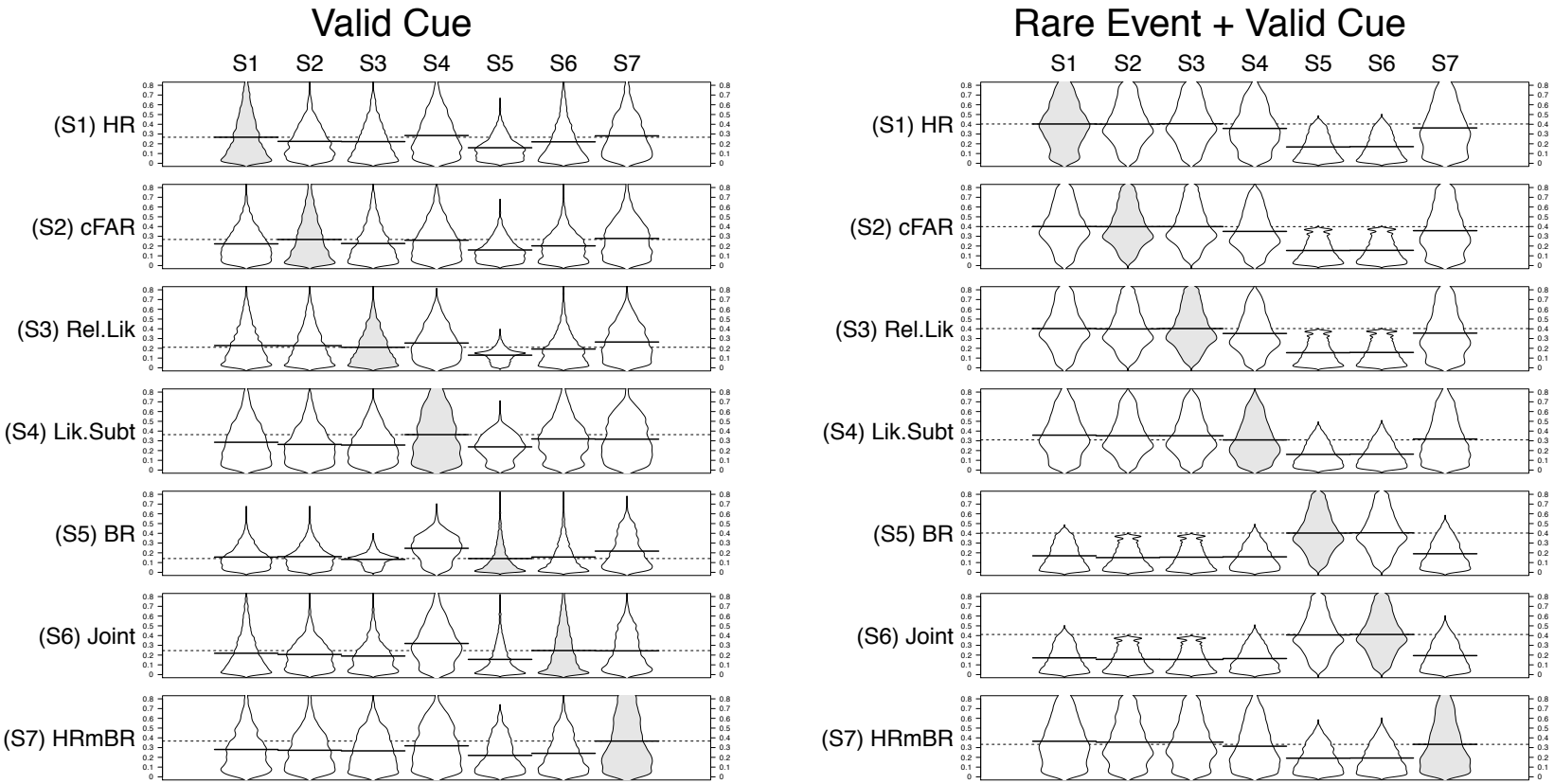


Figure 3: Distributions of absolute deviations for all averaging strategies across Valid Cue (VC, left panel) and Rare Event Plus Valid Cue (RE+, right panel) environments. Each figure in the plot is a "bean": a smoothed and vertically mirrored density (Kampstra, 2008). Lower values indicate better accuracy. Horizontal bars in each bean show median values. The

intersection between a row and column strategy presents the distribution of errors for the average of those two strategies. Distributions on the main diagonal (in gray) represent individual strategies. The horizontal dotted line in each row shows the mean absolute deviation of the individual strategy corresponding to that row.

In Figure 3, the intersection of a row and column shows the distribution of absolute errors for the average of the two strategies named in the respective row and column across all stimuli profiles within an environment. Distributions on the main diagonal correspond to individual strategies while distributions on the off diagonal represent averaging strategies.

We begin by looking at the VC environment (left panel of Figure 3). In the VC environment, some strategies tend to be aided by averaging while others tend to be hurt. For example, the individual strategy Hit Rate minus Base Rate (S7) tends to have large errors, while averaging strategies using Hit Rate minus Base Rate (S7) appear to have lower absolute errors. Thus, the Hit Rate minus Base Rate strategy (S7) lends itself to averaging in VC environments. In contrast, the Base Rate (S5) strategy tends to perform well on its own. Averaging strategies that use Base Rate (S5) tend to have larger absolute errors than the individual Base Rate (S5) strategy. Thus, the Base Rate strategy does not tend to benefit from averaging in valid cue environments.

In the RE+ environments, by contrast, the mass of most averaging strategies tend to be lower than that of the individual strategies, suggesting that averaging usually improves accuracy. For example, the Base Rate (S5) strategy tends to be less accurate on its own than when it is averaged with other strategies. This is in contrast to the VC environment where the strategy did fairly well on its own. Similarly, the Joint (S6) strategy appears to have lower errors when it is averaged with other strategies. These results suggests that, consistent with our predictions, the RE+ environment favors averaging more than the VC environment. To see this relationship more clearly, we

plotted the mean absolute deviation (MAD) of each solitary strategy relative to the MAD of all averaging strategies that contain that solitary strategy in Figure 4.

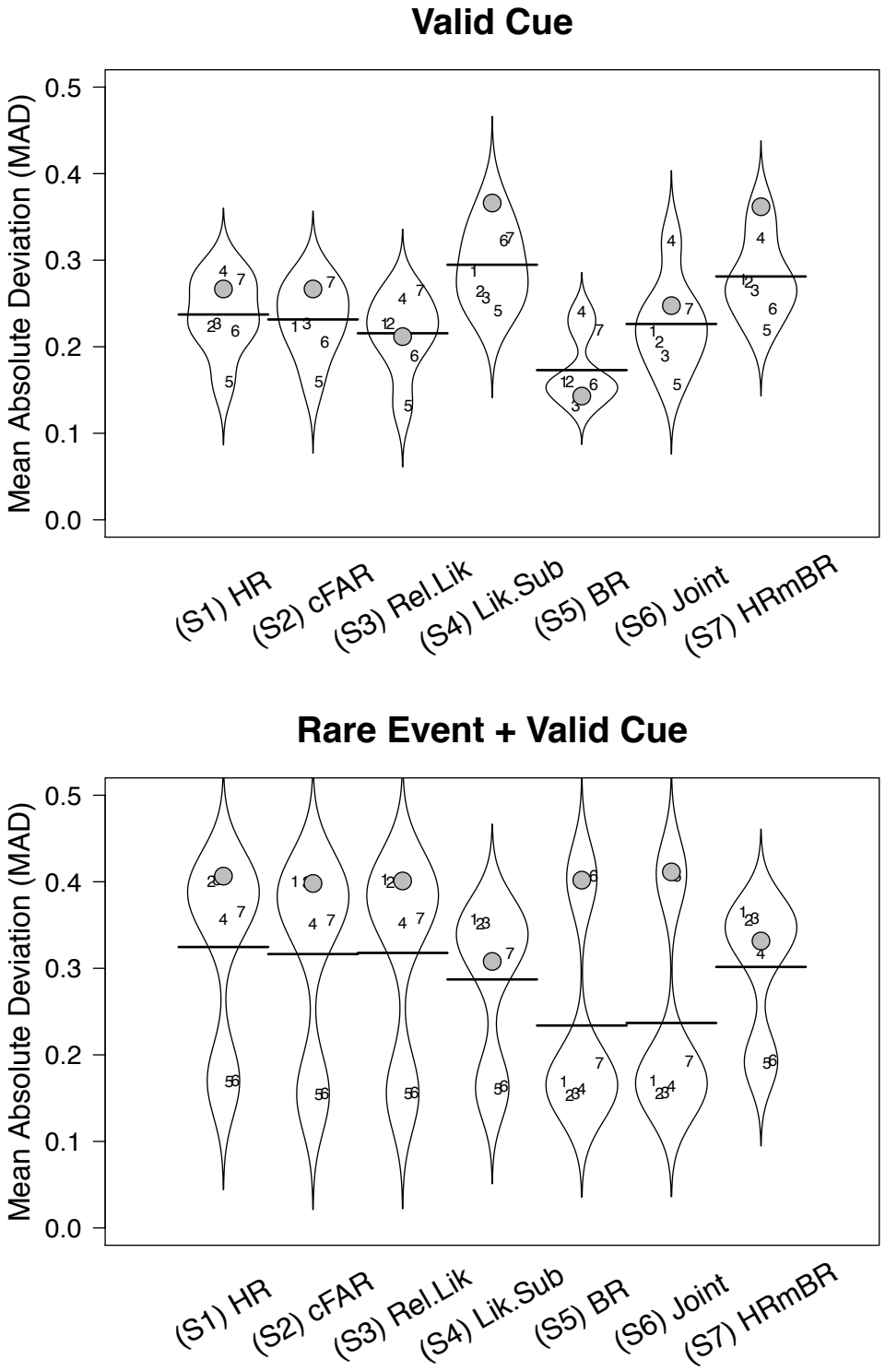


Figure 4: Mean absolute deviation (MAD) of individual strategies and combined strategies in two statistic environments. Each figure in the plot is a "bean": a smoothed and vertically mirrored density (Kampstra, 2008). Gray circles show individual strategies while numbers show combined strategies. Horizontal bars show the median MAD value for all averaging strategies. Each distribution shows the MAD values for all averaging strategies that use a specific "starting" strategy (indicated by the graph labels). The numbers indicate the MAD value of the averaging strategy that combined the starting strategy and the numbered strategy. For example, the number 7 in the far left plot in the left panel represents the MAD value of S1a7 - the averaging strategy that combines Likelihood (S1) and the Hit Rate Minus Base Rate (S7) strategy. The numbers are jittered horizontally to make them easier to read. The gray circles represent the MAD value for the individual starting strategy. When most numbers fall below the gray circle, then the strategy tends to benefit from averaging. When most numbers fall above the green dot, then the strategy tends to be hurt by averaging.

In the aggregated results presented in Figure 4, the distinction between VC and RE+ environments becomes more clear: In VC environments, some strategies tend to benefit from averaging (e.g., Likelihood Subtraction (S4) and Hit Rate Minus Base Rate (S7)) while others do not reliably benefit (e.g.; Relative Likelihood (S1)), and some are even generally hurt by averaging (e.g, Base Rate (S5)). Additionally, one can see that the number 5 falls below most of the green dots. This shows that the Base Rate strategy (S5)

is a good 'pairing' strategy that, when averaged, improves the performance of other strategies.

The picture in RE+ environments is very different. Here because the median MAD of averaging strategies is less than the individual strategy MAD for all strategies, *all* strategies tend to benefit from averaging. Additionally, strategies Joint (S6) and Base Rate (S5) appear to be especially beneficial strategies for other strategies to average with. For all other strategies, averaging with these strategies dramatically decreases average error relative to individual strategy performance. For example, when Likelihood (S1) is averaged with Base Rate (S5), Likelihood (S1) reaps a decrease in mean absolute error from .41 to .16, while Base Rate (S5) shows a decrease in MAD from .41 to .17.

Simulation conclusions. Our simulation showed that averaging gains depend on which strategies are combined and the environment strategies are applied in. In VC statistic environments, there is large variability between strategies in averaging gains: some reap large benefits (Likelihood Subtraction (S4) and Hit Rate Minus Base Rate (S7)), while others do not reliably benefit, and others are even hurt by averaging (Base Rate (S5)). In contrast, we find that in RE+ environments, *all* strategies tend to benefit from averaging. Additionally, we also find that two strategies, Base Rate (S5) and Joint (S6) appear to be especially conducive to averaging in that, when they are averaged with other strategies, they can lead to large reductions in error in one, if not both strategies. We can explain this with reference to whether or not strategies use base-rates. While strategies 1 through 4 do not use base rate information at all, Base Rate (S5) and Joint (S6) are highly influenced by base rates. By using base rate information, strategies Base Rate (S5) and Joint (S6) serve as promising 'helping strategies' that, when averaged,

increase the accuracy of other strategies. Strategy Hit Rate Minus Base Rate (S7) does use base-rate information, but not in a way consistent with Bayes theorem[9]. From this, we can predict that people who switch between using a strategy that does not use base rate information, and a strategy that does use base rate information (excluding Hit Rate Minus Base Rate (S7)), will reap larger averaging gains than those who do not.

Does Dialectical Bootstrapping Increase Strategy Diversity and subsequent averaging gains? Two Empirical Studies

To test the degree to which people could benefit from their inner crowd in a Bayesian reasoning task, we conducted two studies. In each study, participants gave estimates for several estimation problems. Each problem presented participants with a cue profile in the context of a vignette and asked them to estimate the Bayesian posterior probability. In study 1, we used a subset of the problems used by Gigerenzer and Hoffrage (1995). These stimuli contained vignettes from a wide variety of content domains (e.g., medical, eyewitness identification). Moreover, because some of the questions fell into the RE+ environment while others did not, the data allowed us to test whether statistic environments affect averaging gains within participants. In study 2, we created a new set of stimuli using a "balls and boxes" vignette across all questions. This standardized vignette was designed to minimize idiosyncratic differences between questions. In study 2, we constructed two separate sets of stimuli profiles that satisfied RE+ and VC criteria and included stimuli type as a between-participant manipulation.

[9] Hit Rate Minus Base Rate (S7) uses base rate information, but it combines it with other cues in a non-Bayesian manner. Instead of multiplicatively combining base rates with hit rates (as prescribed by Bayes theorem), it subtracts base rates from hit rates. In cue combinations with low base rates, Hit Rate Minus Base Rate effectively ignores the base rate cue and relies solely on the hit rate cue. Of course, this is completely antithetical to Bayes theorem's posterior probability, which is heavily influenced by small base rates. Thus, while Hit Rate Minus Base Rate (S7) uses base rate information, it does not help other strategies the way that Base Rate and Joint do.

In both studies 1 and 2, participants gave two separate estimates to each problem across two phases. After giving a first set of estimates in phase 1, participants were assigned to one of two conditions before giving a second set of estimates on phase 2. In the *control* conditions, participants were told to give a second set of estimates as if they were seeing the problems again for the first time (cf. Herzog & Hertwig, 2009, 2014b). We included this condition to measure baseline changes in strategy use between phases. In the *dialectical* conditions, participants read dialectical instructions, which encouraged participants to think of new ways of deriving estimates by challenging the assumptions they made in their previous estimates (Herzog & Hertwig, 2009). In previous studies on dialectical bootstrapping, dialectical instructions have been found to increase estimate diversity and subsequent averaging gains relative to control instructions. However, no previous study has used probabilistic modeling techniques (Lewandowsky & Farrell, 2010) to measure the effects of dialectical instructions on qualitative changes in strategies. In the current two studies, we use probabilistic modeling techniques to classify the estimation strategy participants used in each of the two phases as well as whether or not participants switched strategies. This classification analysis allowed us to test the effects of dialectical instructions on strategy switching, and the extent to which participants used strategies that, according to our simulations, should lead to averaging gains.

In study 1 we included an additional control condition where, after phase 1, participants spent four minutes solving anagrams unrelated to the reasoning task (*anagram* condition). We included this condition to test whether or not distraction from the Bayesian reasoning task was sufficient to produce strategy changes between

Distraction can provide an incubation period that increases creative problem solving (e.g., Baird, Smallwood, Mrazek, Kam, Franklin & Schooler, 2012). Thus, in order to test if distraction can match dialectical instructions in inducing strategy change, we included the anagram condition in study 1[10]. In study 2, we included an additional third estimation phase that was designed to test how people decide to aggregate their inner crowd (cf. Fraundorf & Benjamin, 2014; Herzog & Hertwig, 2014; Müller-Trede, 2011). For each question in this phase, instead of presenting the cue profile, we showed participants their phase 1 and 2 estimates for each of the problems and asked them to come up with their best estimate in light of those two estimates. We used these response data to determine if people were able to outperform the simple average of their phase 1 and phase 2 estimates.

Hypotheses

Because dialectical instructions should spur participants to think of a new way to make estimates (Herzog & Hertwig, 2009), we expect participants in this dialectical condition to be more likely to switch strategies between phases 1 and 2 (H1). Based on our stimulation study, we predicted that averaging gains would be larger in the RE+ environment than the VC environment (H2) and that participants who switch between a strategy that uses base-rates and another that does not use base-rates in phases 1 and 2 will reap larger averaging gains than participants who do not (H3).

Methods

Participants. 575 participants (350 in study 1 and 225 in study 2) were recruited from Amazon Mechanical Turk (mturk) participated in the study. They each received a flat payment of $3.50 for participating. In addition to the flat payment, they were eligible

[10] However, as the reader will see, participants in the anagram condition in study 1 actually changed strategies less often than those in the control condition. As this discredits the hypothesis that distraction was sufficient to produce strategy change, we did not include the condition in study 2.

for a performance-based bonus reward of up to $2.00. We restricted participants to those located in the United States with at least a 95% mturk work acceptance rate. We informed participants that the study would take between 60 and 90 minutes to compete.

Materials and procedure. In both studies, participants gave estimates to several questions that required them to make a posterior probability estimate in light of three statistical cues contained in a vignette. They were told that each question had a correct answer, and that they would receive a monetary bonus proportional to the agreement between their estimates and those that a statistician would give. The order of questions was randomized for each participant. After reading each question, participants gave a posterior probability estimate in percentage format from 0 to 100 up to two decimal places.

Study 1 stimuli. In study 1, participants gave estimates to ten probability estimate tasks using the standard probability format[11] taken from Gigerenzer and Hoffrage's (1995) stimuli and translated into English[12]. We selected ten out of the fifteen tasks from their study that maximized the variance in predictions from the 7 estimation strategies in Table 1 (and thus increased model identifiability). Each problem was presented as a vignette that asked the participant to indicate the posterior probability of an event given base-rate, hit-rate, and false-alarm rate information. A list of the cue profiles and verbatim texts used in each of the ten questions is shown in Appendix A.

Study 2 stimuli. In study 2, participants were asked to estimate the probability of an event in a "boxes and balls' paradigm (see Appendix A for verbatim instructions). In

[11] The standard probability format of a Bayesian inference task provides information (base-rates, hit-rates, and false-alarm rates) in single event probabilities (Gigerenzer & Hoffrage, 1995). We chose to use this format because it is the one where people are known to have the most difficulty deriving Bayesian estimates.

[12] We thank Ulrich Hoffrage for providing us with the raw questionnaires from their study.

each question, participants were asked to imagine that a "game master" had a "choosing hat" containing "Choose-A" and "Choose-B" tickets, and two boxes labeled "A" and "B" containing different distributions of Green and Red balls. Participants were told to imagine that the game master selected a random ticket from the choosing hat and drew a randomly selected ball from the box named on the ticket. They then answered the question: "Given that the game master drew a Green ball from the selected box, what is the probability that the ball came from box A?" At the beginning of each question, they were given three pieces of information that corresponded to the base-rate, hit-rate and false-alarm rate of the target question: the proportion of "Choose-A" tickets in the choosing hat (base rate), the proportion of Green balls in box A (hit rate), and the proportion of Green balls in box B (false-alarm rate). A screenshot of the game is presented as Figure A1 in Appendix A. We generated 2 sets of cue-profiles corresponding to Valid Cue and Rare Event + Valid Cue conditions[13]. The final stimuli profiles we selected and used for each condition are presented in Table A1 in Appendix A.

Phase 2 Conditions. After giving their first set of estimates to each problem, participants were told that they would be giving a second set of estimates to each of the questions. They did not previously know that they would be making second estimates. They were randomly (and independently) assigned to one of three conditions in study 1, and one of two conditions in study 2. In the *control* condition (study 1 N = 108, study 2 N

[13] To generate the stimuli profiles for each condition, we simulated 10,000 sets of 15 stimuli profiles where each cue value was drawn from a uniform distribution from 0 to 1. After removing stimuli profiles that did not satisfy the cue-profile constraints of the respective environment, we selected the set of 15 stimuli profiles that maximized the standard deviation of the strategy estimates in Table 3. By maximizing the standard deviation of the strategy estimates, we increase the identifiably of strategies.

=109), participants were told that the researchers were interested in the natural variability in their estimates, and would provide second estimates to the questions (cf. Herzog & Hertwig, 2009, 2014b). They were instructed to answer the questions "as if they were seeing them for the first time." In the *dialectical* condition (study 1 N = 131[14], study 2 N = 116), participants read dialectical instructions. Specifically, they were instructed (Herzog & Hertwig, 2009, p. 234):

> First, assume that your first estimate is off the mark. Second, think about a few reasons why that could be. Which assumptions and considerations could have been wrong? Third, what do these new considerations imply? Was the first estimate too high or too low? Fourth, based on this new perspective, make a second, alternative estimate.

In study 1, we included a third *anagram* condition (N = 111), where participants solved anagrams (unrelated to the estimation task) for four minutes. After the four-minute anagram period completed, participants were given the same instructions as those in the reliability condition.

Following the condition specific manipulations, all participants were also told that their performance bonus for each question would be based on the *better* of their two estimates in phases 1 and 2. This was meant to encourage participants to try different estimation strategies (cf. Herzog & Hertwig, 2009, 2014b). Phase 2 then began and participants were presented with the same problems from phase 1 a second time in a newly randomized order. Participants in the dialectical condition were also reminded of

[14] The imbalance in sample sizes per condition was due to sampling error in our independent random assignment.

their original estimates, in addition to being presented with the problem stimuli, in order to further facilitate changes in estimates.

Phase 3 in study 2. In study 2, we included a third estimate phase (cf. Herzog & Hertwig, 2014b; Müller-Trede, 2011). After giving second estimates in phase 2, participants were told (again, without warning) that they would be making a third, and final, set of estimates for each problem. They were told that they would earn an additional performance based bonus for the accuracy of their phase 3 estimates independent of their estimates in phases 1 and 2 (cf. Herzog & Hertwig, 2014b). However, in contrast to phases 1 and 2, participants were *not* be reminded of the cue statistics for each question, and instead had to make an estimate on the basis of their phase 1 and phase 2 estimates alone. We then presented participants with each of their phase 1 and phase 2 estimates for each problem and asked them to make a new set of estimates.

Results

Definitions. In the following formulas, we use the notation x_{ki} to designate an estimate in phase k for question i, and b_i to designate the Bayesian criterion for question i. For each participant, we calculated the mean absolute deviation (MAD) between their estimates and the Bayesian criterion across problems.

$$MAD_k = \frac{\sum_{i=1}^{N} |x_{ki} - b_i|}{N}$$

We separately calculated MAD values for each participant's phase 1 and phase 2 estimates (labeled MAD_1 and MAD_2 respectively). In addition, we calculated the MAD value of each participant's average estimate between phase 1 and phase 2 across problems (labeled MAD_{avg12}):

$$MAD_{avg12} = \frac{\sum_{i=1}^{N} \left| \frac{(x_{1i} + x_{2i})}{2} - b_i \right|}{N}$$

To calculate averaging gains, we subtracted the MAD of each participant's average estimates from the MAD of their phase 1 estimates:

$$Averaging\ Gain = MAD_1 - MAD_{avg12}$$

Next, we calculated each participant's bracketing rate and accuracy ratio. A participant's *bracketing rate* is defined as the proportion of questions where a person's phase 1 and phase 2 estimates fall on either side of the Bayesian estimate (i.e., have different signed errors; Larrick & Soll, 2006). Finally, a participant's *accuracy ratio* is defined as the ratio of her higher phase MAD value to her lower phase MAD value (Soll & Larrick, 2009):

$$Accuracy\ Ratio = \frac{\max(MAD_1, MAD_2)}{\min(MAD_1, MAD_2)}$$

The minimum accuracy ratio is 1 which occurs when a person's phase 1 and phase 2 estimates have the same average error.

We use Bayesian parameter estimation procedures for all analyses (for information on the strengths of a Bayesian approach to statistics, see, for example, Dienes, 2011; Kruschke, 2010, 2011a, 2011b; Wagenmakers, 2007). To make inferences to group means, we use the BEST package in R (Kruschke, 2013). For regression analyses, we conduct Bayesian mixed-level analyses using the MCMCglmm package in R (Baayen, Davidson & Bates, 2008; Hadfield, 2010). When applicable, we include random intercepts for participants and stimuli. For all analyses, we summarize parameter posterior distributions of interest using 95% highest density intervals (HDIs).

Summary statistics. Summary estimate change and accuracy statistics separated by instruction condition and stimuli environment are presented in Table 4. We only

include data from participants (272 out of 350 in study 1, 206 out of 225 in study 2) who were not classified as using a Bayesian strategy in either phase 1 or phase 2. We highlight four key results here. First, participants in dialectical conditions appeared to change their estimates between phases 1 and 2 more than those in the control conditions (and the anagram condition in study 1). This suggests that dialectical instructions were successful in changing participants' estimation strategies (we will test this using a modeling approach in the next section). Second, estimates were generally more accurate (had smaller absolute errors) in the VC environment compared to the RE+ environment. This is consistent with previous research (and our simulation results) showing that people make poorer estimates relative to Bayes theorem for problems with very low base-rates to those with moderate base-rates (McKenzie, 1994). Third, bracketing rates appear consistently higher in RE+ environments compared to VC environments. This suggests that RE+ stimuli do indeed lead strategies to produce different errors than VC environments, which should in turn lead to larger averaging gains. Finally, bracketing rates tended to be fairly low. The smallest mean bracketing rate was only 6% in the control condition of study 1 for VC stimuli and the largest was 19% in the dialectical condition of study 2 for RE+ stimuli. These rates are substantially lower than has been found in previous inner-crowd research. For example, Herzog and Hertwig (2014b) had a smallest group mean bracketing rate of 14% in their reliability condition, and a largest group mean bracketing rate of 22% in their dialectical condition.

	Control		Anagram		Dialectical	
	VC	RE+	VC	RE+	VC	RE+
Mean Estimate Change	9.33 [0, 31.84]	20.28 [1.74, 48.14]	8.77 [0, 25.85]	20.55 [1.66, 46.16]	15.11 [0, 45.31]	24.91 [1.5, 65.61]
MAD_1	18.98 [0.73, 49.71]	51.37 [8.91, 85.58]	15.87 [0.08, 51.34]	51.1 [11.43, 94.68]	19.04 [0.05, 49.76]	50.48 [0.14, 86.52]
MAD_2	18.47 [0.14, 49.46]	48.08 [3.92, 79.54]	15.11 [0, 50.76]	52.53 [2.06, 83.04]	20.93 [0.09, 48.39]	46.01 [3.92, 81.48]
MAD_{avg12}	18.72 [0.21, 42.91]	49.71 [17.99, 84.91]	15.49 [0.63, 50.51]	51.81 [9.69, 86.78]	19.98 [0.67, 45.4]	48.23 [11.42, 85.03]
$MAD_1 - MAD_{avg12}$	0.26 [-9.17, 12.66]	1.66 [-11.11, 19.28]	0.38 [-12.01, 11.31]	-0.71 [-14.86, 15.98]	-0.94 [-17.44, 9.06]	2.25 [-18.49, 32.31]
Bracketing	0.06 [0, 0.25]	0.09 [0, 0.33]	0.09 [0, 0.5]	0.09 [0, 0.5]	0.11 [0, 0.5]	0.12 [0, 0.5]
Accuracy Ratio	2.64 [1, 9.93]	1.58 [1, 3.12]	3.03 [1, 12.53]	1.68 [1.03, 3.04]	2.59 [1, 9]	2.08 [1, 8.33]

	Control		Dialectical	
	VC	RE+	VC	RE+
Mean Estimate Change	7.44 [0, 14.73]	10.85 [0.19, 33.47]	7.45 [0, 21.6]	14.82 [0.93, 45.73]
MAD_1	21.49 [7.04, 36.04]	21.38 [0.06, 38.87]	31.96 [1.61, 43.66]	35.29 [19.31, 45.59]
MAD_2	20.05 [8.56, 36.3]	22.53 [0.24, 48.59]	32.24 [1.35, 45.2]	34.08 [9.9, 53.62]
MAD_{avg12}	20.16 [7.46, 36.17]	21.14 [0.14, 40.99]	31.3 [1.77, 43.92]	32.84 [15.03, 44.03]
$MAD_1 - MAD_{avg12}$	1.32 [-0.8, 5.87]	0.24 [-13.39, 6.49]	0.66 [-4.01, 7.92]	2.45 [-2.27, 19.87]
Bracketing	0.08 [0, 0.27]	0.13 [0, 0.4]	0.08 [0, 0.33]	0.19 [0, 0.73]
Accuracy Ratio	1.2 [1, 1.89]	2.23 [1, 4.1]	1.26 [1, 1.86]	1.55 [1, 2.59]

Table 4a and 4b: Study 1 (4a) and study 2 (4b) summary statistics. Each cell contains the sample mean across participants in a condition and the corresponding 95% highest density interval (HDI). In calculating summary statistics for accuracy ratios, we excluded data from 15 participants with VC accuracy ratios greater than 15 and 3 participants with RE+ accuracy ratios greater than 15. The corresponding table for all 350 participants, including those who used a Bayesian strategy in at least one phase, is in Appendix B.

Strategy classification. In order to measure how dialectical instructions influenced strategy change, and how strategy change affected averaging gains, we conducted a strategy classification analysis. We assume that each participant uses one estimation strategy across most questions within each estimation phase. However, we do allow some application variability due to factors such as calculation errors or spontaneous use of a different strategy. Formally, we define the likelihood of responses for question i from strategy j applied by participant k as a t-distribution with mean equal to the estimate of strategy j, degrees of freedom equal to one, and standard deviation specific to each participant.

$$t(\mu = s_j(BR_i, HR_i, FAR_i), \sigma = \sigma_{kj}, df = 1) \qquad \text{EQ 1.}$$

Where s_j is the output function of strategy j, and BR_i, HR_i, and FAR_i correspond to the base-rate, hit-rate, and false-alarm rate values for question i. We chose to model errors using the t-distribution because its fat tails can accommodate outliers better than the normal distribution. In addition, we truncated the probability density function (PDF) below 0 and above 1 and normalized it to integrate again to 1. For a visual representation or the modeling procedure, we refer to Figure 6 where we compare the likelihoods given by two different strategies A and B

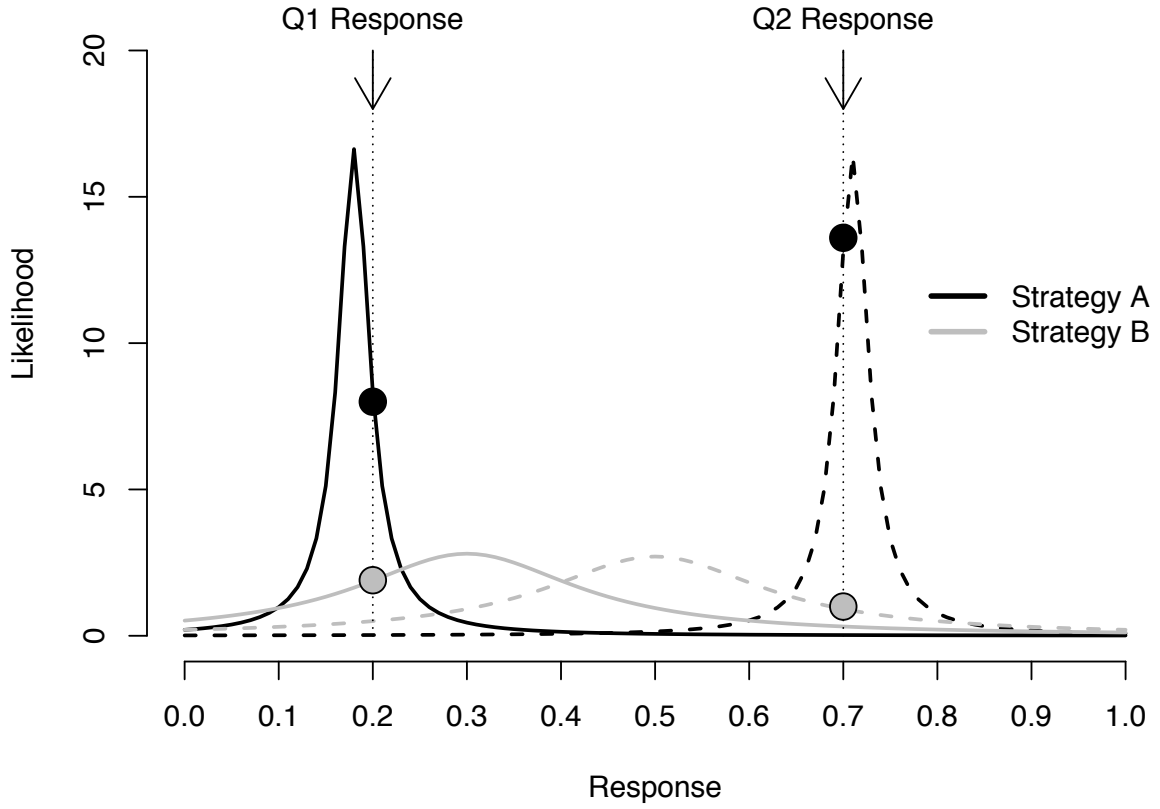


Figure 6. Likelihood functions of two different strategies A and B, plotted in black and gray lines, respectively to two responses of .20 and .70 from a hypothetical participant. The solid lines are the respective likelihood functions for question 1 (Q1) while the dashed lines are the respective likelihood functions for question 2 (Q2). Strategy A depicted predicted values of .18 and .71, while strategy B predicted values of .30 and .50. Because strategy A made predictions close to the participant's responses, its standard deviation can be very low and thus give high likelihood to the responses. In contrast, because strategy B made predictions relatively far from the responses, its standard deviation is forced to be relatively high and it gives low likelihood to the participant's responses. Multiplied across both problems, strategy A gives higher cumulative likelihood to the data than strategy B.

For these two questions, the participant indicated estimates of 0.20 and 0.70, respectively (). Strategy A predicts responses of 0.18 and 0.71, while strategy B predicts responses of 0.30 and 0.50 for questions 1 and 2, respectively. The participant's responses have higher likelihood under strategy A than strategy B.

For each participant and each estimate phase, we compared the maximum-likelihood fits of twelve different strategies in Table 5. In addition to the seven non-averaging strategies from Table 1, we included Bayes theorem, two averaging strategies proposed in prior literature (Gigerenzer & Hoffrage, 1995; McKenzie, 1994), and two base-line models that ignored the cue-profile information. The two averaging strategies were the Relative Likelihood Average (defined as the average of the Base Rate (S5) and Relative Likelihood (S3)), and the Likelihood Average (defined as the average of Likelihood (S1) and Base Rate (S5)). Prior research has suggested that some people spontaneously adopt the strategies (Gigerenzer & Hoffrage, 1995; McKenzie, 1994). For this reason, we elected to include them in our strategy classification analysis. The two base-line models were called "Random" and "Mean." The "Mean" model sets the mean of the t-distribution in EQ 1 to the mean of the participant's responses, while the "Random" model uses a uniform distribution from 0 to 1. A table of all twelve strategies is presented in Table 5.

Strategy Number	Strategy Name	Averaging Strategy?	Cue-Based?	Number of free parameters
0	Bayes	No	Yes	1
1 - 7	See Table 4	No	Yes	1
8	Relative Likelihood	Yes	Yes	1

	Average			
9	Likelihood Average	Yes	Yes	1
10	Mean	No	No	2
11	Random	No	No	0

Table 5. Twelve strategies used in our strategy classification analysis. Averaging strategies average 2 or more cue values. Cue-based strategies make estimates as a function of cue (i.e., base rate, hit rate and/or false-alarm rate) while non-cue-based strategies ignore cue values.

Strategies 0 through 9 have one free parameter, which is the standard deviation of the *t*-distribution; the smaller the standard deviation, the better a strategy captures a participant's estimates across questions (see also Figure 6). Strategy 10 (Mean) has one additional parameter, which is the mean of its *t*-distribution. Strategy 11 (Random) models estimates as a uniform distribution ranging from 0 and 1 (i.e., with no free parameters).

For each participant we calculated maximum-likelihood estimates of the parameters for each strategy for each estimate phase. We then calculated the Bayesian Information Criterion (BIC) for each strategy *m* using the equation:

$$BIC_m = -2\sum_{k=1}^{15} ln\big(lik_m(b_k)\big) + p_m \ln(N)$$

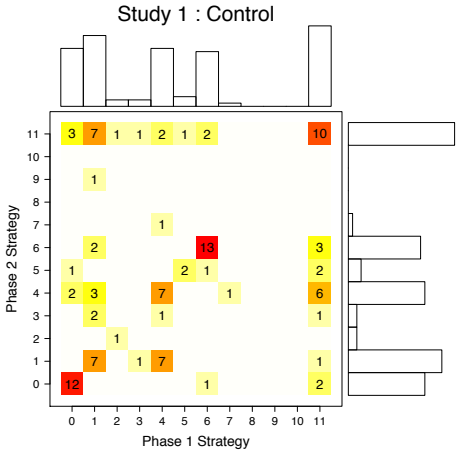
Where *k* is an index for questions, lik_m is the likelihood of the estimate b_k given model *m* using the maximum likelihood estimates for each parameter p_m in model *m*, and N is the number of data points (N = 15 for all models). The BIC measure rewards models that give high maximum-

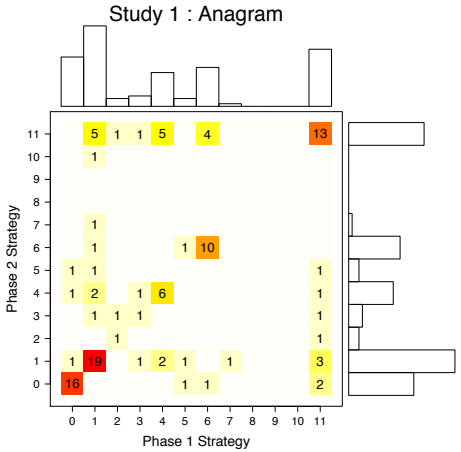
likelihoods to the data while simultaneously punishing models with many free parameters (Lewandowsky & Farrell, 2010). We then calculated ΔBIC values for each model by subtracting the minimum BIC value from each model's BIC value. Finally, we calculated posterior probabilities of each model *m* using the equation:

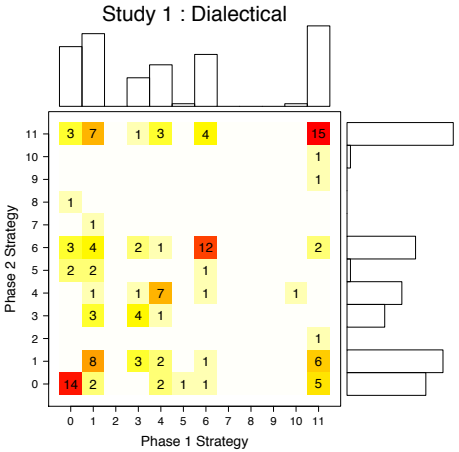
$$Post_m = \frac{e^{-.5*\Delta BIC_m}}{\sum_{i=0}^{6} e^{-.5*\Delta BIC_i}}$$

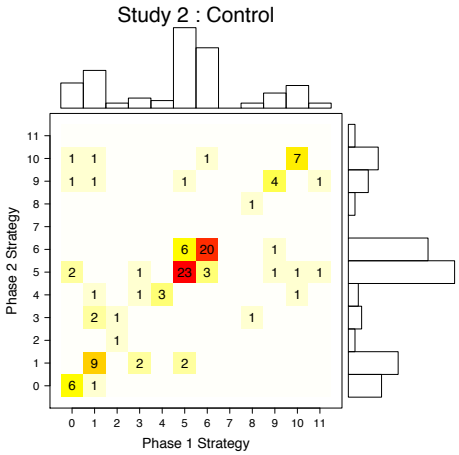
We classified each participant as using the model with the highest posterior model probability. See Appendix C for model recovery simulations supporting the validity of this model classification procedure.

Strategy classification results. Strategy classification results are presented in Figure 7 using heat plots. Vertical axes show results from phase 1 while horizontal axes show results from phase 2. Cells on the diagonal indicate cases where participants were classified as using the same strategy in both phases 1 and 2. Cells in the off diagonal indicate cases where participants were classified as using a different strategy in the two phases. Aggregated classification results are presented in Table 7.









Strategy Index	
0	Bayes
1	Likelihood
2	False Alarm Complement
3	Relative Likelihood
4	Likelihood Subtraction
5	Base Rate
6	Joint Occurrence
7	Hit Rate Minus Base Rate
8	Relative Likelihood Average
9	Likelihood Average
10	Mean
11	Random

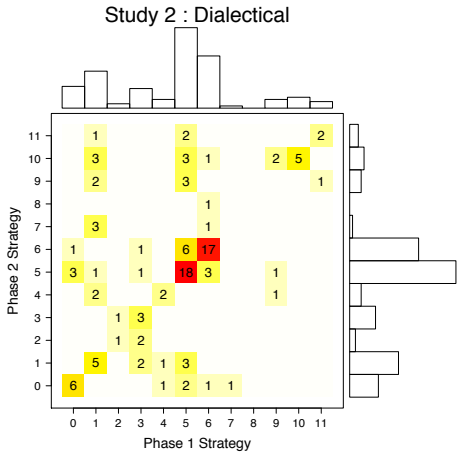


Figure 7. Strategy classification results in phases 1 (horizontal axis) and 2 (vertical axis) separated by study and instructions condition. Values on the main diagonal indicate participants who were classified as using the same strategy in phases 1 and 2. Values off the main diagonal indicate participants who were classified as using different strategies in phases 1 and 2.

Strategy		Study 1			Study 2		
Index	Name	Phase 1	Phase 2	Combined	Phase 1	Phase 2	Combined
0	Bayes	60 (22%)	60 (23%)	120 (23%)	20 (10%)	18 (9%)	38 (9%)
1	Likelihood	80 (30%)	64 (25%)	144 (27%)	32 (16%)	24 (12%)	56 (14%)
2	False alarm complement	5 (2%)	4 (2%)	9 (2%)	4 (2%)	4 (2%)	8 (2%)
3	Relative likelihood	17 (6%)	16 (6%)	33 (6%)	13 (6%)	8 (4%)	21 (5%)
4	Likelihood subtraction	47 (17%)	41 (16%)	88 (17%)	7 (3%)	11 (6%)	18 (4%)
5	Base rate	7 (3%)	15 (6%)	22 (4%)	69 (33%)	59 (30%)	128 (32%)
6	Joint occurrence	52 (19%)	54 (21%)	106 (20%)	48 (23%)	52 (27%)	100 (25%)
7	Hit rate minus base rate	2 (1%)	3 (1%)	5 (1%)	1 (<1%)	4 (2%)	5 (1%)
8	Relative likelihood average	0 (<1%)	1 (<1%)	1 (< 1%)	2 (1%)	2 (1%)	4 (1%)
9	Likelihood average	0 (<1%)	2 (1%)	2 (< 1%)	10 (5%)	14 (7%)	24 (6%)

10	Mean	1	2	3	14	24	38
11	Random	78	89	167	5	5	10

Table 6. Strategy classification results aggregated across conditions separately for each study and phase. Percentages are column percentages for cue-based strategies (strategies 0 through 9) and ignore non cue-based strategies (strategies 10 and 11).

In Study 1, we classified participants to a cue-based strategy in 530 out of 700 (76%) of cases across phases. The remaining 24% of participants were mostly classified as using the Random strategy[15]. Of those participants using cue-based strategies, the most common strategies were (S1) Likelihood (27%), (S0) Bayes (23%), (S6) Joint Occurrence (20%), and (S4) Likelihood Subtraction (17%). These four strategies accounted for 86% of the all cue-based strategies. These classification rates coincide fairly closely with Gigerenzer and Hoffrage (1995), who classified participants using a combination of write-aloud protocols and direct behavioral measures[16].

In study 2, we classified participants to a cue-based strategy in 402 out of 450 (89.3%) of cases across phases. Of those participants using cue-based strategies, the most common strategies were (S5) Base rate (32%), (S6) Joint occurrence (25%), (S1) Likelihood (14%), and (S0) Bayes (9%). These four strategies accounted for 80% of all cue-based strategies. One major difference in strategy use between study 1 and study 2 was the use of the Base rate (S5) strategy. In study 1, this strategy was only used in 4% of cases while in study 2, it was used in 32% of cases. We conjecture that the increase in use of the BR strategy is due to the direct causal function of the base-rate in the Experiment 2 vignette (see Ajzen, 1977).

Strategy switching. Next, we explored the relationship between experimental condition and strategy change between phases 1 and 2. For each participant (including those who were classified as using Bayes in either phase), we calculated whether s/he was classified as using the same strategy or different strategies in the two phases. We used Bayesian graphical modeling to

[15] Participants who were classified to the "Random" strategy were not necessarily responding randomly. They may have been using an estimation strategy other than those in Table 5. Alternatively, they may have violated our modeling assumptions by, for example, alternating between two strategies.

[16] In Gigerenzer and Hoffrage's (1995) first study using the standard probability format, they found the following classification rates (our rates from phase 1 in Study 1 given in the parentheses): Bayes 22% (23%), Joint: 12% (0%), Likelihood: 32% (27%), Likelihood-Subtraction: 10% (17%), Base Rate: 2% (4%). Additionally, they failed to identify the strategy of 27% (22%) of participants.

compare the probability that participants switched strategies as a function of their experimental condition[17]. We then computed 95% HDI for the difference in proportions between conditions. Strategy switching rates for each condition are presented in Figure 8.

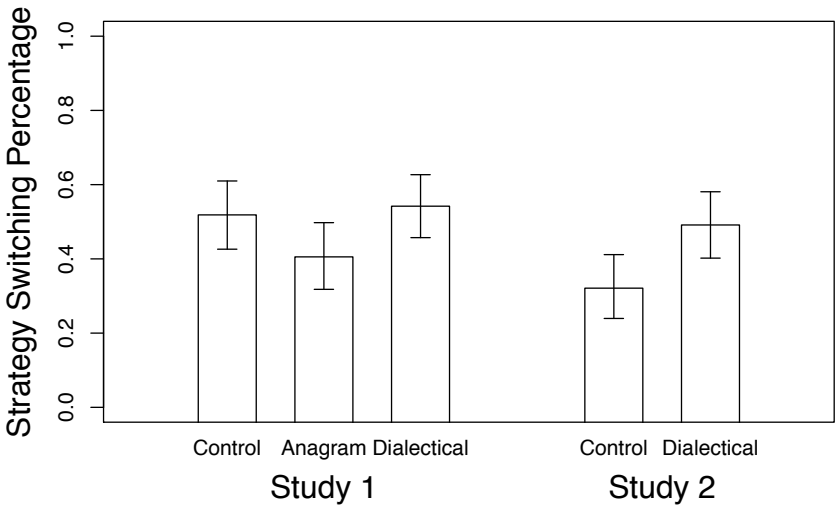


Figure 8: Percentage of participants classified as changing strategies between phases 1 and 2 by instructions condition and study. Error bars represent 95% highest density intervals (HDI).

In study 1, 54% participants in the dialectical condition changed strategies compared to 41% in the anagram condition. The difference between these two conditions was credible (95% HDI of the difference: [.09, .25]). The difference in proportions between the dialectical and the control condition (.023) was positive, but not credibly different from 0 (95% HDI of the difference: [–.10, .15]). In study 2, 49% of participants in the dialectical condition switched strategies compared to 32% in the control condition. This difference was credible (95% HDI of

[17] We used an uninformative beta(1,1) prior distribution for the switching probability in each condition and modeled the likelihood that each participant changed strategies as following a binomial distribution with sample size 1 and probability p corresponding to its experimental group.

difference: [.04, .29]). Thus, with the exception of the non-credible difference between the control and dialectical condition of study 1, we find evidence that dialectical instructions does indeed increase the probability that people change their estimation strategies. To our knowledge, this is the first analysis that uses probabilistic modeling techniques to demonstrate that dialectical instructions induce a qualitative shift in strategy use.

Our simulation results suggested that averaging gains are most pronounced when people switch between strategies with differential base-rate use. Did participants in our studies frequently change their base-rate usage and if so did dialectical instructions increase this tendency? To answer this question, we looked at how often those participants who used non-averaging, cue-based strategies in both phases 1 and 2, switched between strategies using differential base-rate usage in each condition. We excluded participants who were classified as using Bayes in either estimate phase. Results are presented in Table 7.

	Study 1	Study 2
Control	3 / 49 (6%)	3 / 75 (4%)
Dialectical	11 / 55 (20%)	7 / 73 (10%)
Anagram	4 / 51 (8%)	–
Total	18 / 155 (12%)	9 / 148 (6%)

Table 7. Frequencies and proportions of participants who switched between strategies with different base-rate usage. Only participants who used cue-based strategies (S1–S9) in both phases 1 and 2 are included in this table.

Overall, we found low levels of base-rate strategy switching in both study 1 (12% across conditions) and study 2 (6% across conditions). However, we did find credibly higher rate of base-rate switching in the dialectical conditions compared to the control conditions when aggregated over both studies 1 and 2 (95% HDI of difference 2.1%, 16.5%) Thus, we find that in addition to increasing the overall rate of strategy switching, dialectical instructions changed how people on average used base-rate information.

Predictors of averaging gains. Next, we calculated the effect of strategy switching on averaging gains for RE+ and VC stimuli. To do this, we conducted a Bayesian mixed-level regression analyses for each study. In each analysis, we regressed averaging gain on three fixed factors: 1) stimuli type (with RE+ coded as 1 and VC coded as 0), 2) a dummy variable indicating whether or not the participant switched between strategies with different base-rates, and the interaction between the two[18]. Point estimates and 95% HDIs for the beta values in these two regression analyses are presented in Table 8.

	Study 1		Study 2	
	Mean	95% HDI	Mean	95% HDI
Fixed				
RE+ Stimuli	2.75	[–2.72, 8.08]	–0.23	[–1.49, 1.23]
BR Change	6.26	[0.51, 12.48]	2.44	[–0.28, 5.53]
RE+ x BR Change	9.34	[2.67, 15.74]	17.38	[12.42, 22.89]
Random				
Stimuli	16.04	[2.30, 37.35]	0.34	[0.00, 1.43]

[18] We also included random intercepts for participants and stimuli. In order to keep our assumptions as close as possible to our simulation, we only included data from participants who were classified as using a non-Bayesian, non-averaging cue-based strategy (strategies 1 through 7) in both phases 1 and 2.

Participant	37.00	[18.01, 56.28]	7.49	[4.15, 11.26]

Table 8: Averaging gains (measured by the difference in $MAD_1 - MAD_{avg12}$) separately for non-RE+ stimuli (first row) and RE+ stimuli (second row), and for those participants who do not switch strategies (first column) and those who did switch strategies (second column).

In study 1 we found both a credible positive main effect for base-rate cue switching and a credible positive interaction between base-rate strategy switching and stimuli type. In study 2, we replicated the interaction from study 1 but did not quite replicate the main effect for base-rate strategy switching. Together, we find that switching between strategies that differentially use base-rates increases averaging gains over either not switching strategies or switching between strategies with the same base-rate usage. Moreover, these averaging gains are larger for RE+ stimuli compared to VC stimuli[19]. These results are consistent with our simulation-based predictions.

Were participants able to beat their inner crowd? In study 2 we included a third estimate phase where we had participants give their best estimates in light of their estimates in phases 1 and 2. In the next analyses, we focus on how well participants in phase 3 of study 2 were able to beat the average of their inner-crowd. Previous research has found that people are

[19] To see if all strategy switching, irrespective of base-rate use, was related to averaging gains, we repeated the same regression analyses but replaced the BR-Change fixed factor with a variable indicating whether or not people simply changed their strategy between phase 1 and phase 2. In these analyses, the effect of strategy change was not credible (i.e., the 95% HDIs of the strategy change variable included 0) but the interaction between strategy change and stimuli condition were credibly different from 0 for study 1 (95% HDI [0.836, 9.66]) and nearly credibly different form 0 for study 2 (95% HDI [–0.17, 7.01]. Thus, strategy switching (independently of base-rate use) does increase averaging gains to a greater extent for RE+ cue profiles compared to VC cue profiles. However, the magnitude of these interactions was not as large as the interactions between base-rate strategy switching and cue profile presented in Table 11.

largely unable to beat the average of their inner-crowd when they use an alternative strategy such as choosing one of their two estimates (Herzog & Hertwig, 2014b; Müller-Trede, 2011; but see Phillips, Herzog, Kämmer & Hertwig, 2014). However, our analyses of study 2 suggest that the simple average may not be as affective in study 2 relative to past studies (see Table 4). According to the PAR (Soll & Larrick, 2009) model, larger criterion bracketing rates favor averaging strategies while lower criterion bracketing rates favor non-averaging strategies. Previous studies on the inner-crowd found mean bracketing rates ranging from 8% (reliability condition in Herzog & Hertwig, 2009) to 22% (dialectical condition in Herzog & Hertwig, 2014b). In study 2, the median participant produced estimates that bracketed the criterion in only 7% of problems in both the control and dialectical conditions. This suggests that the accuracy of participants' inner-crowd average may not be very high. However, because we know from previous analyses that averaging gains are larger for participants who change their base-rate use, especially for RE+ cue profiles, we expect these participants to be less likely to beat the average of their inner crowd.

To see how often participants were able to beat the average of their inner crowd, we calculated the percent of participants who had MAD_{phase3} values smaller than their MAD_{avg12} values (cf. Herzog & Hertwig, 2014b). These are participants whose phase 3 estimates were, on average, more accurate than the simple average of their phase 1 and phase 2 estimates. Across all conditions and participants, just less than half of all participants (46% or 97 out of 211) were able to beat the simple average of their inner crowd; these results are similar to Herzog and Hertwig (2014b), where the proportions were 47% and 44% for their control and dialectical conditions, respectively. To test the effects of base-rate change and cue-profile environment condition on this effect, we regressed the difference in absolute error between each participant's

phase 3 estimates and the average of their phase 1 and phase 2 estimates on the same three fixed factors and two random factors in our previous regression analysis. We did not find a credible effect of stimuli condition (95% HDI: [–2.57, 0.75]. However, we did find credible negative effects of base-rate cue change (95% HDI: [–8.00, –1.34]) and the interaction term (95% HDI: [–19.67, –8.13]). These negative results suggest that people were *less* able to beat the average of their inner crowd when they changed their base-rate usage between phases 1 and 2. Additionally, this effect was even larger in the RE+ environment.

Discussion

Results from the two empirical studies confirmed three key predictions from the simulation study. First, people can improve the accuracy of their estimates by combining multiple, non-Bayesian strategies. Second, averaging gains are largest when people combine a strategy that does not use base-rate information with another strategy that does use base-rate information. Finally, averaging gains are highest in environments with small base-rates and large hit-rates. Consequently, participants in study 2 were less able to beat their inner crowd when they used strategies with differential base-rate use—especially in the RE+ environment.

There was one major difference between our simulation assumptions and participants' behavior. While our simulation assumed that participants would be equally likely to use each strategy (see Figure 4), our participants clearly preferred some strategies to others. For example, strategy 2 "False Alarm Compliment" was used by about 2% of participants in studies 1 and 2. Moreover, only a small minority of participants (12% in study 1 and 6% in study 2) switched between a strategy that used base rates and a strategy that did not. Because base-rate switching drove much of our expected averaging gains, this meant that participants did not reap as much gains as one would expect based on our simulation.

General Discussion

Psychologists have been comparing human probabilistic inference to Bayes theorem for decades and the result has been a collection of seemingly disparate conclusions in favor of one of two qualitative extremes. While some have argued that people are "conservative Bayesians," (e.g., Edwards, 1968), others claimed that people are "not Bayesian at all" (Kahneman & Tversky, 1972, p. 450). Additionally, some argue that people ignore base-rate information and thus exhibit a categorical base-rate neglect (Lyon & Slovic, 1976), while others argue that people routinely use base rates (e.g., Christensen-Szalanski & Bushyhead, 1981). At the same time, with the notable exception of Gigerenzer and Hoffrage (1995)'s work on natural frequency formats, this debate seems to have ignored a fundamental calling of decision-making research; namely how to help people make better judgments and decisions given their cognitive architecture? To do this, we test how people can use their inner-crowd to improve their Bayesian reasoning judgments. We propose and support the claim that people use a wide variety of estimation strategies that differ in the information they use (i.e., base rates) and the kind of errors they commit. In a simulation and two empirical studies, we find that people can harness this diversity by using their inner crowd (Herzog & Hertwig, 2014a) in order to improve the accuracy of their judgments without any explicit knowledge of Bayes theorem. We find that people can harness the largest gains when they combine strategies with different base-rate use in environments with rare events and a diagnostic cue ("Rare event plus valid cue" environments). Moreover, based on formal strategy classification analyses we find evidence that dialectical bootstrapping, a method of increasing the diversity of the inner crowd (Herzog & Hertwig, 2009), increases both the diversity of strategies used and the probability that people chase their use of the base-rate cue in Bayesian reasoning tasks.

People do not neglect base-rate information

Our simulation and empirical results suggest that it is crucial for people to use base-rate information, especially in rare-event environments with very small base rates. Early research in the heuristic and biases movement of the 1970s and 1980s concluded that people ignore base-rate information (i.e., exhibit "base-rate neglect") and thus are "not Bayesian at all." (Kahneman & Tversky, 1972, p. 450) Our model based classification procedure does not support this conclusion. A substantial number of our participants were classified as using strategies that are sensitive to base rates. In study 1, almost half of our participants (46%) used a strategy that used base-rates (strategies 0, 5 and 6), with almost half of those using a Bayesian strategy (23%). Moreover, in study 2, the percentage of participants who used a strategy using base rates increased to a full 66% of participants. We are certainly not the first to demonstrate that a substantial number of people use base rates. For example, Gigerenzer and Hoffrage (1995) found that 36%[20] of participants used base-rates, while Stanovich and West (1998) found that 42%[21] of participants view base rates as necessary for estimating posterior probabilities. Thus, it is clear that there is substantial variability between people in the strategies they use in Bayesian reasoning tasks. However, as far as we know, we are the first to show that people can generate strategy diversity within one mind, and harness that diversity to improve their judgments.

Spurring and Modeling Strategy Variability

Previous research on the inner crowd (Herzog & Hertwig, 2014a) has found that interventions such as time delay (Vul & Pashler, 2008) and dialectical instructions (Herzog & Hertwig, 2009, 2014b) can decrease error correlations between estimates from the same mind and subsequently improve averaging gains. But where do these decreases in error come from and

[20] We calculated this percentage from the standard probability format column in their Table 3 on page 695.
[21] Study 1 of Stanovich and West (1998).

when can we predict when they are most likely to occur? At a conceptual level, researchers predicted a decrease in errors when people generate second estimates "using knowledge that is at least partially different from the knowledge they used to generate the second estimate" (Herzog & Hertwig, 2009, p. 233). However, the specific estimation process underlying both initial and dialectical estimates has largely been undefined. As a result, previous inner-crowd research has not addressed *how* people change their estimation strategies from initial to second estimates, and how different methods of strategy change affect averaging gains.

In this paper, we took an initial step in answering these questions by modeling the specific estimation strategies people used in both estimation phases. We found that dialectical instructions increase the probability that people adopt new strategies. It would be valuable to use a similar modeling procedure to test how people change their strategies in other estimation domains, such as general knowledge estimation tasks. In the Bayesian reasoning paradigm, all relevant information about the question is presented in the problem (for an alternative view, see Birnbaum, 1983; Gigerenzer, 1996) and all participants have access to the same information. However, when answering general knowledge questions, such as "How tall is the Eiffel Tower?" no statistical information is given and individuals must generate estimates based on their own idiosyncratic knowledge and idiosyncratic estimation strategies (Brown, 2002). For example, in exemplar-based models of judgment, people answer general knowledge questions by first selecting cues relevant to the criterion (e.g., major landmark, building in France) and then retrieve exemplars from long-term memory with similar cue values (e.g., one building I know in France is 80m tall, and the other is 20m tall) and use that distribution to form an estimate (Juslin et al., 2007). How do dialectical instructions affect the process of answering such general knowledge questions? Do they cause people to use different cues or use the same cues and use

different exemplars? In another paper, we address these questions in the context of a population-estimation task and model the processes underlying changes in estimates induced by dialectical bootstrapping (Phillips et al., 2014).

Our model-based simulation results (here and in Phillips et al., 2014) allow us to predict which combination of strategies would lead to the largest averaging gains in which environment. We believe that our method of combining simulation (or analytical) predictions with formal strategy classification methods improves our ability to both predict when averaging gains will occur and to test the empirical accuracy of those predictions with participants.

Judging From Experience: The Normative Adequacy of Bayes Theorem When Probabilities Need to Be Learned

In this paper, we assumed that Bayes theorem is the appropriate normative solution to diagnostic reasoning tasks when the relevant probabilities are given (i.e., base rate, hit rate, and false-alarm rate). However, in everyday life, people often have to learn relevant probabilities by experience because those probabilities are not already conveniently summarized (Hertwig & Erev, 2009). When base rates, hit rates, and false-alarm rates have to be learned based on relatively small samples of experience, a naïve estimate of the Bayesian posterior probability—taking the observed rates at face—is no longer the gold standard and simpler, "non-Bayesian" strategies can outperform it (Juslin, Nilsson, & Winman, 2009). The basic explanation is that multiplication (as used in the Bayes theorem) can exacerbate random error in its noisy inputs, while linear weighting (e.g., such as taking a simple average of the base rate (S5) and hit rate (S1); likelihood average, S9) benefit from the cancellation of opposing random errors and can thus make better out-of-sample predictions than a naïve implementation of Bayes theorem that does not account for sampling error. Because averaging can "tame" random error better than

multiplication can, the accuracy benefits of averaging intuitive non-Bayesian strategies in the inner crowd should be even larger (than we have already observed in our simulations) when base rates, hit rates and false-alarm rates have to be learned from experience. People who use the same strategy in two separate estimation phases could potentially reap inner-crowd averaging gains through the cancelation of random error if their estimates of the relevant probabilities are based on different samples from the underlying populations. Furthermore, averaging the same or different non-Bayesian strategies could potentially even outperform a naïve Bayesian estimator.

Dialectical Bootstrapping And Natural Frequencies: Two Different Ways to Boost Bayesian inferences?

The inner-crowd in general, and dialectical bootstrapping in particular, is one of many methods of bringing people's judgments closer to the Bayesian criterion. The purpose of this research was not to supplant other methods of improving Bayesian inference with dialectical bootstrapping. Indeed, the benefits our participants gained from their inner crowd were relatively small and do not surpass the gains from other methods, such as *natural frequencies* (Gigerenzer and Hoffrage, 1995). Information in natural frequency formats is presented as summary frequencies of events instead of marginal probabilities. When people are given cue profiles transformed into natural frequencies, up to 50% of participants adopt a strategy that makes estimates identical to Bayes theorem (Gigerenzer & Hoffrage, 1995). However, we argue that the natural frequencies and dialectical bootstrapping represent two separate solutions to the same problem. Gigerenzer and Hoffrage argued that probabilities are a relatively recent invention that the human brain did not evolve to process them. Thus, using Hogarth's (2005) terminology, environments where information is only conveyed in a probability format could be seen as

"wicked," where the environment does not provide transparent information to decision makers.[22] Gigerenzer and Hoffrage's solution to the problem was to change the environment by using natural frequencies instead of probabilities as the representational format. In doing this, they changed a wicked information environment into a "kind" information environment that "invite[s] Bayesian algorithms" (Gigerenzer & Hoffrage, 1995, p. 695). In this paper, we asked, given that people find themselves in a 'wicked' information environment (i.e., probability information), what strategies could they use to improve their judgments? We proposed and tested the extent to which people could bootstrap themselves out of a wicked probability environment using their inner crowd. These two approaches represent two different, but mutually complimentary ways to improve human judgment.

Conclusion

In the 1960s, Edwards claimed that people are conservative Bayesians who rely too much on base-rates. In the 1970s, Kahneman and Tversky (1972) argued that people ignore base rates and are thus "not Bayesian at all" (p. 450). We argue that both views are too extreme. People use a variety of strategies in Bayesian reasoning tasks. At times they show base-rate neglect while at other times they show base-rate sensitivity. People can harness the diversity of their inner crowd of non-Bayesian strategies using dialectical bootstrapping to become more Bayesian without any explicit knowledge of Bayes theorem. Averaging two wrongs make it (almost) right.

[22] Hogarth's (2005) original definition of "wicked" versus "kind" environments focused on learning tasks, where kind environments provide immediate and transparent information to the decision maker and wicked environments provide information that is delayed or misleading. While the vignette-based Bayesian reasoning tasks we use are not learning tasks, we find the transparency aspect of the wicked-kind distinction to be relevant.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35*, 303-314. doi:10.1037/0022-3514.35.5.303

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). Norwell, MA: Kluwer Academic Publishers.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. doi:10.1016/j.jml.2007.12.005.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science*, 23, 1117-1122. doi:10.1177/0956797612446024

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211-233. doi:10.1016/0001-6918(80)90046-3

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences, 30*(3), 241-254. doi:10.1017/S0140525X07001653

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *The American Journal of Psychology*, *96*(1), 85-94. doi:10.2307/1422211

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. In B. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 41, pp. 321–359). San Diego, CA: Academic Press.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, *50*(3), 255-272. doi:10.1037/h0060889

Christensen-Szalanski, J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7*(4), 928-935. doi:10.1037//0096-1523.7.4.928

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision, 1*, 79–101. doi:10.1037/dec0000004

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290. doi:10.1177/1745691611406920

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment*. New York: Wiley, 17-52.

Fraundorf, S. H., & Benjamin, A. S. (2014). Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language, 71*(1), 17-38. doi:10.1016/j.jml.2013.10.002

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684-704. doi:10.1037/0033-295X.102.4.684

Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky. *Psychological Review, 103*(3), 592-596. doi:10.1037/0033-295X.103.3.592

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587-606. doi:xxx

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software, 33*, 1–22. Retrieved from http://www.jstatsoft.org/v33/i02/paper

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13*, 517–523. doi:10.1016/j.tics.2009.09.004

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*, 231–237. doi:10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2014a). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*. doi:xxx

Herzog, S. M., & Hertwig, R. (2014b). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 218–232. doi:10.1037/a0034054

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance, 21*, 40–46. doi:10.1016/0030-5073(78)90037-5

Hogarth, R. M. (2005). Deciding analytically or trusting your intuition? The advantages and disadvantages of analytic and intuitive thought. In T. Betsch & S. Haberstroh (Eds.), *The routines of decision making* (67-82). East Sussex: Psychology Press.

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*(4), 856-874. doi:10.1037/a0016979

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114*(3), 678-703. doi: 10.1037/0033-295X.114.3.678

Kahneman D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454. doi:10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237-251. doi:10.1037/h0034747

Kahneman, D., & Tversky, A. (1980). Subjective probability: A judgment of representativeness. In D. Kahneman, P. Slovic, & A. Tversky, *Judgment under uncertainty: Heuristics and biases*. (1st ed., pp 32-47). New York, NY: Cambridge University Press.

Kampstra, P. (2008). Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software, 28*, CS1. doi:xxx

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences,19*(1), 1-17. doi:xxx

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14,* 293–300. doi:10.1016/j.tics.2010.05.001

Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312. doi:10.1177/1745691611406925

Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press/Elsevier.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142,* 573–603. doi:10.1037/a0029146

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers in social psychology: Social judgment and decision making* (pp. 227–242). New York, NY: Psychology Press.

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. California: Sage.

Lindskog, M., Winman, A., & Juslin, P. (2013). Naïve point estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3), 782-800. doi:10.1037/a0029670

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of Personality and Social Psychology*, 47(6), 1231-1243. doi:10.1037//0022-3514.47.6.1231

Lyon, D., & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica*, *40*(4), 287-298. doi:10.1016/0001-6918(76)90032-9

Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *The Quarterly Journal of Experimental Psychology*, 48(1), 188-207. doi:xxx

McKenzie, C. R. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology*, 26(3), 209-239. doi:10.1006/cogp.1994.1007

Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. *Judgment and Decision Making*, 6(4), 283-294. doi:xxx

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, 19(3), 528-534. doi:10.3758/s13423-012-0249-3

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*(1), 29-46. doi:10.1037/h0024722

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*(3), 346-354. doi:10.1037/h0023653

Phillips, N., Herzog, S., Kämmer, J., & Hertwig, R. (2014). Confidence and Dialectical Bootstrapping Facilitates Choosing in The Inner-Crowd. *Unpublished manuscript*.

Ray, R. (2006). Prediction markets and the financial "wisdom of crowds". *The Journal of Behavioral Finance, 7*(1), 2-4. doi:xxx

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 99-118. doi:10.2307/1884852

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 780-805. doi:10.1037/a0015145

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*(2), 161-188. doi:10.1037/0096-3445.127.2.161

Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (in press). Measuring the crowd within again: A pre-registered replication study. *Frontiers in Psychology*. doi:10.3389/fpsyg.2014.00786

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology, 53*(1), 1-26. doi:10.1016/j.cogpsych.2005.10.003

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Random House LLC.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207-232. doi:10.1016/0010-0285(73)90033-9

von Helversen, B., & Rieskamp, J. (2008). The mapping model: a cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General, 137*(1), 73-96. doi:10.1037/0096-3445.137.1.73

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*(7), 645-647. doi:10.1111/j.1467-9280.2008.02136.x

Appendix A

Stimuli Used in Study 1 (from Gigerenzer & Hoffrage, 1995)

The following stimuli were taken from Gigerenzer and Hoffrage's (1995) standard probability formats.

Question 1: AIDS. A consulting center for AIDS specializes in testing people for the AIDS virus. Before an employee of this consulting center informs somebody about a positive result, he wants to know how great the risk is that a person who tests positive for AIDS really is infected with the AIDS virus. He has the following information to make this judgment. The probability is 0.01%, that a man, who is testing for AIDS, is really infected with the AIDS virus. If a man, who is testing for AIDS, really IS infected, the probability is 100.00% that he will get a positive result. If a man, who is testing for AIDS, really is NOT infected, the probability is 0.10% that he will get a positive result. If a man, who is testing for AIDS, got a positive result, what is the probability that he is really infected? ___%

Question 2: Heroin. A hospital has the problem that a lot of men between 20 and 30 years are frequently admitted while unconscious with symptoms of illegal drug use. Doctors often find in such cases pinholes in the patient`s elbow. Doctors then want to know whether or not such patients are heroin users in order to provide appropriate medication. Consider the following statistics relating heroin use and pinholes. The probability is 0.01% that a man admitted to the hospital in this age range is a heroin user. If a man this age IS a heroin user, the probability is 100.00%, that he will have one or more pinholes in his elbow. If a man this age is NOT a heroin user, the probability is 0.19% that he will have one or more pinholes in his elbow.

If a man has one or more pinholes in this elbow, what is the probability that he is a heroin user? ___%

Question 3: Pregnancy. A lab assistant at a gynecological clinic is interested in how accurate pregnancy tests are. She knows the following statistics related to true pregnancies and the results of a pregnancy test: The probability that a woman who gets a pregnancy test at the clinic is really pregnant is 2.00%. If a woman who takes a pregnancy test IS pregnant, the probability is 95.00% that she will get a positive result. If a woman who takes a pregnancy test, is NOT pregnant, the probability is 0.51% that she will get a positive result. If a woman at the clinic has a positive pregnancy test, what is the probability that she is really pregnant? ___%

Question 4: Pimp. There are stereotypes in the US regarding pimps and Rolex watches. One common belief is that American pimps wear a Rolex watch because if they are on the run, they will have still money in the form of an expensive watch. Consider the following statistics regarding pimps and Rolex watches: The probability is 0.005% that an American man is a pimp. If an American man IS a pimp, the probability is 80% that he is wearing a Rolex. If an American man is NOT a pimp, the probability is 0.05% that he is wearing a Rolex. If an American man is wearing a Rolex, what is the probability that he is a pimp? ___%

Question 5: Mammogram. A reporter of a woman's health magazine would like to write an article about breast cancer. He is doing some research about a test that is normally used to identify breast cancer. Because he knows that the test is not perfect and can make errors, he is interested in what it means if a woman has a positive result in a breast cancer test. Consider the following statistics regarding breast cancer and mammogram tests: The probability that a woman who participants in routine breast cancer screening has breast cancer is 1.00%. If a woman who participates in routine screening HAS breast cancer, the probability is 80.00% that she will get a

positive result in the test. If a woman who participates in routine screening does NOT have breast cancer, the probability is 9.60% that the test will make a mistake and she will get a positive result. If a woman, who is going to a routine examination receives a positive test result, what is the probability that she really has breast cancer? ___%

Question 6: Rubella. During a mother's pregnancy, doctor's are concerned about whether or not she will have an outbreak of the disease rubella, as a rubella outbreak during pregnancy can cause a variety in illnesses for her baby. One doctor wants to what the probability is that a baby will have an illness after birth if its mother has an rubella outbreak during pregnancy. Consider the following statistics: The probability is 0.21%, that a baby has an illness after it has been born. If a baby HAS an illness after it has been born, The probability is 48.00% that its mother had a rubella outbreak during pregnancy. If a baby does NOT have an illness after it has been born, the probability is 0.50% that its mother had a rubella outbreak during pregnancy. What is the probability that the baby has an illness after it has been born if its mother had a rubella outbreak during pregnancy? ___%

Question 7: Drunken. There are many accidents at the intersection of Pine street and Oak street. A group of police officers are trying to reduce the number of accidents at this intersection. As there were a lot of drunk drivers involving in these accidents, they are thinking about introducing a breath test. Before they start using the test, they want to know how important it is to know whether or not drivers are drunk. That is, they want to know the relationship between being drunk and getting in a car crash. Consider the following statistics related to drunkenness and car accidents: The probability is 1.00% that somebody crashes with the car on this road at night. If somebody DOES crash his car on this road at night, the probability is 55.00% that he is drunk. If somebody does NOT crash his car on this road at night, the

probability is 5.00% that he is drunk. If somebody drives drunk on this road at night, what is the probability that he crashes the car? ___%

Question 8: Accident. The department of education receives statistics that give information about the causes of accidents during childrens' trips from home to school. There is evidence that children's neighborhood type (urban or rural) is an important risk factor. The department wants to know what the relationship between neighborhood type and the likelihood of an accident is. Consider the following statistics relating neighborhood type and the rate of accidents: The probability is 3.00%, that a child has an accident on the way to school within one year. If a child HAS an accident on his way to school, the probability is 90.00% that he is living in a city. If a child does NOT have an accident on his way to school, the probability is 40.00% that he is living in a city. If a child lives in a city, what is the probability that he has an accident on his way to school within one year? ___%

Question 9: Cab. In an American city there are two taxi companies. One of them has only green taxis, the other one has only blue taxis. One day a taxi caused an accident and the driver took off. The case eventually comes to a trial. One witness identified the taxi as a blue one. The court is now investigating the witness`s ability to correctly identify a blue taxi by night. To help understand how accurate the witness is at distinguishing blue and green taxis, a court usher and the witness went to the location of the accident. The witness then proceeded to identify the color of random cars that were driving by. Consider the following statistics that the court usher gathered relating witness identification and taxi color: The probability that a blue taxi passes by is 15.00%. If a passing taxi really IS blue, the probability is 80.00% that the witness identifies it as blue. If a passing taxi is green and thus is NOT blue, the probability is 20.00% that

the witness identifies it as blue. If the witness identifies this taxi as blue, what is the probability that the taxi really is blue? ___%

Question 10: Feminist. The editors of an Austrian Radio station are planning a broadcast regarding "Feminism and Profession". Participants of this broadcast are asked to guess the representation of feminists (between 20 and 30 years old) in several professions. To evaluate the answers of the participants, the editors obtain the following data: The probability is 5.00%, that a woman this age is an active feminist. If a woman this age IS an active feminist, the probability is 0.40% that she is a bank employee. If a woman this age is NOT an active feminist, the probability is 2.00% that she is a bank employee. If a woman in this age range is bank employee, what is the probability that she is an active feminist? ___%

	Base Rate	Hit Rate	False Alarm Rate	Rare Event + Valid Cue Stimuli
1	0.01%	100%	0.10%	Y
2	0.01%	100%	0.19%	Y
3	2.0%	95.00%	0.51%	Y
4	0.005%	80.00%	0.05%	Y
5	1.00%	80.00%	9.60%	Y
6	0.21%	48.00%	0.50%	N
7	1.00%	55.00%	5.00%	N
8	3.00%	90.00%	40.00%	N
9	15.00%	80.00%	20.00%	Y

10	5.00%	0.40%	2.00%	N

Table A1: Cue patterns contained in the ten vignettes used in study 1. The last column states whether or not the question satisfies the criteria for being in the Rare Event + Diagnostic Cue environment (BR <= .2, HR >= .8, FAR <= .2)

Study 2 Instructions

In this HIT, you will be playing several games that have the same structure. I will tell you about 2 boxes: Box A and Box B. These boxes are filled with balls. In each game, I will choose a box, show you a ball from that box, then ask you to tell me how likely it is that I chose each box. Here is how the games will look:

I will begin each game by filling each box with 100 balls, some of which are White and some of which are Black. The two boxes will have different mixtures of Black and White balls. I will tell you the mixtures in both boxes A and B. For example, I might tell you that Box A has 80 Black balls and 20 White balls, and Box B has 10 Black balls and 90 White Balls.

After I fill the two boxes with 100 balls, I will choose one of the boxes (without telling you which one I chose!) and take a random ball out of that box.

Here is how I will choose a box. I have a hat with 100 tickets. Some of these tickets in the hat say “Choose Box A” and some say “Choose Box B”. I will draw a random ticket out of the hat. I will read the ticket, then choose the box that the ticket tells me. Then I will take a random ball out of that box and tell you what color the ball was.

The number of tickets that say “Choose Box A” and “Choose Box B” will be different in different games. For example, in one game the hat might have 30 tickets that say “Choose Box

A" and 70 that say "Choose Box B." In this game, there would be a 30% chance that I would choose a ticket that says "Choose Box A" and subsequently show you a ball from box A.

Again, after I tell you the color of the ball, your job is to estimate the probability that the ball came from box A

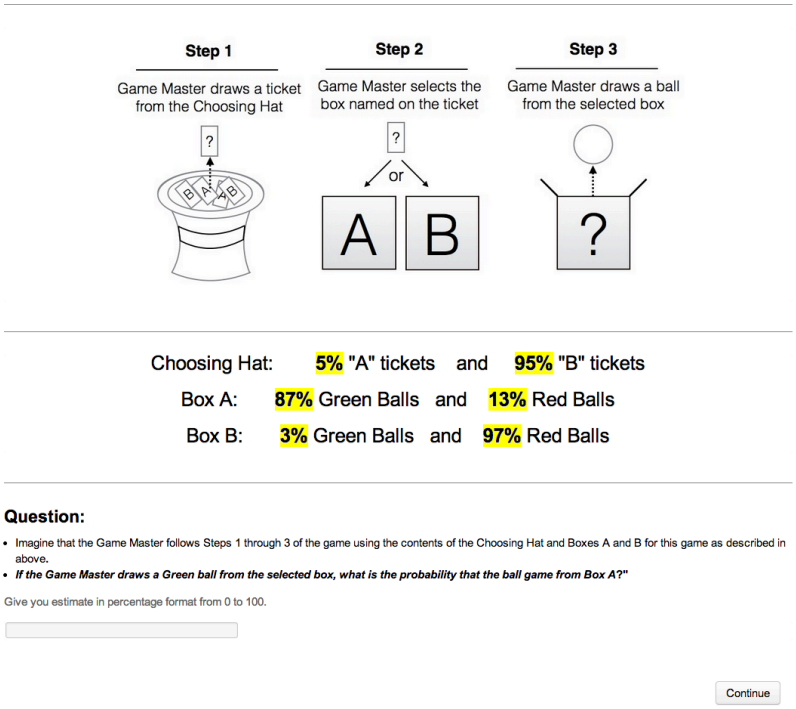


Figure A1: Screen shot of an example stimuli in study 2.

	Valid Cue			Rare Event + Valid Cue		
	BR	HR	FAR	BR	HR	FAR
1	0.87	0.71	0.24	0.05	0.87	0.03
2	0.44	0.72	0.17	0.2	0.89	0.03
3	0.02	0.74	0.23	0.17	0.99	0.07

4	0.19	0.79	0.14	0.2	0.83	0.05
5	0.12	0.88	0.18	0.19	0.97	0.09
6	0.7	0.62	0.14	0.01	0.95	0.1
7	0.33	0.82	0.47	0.16	0.92	0.02
8	1	0.57	0.19	0.15	0.87	0.16
9	0.99	0.54	0.24	0.05	0.8	0.18
10	0.03	0.93	0.18	0.18	0.86	0.15
11	0.94	0.88	0.03	0.04	0.89	0.06
12	0.06	0.85	0.29	0.07	0.97	0.08
13	0.96	0.97	0.14	0.13	0.86	0.18
14	0.95	0.72	0.15	0.15	0.84	0.17
15	0.32	0.51	0.39	0.02	0.98	0.17

Table A2. Cue profiles used in study 2.

Appendix B

	Control		Anagram		Dialectical	
	VC	RE+	VC	RE+	VC	RE+
Mean Estimate Change	9.55 [0, 29.19]	19.65 [1.74, 48.14]	8.78 [0, 26.71]	18.93 [1.66, 46.16]	16.6 [0, 47.75]	26.4 [3.32, 74.33]
MAD_1	22.6 [0.11, 49.71]	59.28 [17.97, 89.23]	20.47 [0.08, 50.26]	61.91 [15.3, 94.68]	20.63 [0.05, 49.91]	58.99 [11.72, 93.93]
MAD_2	21.65 [0.14, 52.32]	54.58 [9.73, 87.4]	20.45 [0.79, 58.16]	60.34 [19.93, 90.54]	22.04 [0.09, 51.91]	51.17 [5.87, 84.49]
MAD_{avg12}	22.12 [0.21, 49.88]	56.92 [19.03, 84.87]	20.46 [0.63, 52.78]	61.12 [19.56, 92.92]	21.33 [0.67, 48.1]	55.07 [17.35, 90.57]
$MAD_1 - MAD_{avg12}$	0.48 [-9.17, 13.2]	2.36 [-11.11, 19.28]	0.02 [-12.01, 12.4]	0.8 [-14.86, 15.98]	-0.7 [-15.56, 15.56]	3.92 [-18.49, 32.31]
Bracketing	0.05 [0, 0.25]	0.08 [0, 0.33]	0.07 [0, 0.25]	0.07 [0, 0.33]	0.1 [0, 0.5]	0.11 [0, 0.33]
Accuracy Ratio	2.78 [1, 11.11]	1.48 [1, 2.36]	3904.17 [1, 20.27]	1.72 [1.01, 3.04]	8.25 [1, 22.9]	6.73 [1, 8.42]

Table B1a: Study 1 summary statistics (with Bayesians)

	Control		Dialectical	
	VC	RE+	VC	RE+
Mean Estimate Change	7.93 [0, 24.8]	8.63 [0, 31.2]	12.9 [0.19, 38.87]	16.6 [0.93, 67.67]
MAD_1	23.44 [7.04, 36.8]	32.77 [1.61, 43.66]	21.7 [0.83, 38.87]	35.53 [18.19, 43.74]
MAD_2	22.14 [8.56, 36.3]	32.76 [1.35, 45.2]	23.94 [4.67, 48.59]	34.11 [16.96, 53.62]
MAD_{avg12}	22.18 [7.46, 36.17]	31.63 [1.77, 43.92]	21.69 [1.08, 40.99]	32.46 [18.15, 44.03]
$MAD_1 - MAD_{avg12}$	1.26 [-3.44, 8.16]	1.14 [-4.78, 7.92]	0.01 [-13.39, 4.87]	3.07 [-2.27, 19.87]
Bracketing	0.08 [0, 0.4]	0.1 [0, 0.33]	0.14 [0, 0.4]	0.21 [0, 0.8]
Accuracy Ratio	1.22 [1, 2.17]	1.24 [1, 1.86]	2.17 [1, 3.25]	1.47 [1, 2.13]

Table B1b: Study 2 summary statistics (with Bayesians)

Appendix C

Formal Strategy Classification Method: Model Recovery Analysis

We conducted a model recovery analysis in order to test the efficacy of our model classification procedure. In the analysis, we simulated the estimates of agents who used the twelve estimation strategies in Table 8 across 10 randomly generated cue profiles in the VC stimuli environment.

For the ten cue-based strategies (including both averaging strategies), we added two sources of error to each agent's estimates: random noise added to each estimate, and a number of contamination responses (i.e., where the estimates are not based on the strategy proper, but represent an "erratic" random response). To model random noise, we added normally distributed error to each agent's estimate with mean of 0 and a standard deviation σ_i (where σ_i represents the agent's level of noise; randomly drawn from the set of values {0, 0.05, 0.10, 0.15, 0.20}). To simulate contamination responses, we replaced c_i of the agent's 10 responses (where c_i is drawn randomly from the set {0, 1, 2, 3, 4}) with a random draw from a uniform distribution ranging from 0 to 1.

For agents using strategy 10 (Mean) we generated t-distributed responses with mean μ_i (where μ_i was drawn from a uniform distribution with bounds at 0 and 1), df = 1, and standard deviation equal to σ_i (randomly drawn from the same set as depicted above). For agents using strategy 11 (Random) we generated responses from a uniform distribution with bounds at 0 and 1.

After generating phase 1 and phase 2 responses from all 5,000 agents, we classified their estimation strategy using the classification procedure outlined in the Results section.

Results. To see how well we were able to recover each agent's true generating strategy, we looked at how often the classified strategy was the same as the generating strategy. Across all agents, the mean recovery rate was fairly high at 86%. As expected, we also found that the recovery rate decreased as the level of error increased (for both random error and contamination responses). Mean model recovery rates by error levels are presented in Table C1:

		Contamination Responses				
		0	1	2	3	4
Noise standard deviation	0	99%	97%	98%	98%	97%
	.05	95%	94%	93%	96%	97%
	.1	91%	87%	89%	87%	82%
	.15	81%	83%	76%	78%	74%
	.20	77%	80%	71%	74%	62%

Table C1. Rates of correct recovering the true underlying model as a function of contamination responses and the standard deviation of random noise. Chance level is 9% (1 : 11 strategies).

Table C2 shows that our classification procedure was able to produce high model recovery rates across a wide range of error values. The procedure seemed particularly robust against contamination responses.

Next, we looked at how often we were able to recover strategy-switching behavior. For each agent, we classified whether it *truly* used two different strategies in phases 1 and 2, and whether it was classified as using two different strategies. We call this variable "strategy switch

recovery." Across agents, the strategy switch recovery rate was very high at 96%. As before, we also found that this measure varied as a function of error rates. In Table C2, we present mean strategy switch recovery percentages as a function of error levels.

		Contamination Responses				
		0	1	2	3	4
Noise standard deviation	0	99%	98%	98%	100%	98%
	.05	98%	100%	98%	98%	99%
	.1	97%	97%	96%	97%	95%
	.15	97%	95%	93%	92%	94%
	.20	95%	94%	96%	95%	90%

Table C2. Rates of correctly recovering strategy switching as a function of contamination responses and the standard deviation of random noise. Chance level is 9% (1 : 11 strategies).

The results in table C2 show that strategy switch recovery rates were quite high and robust to a wide range of error values. Thus, this provides evidence that our modeling procedure is useful in detecting qualitative strategy change in noisy data.

Appendix D

Statistic distribution of McKenzie's (1994) stimuli

McKenzie's (1994) created cue profiles by generating 6.24 million combinations of 2x2-contingency frequency tables where each cell, labeled A through D, contained a frequency from 1 to 50. He defined statistics for each contingency table as follows: base-rate = (A + B) / (A + B + C + D), hit-rate = A / (A + B), false-alarm rate = C / (C + D). We replicated his calculations and obtained the following (non-uniform) marginal distributions of base-rates, hit-rates, and false-alarm rates (Figure D1)

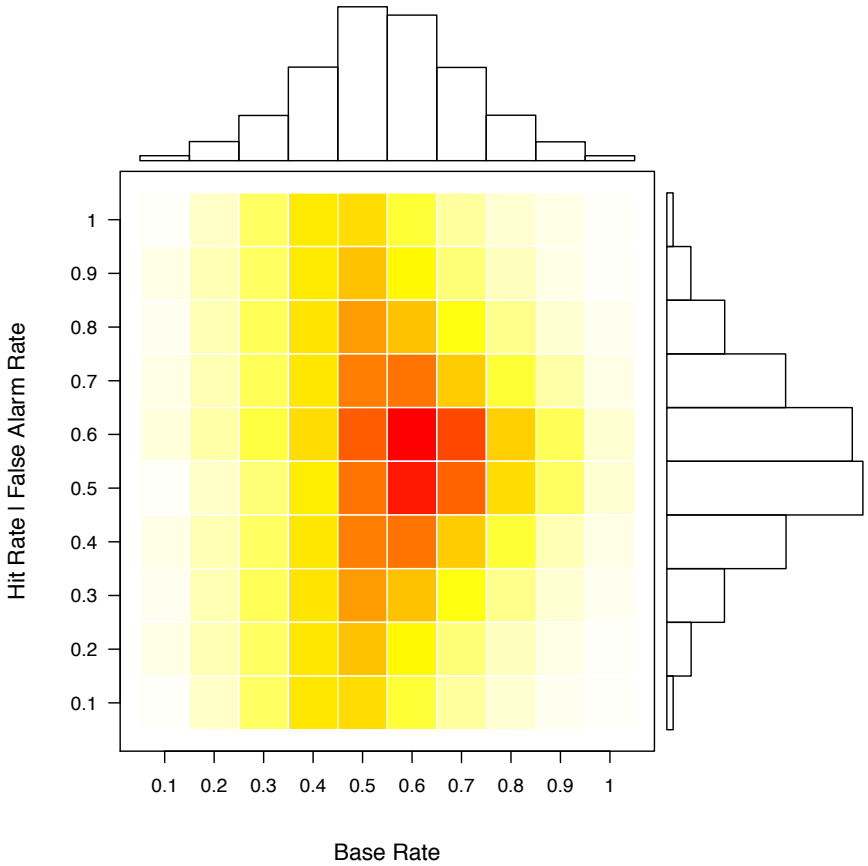


Figure D1. Marginal and joint distributions of base rates and hit rates / false alarm rates that result from McKenzie's (1994) contingency table method. Hit rate and false alarm

rate distributions are identical and are represented here in one plot. Red values indicate higher density.

Confidence and Dialectical Bootstrapping Facilitates Choosing in The Inner-Crowd

Nathaniel D. Phillips, Stefan M. Herzog, Juliane Kämmer, and Ralph Hertwig

Max Planck Institute for Human Development, Berlin, Germany

Author Note

Nathaniel D. Phillips, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development; Stefan M. Herzog, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development; Juliane Kämmer, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development; Ralph Hertwig, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development.

Nathaniel D. Phillips is now at Social Psychology and Decision Sciences, University of Konstanz.

This research was supported in part by grants form the Swiss National Science Foundation 100014_129572/1 to Stefan M. Herzog and Ralph Hertwig.

Correspondence concerning this article should be addressed to Nathaniel Phillips, Department of Social Psychology and Decision Sciences, University of Konstanz, 78457 Konstanz, Germany. Email: nathaniel.phillips@uni.konstanz.de

Abstract

Individuals can improve their judgments by invoking an "inner crowd", that is, by generating multiple, non-redundant estimates and then combining them into a single one. People are often unable to outperform their own inner-crowd when they deviate from averaging their estimates—in part because they cannot reliably identify their more accurate estimates and the statistical structure of estimation environments often favors averaging. In this paper, we explore if people can use their own confidence judgments to identify their more accurate judgments and outperform their average by choosing. We derive predictions for how confidence and accuracy are related in a cue-based estimation task using the Naïve Sampling Model (Juslin et al., 2007). In an empirical study, we confirm key simulation predictions: confidence predicts accuracy in the inner crowd and "high-confidence choosing" outperforms a simple averaging strategy. Furthermore, "dialectical bootstrapping" (i.e., boosting the inner crowd by actively increasing the diversity of estimates) increases the gains reaped by confidence-based estimation.

Keywords: inner crowd, dialectical bootstrapping, wisdom of crowds, judgment aggregation, multiple-cue judgments, confidence,

Confidence and Dialectical Bootstrapping Facilitates Choosing in The Inner-Crowd

Individuals can improve their judgments by invoking an "inner crowd", that is, by generating multiple, non-redundant estimates and then combining them into a single one (Herzog & Herzog, 2014a). For example, given the question "What is the height of the Eifel tower?" a judge can improve the accuracy of a single estimate by generating multiple estimates and averaging them. The driving force underlying inner-crowd benefits is the same underlying the "wisdom of the crowds" phenomena (Davis-Stober, Budescu, Dana, & Broomell, 2014; Larrick, Mannes & Soll 2012; Surowiecki, 2004), where the average of estimates from different people can outperform most, if not all, of its members because non-redundant errors cancel each other out (Larrick & Soll, 2006).

People are, however, reluctant to trust the crowd average and often rather try to "chase the expert" (Soll & Larrick, 2009). Although not using the whole crowd can be a profitable strategy if done wisely (Mannes, Soll & Larrick, in press), chasing a single expert can a risky and often inferior strategy (Davis-Stober et al., 2014; Mannes et al., in press). Similarly, people are often unable to outperform their own inner crowd when they deviate from a simple equal-weight-averaging strategy (Fraundorf & Benjamin, 2014; Herzog & Hertwig, 2014b; Müller-Trede, 2011) because their low skill in identifying their more accurate estimates and the statistical structure of the environment (i.e., somewhat non-redundant errors and not much differences in accuracy to be exploited) favors averaging (Fraundorf & Benjamin, 2014; Herzog & Hertwig, 2014b; Soll & Larrick, 2009).

However, averaging is not always the best aggregation strategy. Many optimal aggregation strategies depart from simple averaging by weighting estimates as a function of *confidence*. When confidence is positively correlated with accuracy, giving more weight to

high-confidence estimates (Yaniv, 2004a) or even choosing high-confidence estimates and ignoring low-confidence estimates (Koriat, 2012b) can outperform averaging. Does this benefit of confidence-based aggregation apply to the inner-crowd?

In this paper we test the benefits of confidence-based aggregation in the inner-crowd by simultaneously modeling the cognitive processes underlying estimate and confidence generation. We use both simulation and experimental data to investigate how confidence is related to accuracy in the inner-crowd and test how people can outperform the simple average of their inner crowd by choosing their more confident estimate (cf. Koriat, 2012b). Additionally, we explore whether "dialectical bootstrapping," a method of boosting the averaging gains of the inner crowd by actively increasing the diversity of estimates (Herzog & Hertwig, 2009) can increase the gains reaped by confidence-based estimation.

Aggregating Different Opinions of the Inner Crowd: Combining, Choosing and Confidence

Just as a group can improve its accuracy by combining multiple estimates (Davis-Stober et al., 2014; Larrick, Mannes & Soll, 2012; Surowiecki, 2004; Yaniv, 2004b; Ariely et al., 2000), an individual can improve her accuracy by generating multiple estimates and combining them into a single, average "inner-crowd" estimate. A typical inner-crowd research task follows three separate phases. In phase 1, participants give initial quantitative estimate to estimation questions, such as "What percent of the Earth's surface is covered by water?[1]" In phase 2, participants give second estimates to the same set of problems, possibly after a time-delay (Vul & Pashler, 2008), or *dialectical* processing instructions that encourage judges to use a different estimation strategy (e.g., Herzog & Hertwig, 2009, 2014b). Finally, in phase 3, a

[1] Try thinking of a few different estimates before reading the answer (it's 71%).

participant's average estimate across phases 1 and 2 for each question is computed and its subsequent accuracy (e.g., in terms of mean absolute deviation from the true values across questions) is compared to a reference such as the average accuracy of phase 1 and phase 2 estimates, or the accuracy of phase 1 estimates. If the inner-crowd works, then the participant's average estimate will be, on average, more accurate than these references.

The reason why the average estimate from an inner-crowd (or any crowd) leads to accuracy improvements is error cancellation: when two estimates have opposing errors, their average will have a smaller absolute error than the average error of the original two estimates. For example, consider a problem with a true answer of 100, and two estimates in a crowd (of different people or within one mind) of 50 and 150. These two estimates have signed errors of -50 and +50 respectively. On average their absolute deviation from the true answer is 50. However, the *average* estimate of the inner-crowd of 100 (50 + 150 / 2) is identical to the true answer and has no error. Thus, to the extent that the estimates from a crowd have opposing errors, whether the crowd is derived from multiple individuals or one mind, the average estimate from the crowd will likely outperform most if not all of its individual members. Because error cancelation drives averaging benefits, individuals reap maximum crowd benefits when the correlation of signed estimate errors is low. This can be achieved by interventions such as a time-delay (Vul & Pashler, 2008) or *dialectical* instructions in the context of dialectical-bootstrapping (Herzog & Hertwig, 2009); a method of increasing inner-crowd benefits by having judges generate dialectical estimates that derive from different knowledge or estimate strategies (Herzog & Hertwig, 2009; 2014a; 2014b).

How Do and Should People Aggregate Multiple Estimates Themselves?

Averaging is a robust strategy in a wide variety of contexts (Soll and Larrick, 2009). However, in advice-taking tasks, people often fallaciously believe that the average judgment is no better than the average judge and try to find the best estimate (aka, "chase-the-expert) in a group (Soll & Larrick, 2009). As a result, people rarely outperform the group average. However, there are specific estimation environments where averaging leads to less accuracy than other aggregation strategies such as weighted-averaging or choosing. Normatively, the probability, accuracy, redundancy (PAR) model (Soll and Larrick, 2009) establishes the environments where averaging outperforms choosing. Generally, the model states that averaging is better when two estimate sources (either two separate judges or two separate estimates from the same mind) have similar overall levels of accuracy, have relatively uncorrelated estimate errors, and when it is difficult to know *a priori* which judge is more accurate (Soll and Larrick, 2009).

While it is clear that people could potentially benefit from averaging their inner-crowd, it is less clear *when* people decide to use their average instead of, for example, choosing one of their individual estimates. Papers that have explored this question showed mixed results, with some studies showing that people do not consistently average their inner-crowd (Müller-Trede, 2011) and others finding that averaging is indeed a common strategy (Fraundorf & Benjamin, 2014; Herzog & Hertwig, 2014a). However, all studies found that people can only rarely outperform the average of their inner-crowd when they decide not to average their estimates (Müller-Trede, 2011; Fraundorf & Benjamin, 2014; Herzog & Hertwig, 2014a).

Can The Meta-Cognitive Cue of Confidence Identify The "Expert in Your Head"?

Previous inner-crowd research has ignored a key variable that could improve people's ability to outperform averaging: *confidence*. On the one hand, if high-confidence estimates in the inner-crowd are substantially more accurate than low-confidence estimates, then choosing high confidence estimates could outperform averaging. On the other hand, if confidence is unrelated to accuracy, then confidence-based choice could be another form of 'chasing the expert' that fools people into thinking they can outperform the average. How do, and should, people use confidence in managing their inner-crowd?

In advice-taking tasks, people weigh advice as a function of the confidence with which it is given. When receiving advice from two advisers, people trust high confidence advisers more (Sniezek & Van Swol, 2001; Bonaccio & Dalal, 2006; Price & Stone, 2004) and weight advice from more confident advisers more than less confident advisers (Yaniv, 1997). When combining their own estimates with the advice of others, people weight their own estimates more as their own confidence increases, and less as their adviser's confidence increases (e.g., Moussaïd, Kämmer, Analytis & Neth, 2013; Soll & Larrick, 2009). Thus, people act as if they believe that estimates given with high confidence are more accurate than those given with low confidence.

While people use confidence as a cue for accuracy, the actual statistical relationship between confidence and accuracy is less clear. Researchers typically use two criteria to evaluate confidence: *calibration*, and *resolution* (or discriminability; Yates, 1990; Liberman & Tversky, 1993). Calibration measures the difference between a judge's predicted probability of an event's occurrence, and the empirical (or true) probability. The smaller the difference between predicted and true probabilities, the higher the calibration. Many studies show that

people exhibit over-confidence and thus poor calibration (e.g.; Lichtenstein & Fischoff, 1977; Hall, Ariss & Todorov, 2007; Soll & Klayman, 2004; Klayman et al., 1999; Block & Harper, 1991; Christensen-Szlanski & Bushyhead, 1981; Glaser, Langer, & Weber, 2013; Juslin et al., 2007; Moore, Tenney, & Haran, in press; Soll & Klayman, 2004). However, there is substantial controversy regarding whether or not measured overconfidence truly reflects a persistent cognitive bias or is due to, for example, random error in the estimation process (Erev, Wallsten & Budescu, 1994; Winman, Hansson & Juslin, 2004), or measurement biases (Gigerenzer, Hoffrage & Kleinbölting, 1991; for the discussion, see Griffin & Brenner, 2004; Keren, 1997; Merkle, Van Zandt, 2008a, 2008b; Olsson, Juslin, & Winman, 2008; Moore & Healy, 2008).

Confidence resolution refers to the ability to discriminate between high and low probability events. The better one's confidence judgments can distinguish between high and low probability events, the higher the judge's resolution. Despite the ongoing controversy about the nature, reality and implications of people's overconfidence in their choices, it is clear that people's confidence intervals show above-chance-level resolution. Many studies have found that accuracy is a monotonically increasing function of confidence (e.g., Yates, 1990; Winkler, 1971; Christensen-Szalanski & Bushyhead, 1981; Koriat, 1980). For example while people's confidence intervals are generally too narrow (43% hit rate vs. 90% confidence intervals), the width of a confidence interval reliably predicts the accuracy of the point estimate it contains (Yaniv & Foster, 1997). Psychologically, this suggests that people have some ability to monitor the accuracy of their decisions and convey that monitoring process through a confidence rating. Practically, this means that confidence can be a valid cue to accuracy, which in turns means that it can benefit judges to weight advice as a function of the their associated level of confidence. Indeed, many studies have found that weighting and adding estimates a function of

each judge's confidence improve accuracy (Bang et al., 2014; Bahrami, Olsen, Latham, Roepstorff, Rees & Frith, 2010; Koriat, 2012b; Yaniv, 1997).

One method to use confidence is to weight high confidence estimates more than low confidence estimates, another is to completely ignore low confidence estimates and *choose* high-confidence estimates. Koriat (2012b) found normative support for high-confidence choosing by showing that a dyad of judges can benefit by using a *maximum confidence slating* (MCS) heuristic, which states that a judge should choose the estimate from the high-confidence adviser and completely ignore the estimates from the low-confidence adviser (but see Bang et al, 2014). Thus, choosing can outperform averaging in a crowd if confidence is sufficiently and positively related to accuracy.

This leads us to our critical question: can the inner-crowd parallel to Koriat's (2012b) MCS heuristic, which we call *inner-MCS*, also outperform an averaging strategy[2]? Or is high-confidence choosing within one mind another method of 'chasing-the-expert" that leads to poorer estimation than simple averaging? Additionally, we test the extent to which dialectical bootstrapping, a method of increasing estimate diversity in the inner-crowd (Herzog & Hertwig, 2009; 2014a; 2014b) can increase the benefits or the inner-MCS heuristic. We will answer these questions using a cue-based estimation task. In the task, judges are asked to estimate the population of several US counties on the basis of four binary statistics from that county. In addition to providing best estimates, judges provide 90% confidence intervals. Judges provide their estimates to each county twice (in phases 1 and 2). In a third phase, we present judges with their estimates and, in absence of cue values, ask them to provide their best

[2] Koriat (2012b) did test the effectiveness if an inner-MCS heuristic on perceptual decisions but did not find substantial differences between averaging. However, is unclear if Koriat's (2012b) results will replicate in estimation tasks and whether people actually *use* confidence in harnessing their inner-crowd.

possible estimate of the county population. To make predictions for behavior in the task, we take a cognitive modeling approach to both estimation and accuracy.

A Cognitive Perspective On the Inner Crowd: Modeling the Estimation Process

With one notable exception (Hourihan & Benjamin, 2010), previous research on the inner-crowd has focused on how people should aggregate their estimates and has largely ignored *how* estimates are generated in the first place. This is both a major practical and theoretical research gap: without knowing the processes underlying estimates, one cannot make more precise predictions for when one should average their inner-crowd or use an alternative strategy. In this paper, we focus on sampling models of estimation which assume that people form estimates based on a small sample of problem-relevant information drawn from long-term memory (e.g., Wallsten & Gonzalez-Vallejo, 1994; Juslin, Olsson & Olsson, 2003; Koriat, 2012a). Specifically, we rely on predictions from the Naïve Sampling Model (NSM, Juslin, Winman & Hansson, 2007). The MNSM simultaneously describes how estimates *and* confidence judgments are generated. The model assumes that people estimate unknown criterion values by sampling a small number of examples similar to a target from long-term memory, and use the distribution of criterion values tied to those observations to derive both best estimates and confidence intervals for the criterion. Importantly, the judge's working memory capacity constrains the number of observations that can be used at the time of judgment. Thus, only a finite number of examples from long-term memory will be used (e.g., Miller, 1956; Cowan, 2001). This class of models has recently been used to correctly predict that people with smaller working memory spans will experience larger inner-crowd benefits than those with larger working memory spans (Hourihan & Benjamin, 2010). More generally, their results provide support for models that assume estimates and confidence judgments are

based on noisy information sampled from long-term memory (Hansson, Juslin, & Winman, 2008; Lindskog & Winman, 2014; Lindskog, Winman, & Juslin, 2013a, 2013b; Vul & Pashler, 2008; but see Rauhut & Lorenz, 2011).

Simulation Study: Extending The Naïve Sampling Model to Model The Inner Crowd, Confidence and Dialectical Bootstrapping

The simulation study served two purposes: First, by simultaneously modeling estimation and confidence using the Naïve Sampling Model (Juslin et al., 2007), we can predict how confidence should be related to accuracy in the inner-crowd for the US county stimuli we use in a subsequent behavioral study. Second, the simulation allowed us to predict the potential effects of *dialectical* instructions, those that encourage judges to generate a diverse inner-crowd, on estimate accuracy and confidence (Herzog & Hertwig, 2009). Before we describe the NSM in more detail, we first introduce the task environment used in the simulations and in the empirical study.

The Task Environment: Estimating US County Populations Based On Binary Cues

Participants (called "agents" in the simulation) estimated the populations of 16 US counties, based on four statistical cues taken from the 2010 US census database[3] (see Juslin, et al., 2007, for a similar simulation based on one single cue). Importantly, participants did not know the names of the counties and instead were forced to make estimates based on cue values alone[4] (cf. Peterson & Pitz, 1986). We calculated cue values and populations from the complete database of all 3,007 US counties as of 2010. We chose four statistics to use as population cues:

[3] Retrieved from http://web.archive.org/web/20130921075947/http://quickfacts.census.gov/qfd/download_data.html).

[4] We chose to hide the county names for two reasons. First, we wanted to prevent participants from directly retrieving the true population from memory in order to force them to make estimates under uncertainty. Second, because we collected experimental data online and rewarded participants for their accuracy, we wanted to prevent them from doing an online search for the true population of each county.

poverty % (i.e., the percentage of the county population living at or below the poverty level), *# housing units* (i.e., the number of housing units in the county), *bachelor's %* (i.e., percent of residents with Bachelor's degrees), and *population density* (i.e., the number of inhabitant's per square mile). We conducted a median split on each cue, with values below the median set to 0 and values above the median set to 1, to create binary value across all counties[5]. We define an individual county's *cue profile* as its combination of all four binarized cues. For example, a cue profile of [0, 0, 0, 0] represents a value in the bottom 50% on all four cues. There were 16 unique cue profiles corresponding to all $2^4 = 16$ possible combinations of the four binarized cues. Next, we grouped counties with the same cue profile and calculated the median county population in each of the 16 groups. This median population value for a given cue-profile represented the criterion value for that profile. Table 1 shows all cue profiles including the number of counties that satisfied each cue profile, and the median county population (criterion):

Stimulus	Bachelor's %	Poverty %	Population Density	Housing Units	N	Median (population)
1	Low	Low	Low	Low	275	8,804
2	Low	Low	Low	High	31	33,052
3	Low	Low	High	Low	47	20,624
4	Low	Low	High	High	120	46,904
5	Low	High	Low	Low	539	11,961
6	Low	High	Low	High	72	36,982
7	Low	High	High	Low	131	20,210

[5] Cue values at the median were independently and randomly assigned to 0 or 1.

8	Low	High	High	High	285	45,573
9	High	Low	Low	Low	331	7,187
10	High	Low	Low	High	78	36,492
11	High	Low	High	Low	42	20,211
12	High	Low	High	High	571	130,016
13	High	High	Low	Low	124	7,765
14	High	High	Low	High	51	39,321
15	High	High	High	Low	14	22,460
16	High	High	High	High	296	113,917

Table 1: 16 stimuli used in the study. Low values indicate being in the bottom 50% of the statistic, while High values indicate being in the top 50% of the statistic. N shows the number of counties in the dataset that satisfy each cue profile. Median (population) shows the median population of all counties matching a cue profile.

The Naïve Sampling Model (Juslin et al., 2007)

The Naïve Sampling Model (Juslin et al., 2007) assumes three sequential phases in the estimation and confidence interval process. We depict both the general process and a specific example for the county estimation task in Figure 1.

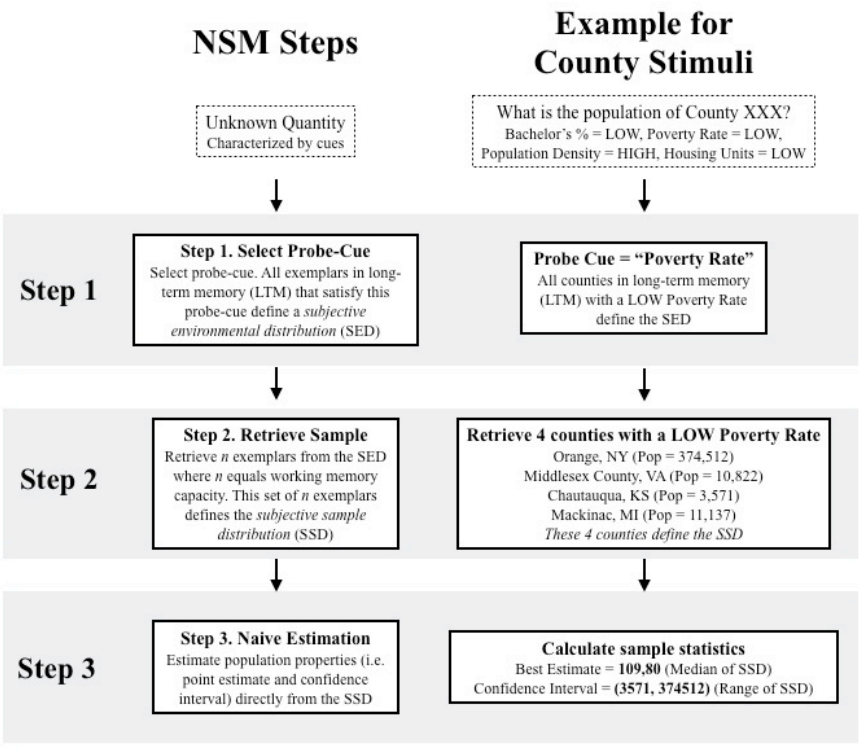


Figure 1: General steps in the NSM and an example estimation procedure for the county population estimation task. For the example, we assumed that the "Poverty Rate" cue was randomly selected in step 1 of the example procedure.

In Step 1 (after being presented with the stimulus) judges select a cue and observe its value. For example, a judge could select the cue "poverty rate" and observe the value "LOW". This cue-value combination becomes the *memory probe*. Next, judges search their long-term memory for exemplars that with cue-value combinations that match the memory. This distribution of exemplars in long-term memory defines a *subjective environmental distribution* (SED). For example, this distribution could represent all counties that a judge knows with a low poverty rate. In Step 2, the judge selects a small sample of observations from the subjective

environmental distribution and brings those samples into working memory. This sample comprises the judge's *subjective sampling distribution* (SSD) for that question. In Step 3, the judge uses the sample distribution of criterion values in their SSD to generate both best estimates and confidence intervals for the criterion. Judges are assumed to use a measure of central tendency such as the sample median to calculate best estimates (Lindskog, Winman & Juslin, 2013), and a measure of variability such as the sample range to calculate a confidence interval (cf. Juslin et al., 2007). For example, a person with an SSD corresponding to the four counties Orange (NY), Middlesex (VA), Cautauqua (KS) and Mackinac (MI), could use the median and range[6] of those four counties as a best estimate and confidence interval for the criterion (see the example procedure in Figure 1 for details).

Modeling dialectical instructions in the NSM. In previous dialectical bootstrapping research, dialectical instructions, usually those inspired the consider-the-opposite technique (Lord, Ross & Lepper, 1984), were designed to spur judges to produce phase 2 estimates using different knowledge from what produced their phase 1 estimates (Herzog & Hertwig, 2009). While previous studies have found that dialectical instructions lead to increased estimate change and decreased signed error correlations between phases 1 and 2 (Herzog & Hertwig, 2009; Herzog & Hertwig, 2014b), no studies have modeled the exact process by which dialectical instructions influence estimate change (but see Phillips, Herzog & Hertwig, in prep). In our simulation, we implemented one potential process of dialectical strategy change relating to cue use in Step 1 of phase 2. When *control* participants (agents) start Step 1 in phase 2, we

[6] The original formulation NSM (Juslin et al., 2007) proposed that people generate X% confidence intervals with a width equal to $w = 2 * sd * z_p$, where sd is the sample standard deviation of the SSD and z_p is the z score delimiting the central proportion *p* of the normal distributions. In our simulations, we assume, for psychological simplicity, that people do not have access to a standard normal table and simply use the range of their SSD as their confidence interval.

assume that they use the *same* probe cue from phase 1. The control procedure was designed to capture random variability in the estimation process. When those in the *dialectical* condition start Step 1 of phase 2, they select a new probe cue different from the one they (randomly) selected in phase 1[7]. This dialectical method was designed to capture how agents, and people, could increase the diversity of their estimates. We expect the increased diversity of cue use in the dialectical condition to lead to larger estimate changes and lower signed error correlations between phases 1 and 2.

Simulation Procedure

We present a general description of the simulation process in this section. Full details are presented in Appendix A. We generated 5,000 agents that produced both best estimates and confidence intervals to the 16 stimuli in Table 1 across two separate phases. Agents differed in several parameters relevant to long-term memory knowledge and working memory capacity[8]. Each agent began by generating phase 1 best estimates and confidence intervals for each of the 16 stimuli in Table 1 by following the NSM procedure outlined in Figure 2. Each agent was then randomly assigned to one of two phase 2 conditions. In the control condition, agents skipped Step 1 of phase 2 and used the same memory probe they used in Step 1 of phase 2. They then drew a new random set of exemplars matching the memory probe (with replacement). In the dialectical condition, agents selected a *new* random cue in Step 1 of phase 2 that differed from the cue they selected in Step 1 of phase 1. They then used this new cue to generate a new probe cue for Step 2 of phase 2.

---

[7] There are certainly alternative valid methods of simulating the effects of dialectical instructions on the estimation procedure and we do not mean to suggest that our method is the only one. Rather, we use it as a starting point to model dialectical estimates.

[8] Each agent was assigned a long-term memory storage consisting of a subset of exemplars from the total county database. In order to capture errors in memory, we assigned each agent a bias and random error term that was added to their memory of county populations. Each agent also had a set working memory capacity that constrained the number of exemplars it could process in step 2 of the NSM procedure. See Appendix A for full details.

Simulation Analyses: Defining Measures of Interest

Each agent provided three values for each question in each phase, totaling six values across both phases: two best estimates (b_1 and b_2), a lower and upper bound for phase 1 (l_1, and u_1) and a lower and upper bound for phase 2 (l_2, u_2). We label the criterion ("Truth") value for a question k as T_k. We use the index *j* for estimate phase (j in {1, 2}) and *k* for questions (k in {1, 2, …. 16}).

Estimation error: Absolute deviation (AD), and Mean Absolute Deviation (MAD). We define the absolute deviation of an estimate for a question as the $\log_{10}$-transformed absolute difference between the best estimate in a phase and the criterion value (i.e.; the median county population) corresponding to the question.

$$AD_{jk} = log_{10}\left|b_{jk} - T_k\right|$$

We further define mean absolute deviation (MAD) values across stimuli separately for phases 1 and 2.

$$MAD_{Phase1} = \frac{\sum_{k=1}^{16} AD_{1k}}{16},\ MAD_{Phase2} = \frac{\sum_{k=1}^{16} AD_{2k}}{16}$$

Confidence (C). We defined confidence in a question as a decreasing function of the width of an agent's confidence interval (i.e., the absolute difference between lower and upper value). Because wider confidence intervals indicate less confidence, we multiplied the absolute difference between the maximum and minimum confidence interval values by -1 so that larger values in the measure indicate more confidence:

$$C_{jk} = -log_{10}\left(u_{jk} - l_{jk}\right)$$

After an agent completed phases 1 and 2, we then defined its *high-confidence best estimate* (b_H) as the best estimate in the phase with higher confidence, and its *low-confidence best estimate* (b_L) as the best estimate in the phase with lower confidence. For example, if an agent produced estimates b_1 = 20,000, l_1 = 10,000, u_1 = 30,000 and b_2 = 50,000, l_1 = 49,000, l_2 = 51,000, then its high-confidence estimate would be 50,000 (the phase 2 best estimate) and its low-confidence estimate 20,000 (the phase 1 best estimate). If an agent gives the same confidence to both estimates, then b_H and b_L are undefined. We then defined each agent's average high-confidence estimate error (MAD_{HC}) and average low-confidence estimate error (MAD_{LC}) values by calculating the mean absolute deviation between the agent's higher and lower confidence best estimates across problems. If an agent always gave its low-confidence estimates in phase 1 (or phase 2), then its MAD_{LC} values would be equal to its MAD_{Phase1} (or MAD_{Phase2}) values.

Changing one's opinion between phase 1 and 2: Estimate change (Δb). Estimate change for each question is defined as the (log_{10} transformed) absolute difference between an agent's best estimate in phase 1 and its best estimate in phase 2.

$$\Delta b_{12} = log_{10}|b_1 - b_2|$$

The next three variables, bracketing, phase accuracy ratio, and confidence accuracy ratio, each relate to Soll and Larrick's (2009) PAR model and help dictate when choosing outperforms averaging.

Error cancelation: Bracketing (Br). Bracketing is a binary value indicating whether or not, for a specific question, the criterion value falls between phase 1 and 2 best estimates (Soll & Larrick, 2009).

$$Br = 1, if\ \{b_1 < T < b_2 | b_2 < T < b_1\}$$

$$Br = 0, if\ \{(b_1 < T\ \&\ b_2 < T)|(T < b_1\ \&\ T < b_2)\}$$

We can subsequently define an agent's *bracketing percentage* (BP) as its average bracketing value across stimuli:

$$BP = \frac{\sum_{k=1}^{16} Br_k}{16}$$

Phase accuracy ratio (A_{phase}). An agent's phase accuracy ratio measures the ratio of mean errors in phases 1 and 2. It is defined as the ratio of the higher to the lower phase-based MAD value (cf. Soll & Larrick, 2009):

$$A_{phase} = \frac{\max\ (MAD_{Phase1}, MAD_{Phase2})}{\min\ (MAD_{Phase1}, MAD_{Phase2})}$$

A large phase accuracy ratio suggests an agent's set of estimates in one phase is is much more accurate than its other phase. However, it does not show whether phase 1 or phase 2 estimates are more accurate.

Confidence accuracy ratio (A_{conf}). In the same way that an agent's phase accuracy ratio measures the relative accuracy of its phase 1 to phase 2 estimates, its *confidence accuracy ratio* measures the relative accuracy of its high confidence estimates to its low confidence estimates. To calculate an agent's confidence accuracy ratio we calculated the ratio of its higher to lower confidence-based MAD values:

$$A_{conf} = \frac{\max\ (MAD_{HC}, MAD_{LC})}{\min\ (MAD_{HC}, MAD_{LC})}$$

High confidence accuracy ratios suggest that an agent's high (or low) confidence estimates are much more accurate than its low (or high) confidence estimates. Just as A_{phase} does not show *which* phase has a lower MAD value, A_{conf} does not show whether an agent's high-confidence or low-confidence estimates are more accurate.

Simulation Results

We analyzed[9] the simulation results with Bayesian mixed-level regression analyses (Baayen, Davidson & Bates, 2008) using the glmmMCMC package in R (Hadfield, 2010). We included random intercepts for agents and stimuli in each analysis except when specified otherwise. We report 95% highest density intervals (HDI) for the posterior distribution of coefficients (Kruschke, 2011).

Estimate change. Agents in the dialectical condition changed their estimates more than control agents; and the more confident agents were in phase 1, the less they changed their estimates between phases 1 and 2. We regressed estimate change (Δb_{12}) on two fixed factors: an indicator variable indicating the agent's phase 2 estimate procedure condition (with the repeated condition coded as 0 and the dialectical condition coded as 1) and phase 1 confidence (C_1). We found credible positive effects for the dialectical condition (95% HDI: 0.10, 0.13) and a credible negative effect of phase 1 confidence (95% HDI: -0.30, -0.32).

Phase 2 confidence. The more confident agents were in their phase 1 estimates, the more confident they were in their phase 2 estimates. However, there was no effect of dialectical instructions on phase 2 confidence. We regressed phase 2 confidence (C_2) on two fixed factors: dummy-coded phase 2 estimate procedure condition and phase 1 confidence (C_1). We found a credible positive effect for phase 1 confidence (95% HDI: 0.20, .21) suggesting that the more confident an agent was for a question in phase 1, the more confident the agent was that question in phase 2. We did not find a credible effect for the dialectical condition (95% HDI: -0.02, -0.01).

---

[9] Technically, we could increase the number of agents in our simulation to such a large size that we would not need to conduct any inferential tests. However, because the simulations were computationally intensive, we elected to restrict the number of agents to 5,000.

PAR parameters. Dialectical agents had higher phase accuracy ratios (A_{phase}), confidence accuracy ratios (A_{conf}), and bracketing percentages (BP) relative to controls: We conducted three separate mixed-level regression analyses on phase accuracy ratios, confidence accuracy ratios and bracketing percentages. For each regression, we entered an indicator variable for phase 2 condition as a fixed factor and random intercepts for each agent. The effect of condition was positive and credible for phase accuracy ratios (95% HDI: 0.00, 0.13), confidence accuracy ratios (95% HDI: 0.24, 0.47), and bracketing percentages (95% HDI: 0.08, 0.10).

Confidence calibration and resolution. Agents gave confidence intervals that were poorly calibrated (relative to 90%). Across both estimate phases, the average agent produced confidence intervals that captured the criterion in 60.6% [IQR: 43.8%, 81.3%] of cases. However, and most importantly, agents' confidence intervals had positive resolution: the more confident agents were, the more accurate their best estimates were. For each estimate phase, we regressed the absolute deviation (AD) on confidence (C). In both phases 1 and 2 we found a credible negative effect of confidence on absolute deviation (Phase 1: 95% HDI: 0.007, 0.070; Phase 2 95% HDI: 0.103, 0.106). This suggests that confidence is indeed a valid cue to accuracy in the county estimation task. This finding is critical. If confidence was not a valid cue to accuracy, than confidence-based aggregation could not outperform the simple average.

Comparing inner-MCS to averaging. Taken together, our simulation suggests that confidence intervals provide a window into the estimation process: confidence in phase 1 predicts estimate change and phase 2 confidence. Most importantly, they also suggest that confidence is indeed correlated with accuracy within (simulated) one mind. Thus, confidence could potentially be used by an agent (or person) to detect its most accurate estimates (Soll &

Larrick, 2009). To our knowledge, this is the first time a cognitive model has simultaneously modeled the processes underlying confidence and estimation and found a positive relationship between the two.

Next, we explore how agents should aggregate their estimates to generate a single estimate from the original two. Our finding that confidence and accuracy are correlated suggests that confidence could be used as an estimate-weighting cue for these stimuli. But will confidence-based estimation beat the simple average? To test this, we compared the performance of two strategies: *Average*, where agents take the average of their best estimates from phases 1 and 2, and *Inner-MCS*, a heuristic inspired by Koriat's (2012b) maximum-confidence slating heuristic, where agents choose their estimate with the higher confidence for a given question[10]. For each agent, we compared the mean absolute deviations (MAD) values of these strategies to two reference strategies: *Choose randomly*, where agents randomly choose between their phase 1 and phase 2 estimates for each problem, and *Choose first*, where agents always chose their phase 1 estimates. See Table 2 for a list of strategies and error labels.

Strategy Name	**Description**	**Error Label**
Average	Average best estimates for a problem across phases 1 and 2.	$MAD_{Average}$
Inner-MCS	Choose the best estimate corresponding to the phase with the smallest confidence interval.	$MAD_{InnerMCS}$

[10] If agents have the same confidence levels for both estimates to a question, they choose randomly.

Choose randomly	Choose a phase 1 or phase 2 estimate at random	$MAD_{ChooseRandom}$
Choose first	Choose the phase 1 best estimate	$MAD_{ChooseFirst}$

Table 2: Description of strategies for managing the inner-crowd.

For each agent, we calculated its accuracy improvement for the two target strategies (Average and Inner-MCS) over the reference strategies by subtracting the MAD of the target strategy from the MAD of the reference strategy. For example, the accuracy gain of Inner-MCS over Choose randomly is calculated as $MAD_{ChooseRandom} - MAD_{InnerMCS}$. Positive values in this difference indicate less error and higher accuracy for the target strategy. A distribution of these gains separated by phase 2 procedures is presented in Figure 2.

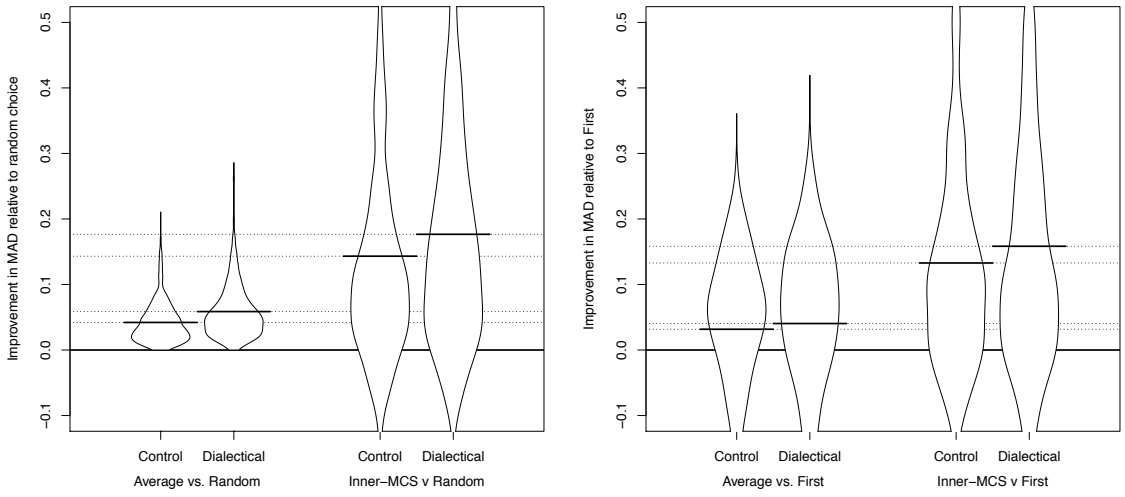


Figure 2: Distribution of accuracy gains in the simulation. The left and right panels show gains relative to *Choose randomly* and *Choose first* respectively. Higher values

indicate better performance by Average and Inner-MCS. The horizontal lines show median values.

We found much higher accuracy gains from Inner-MCS than from Average relative to both reference strategies. Across both phase 2 conditions, the mean accuracy gain from Average over Choose randomly was 0.044 [IQR: 0.019, 0.059], while the mean accuracy gain from Inner-MCS was almost four times as large at 0.154 [IQR: 0.031, 0.239]. Gains were similar relative to Choose first. Additionally, relative to both reference strategies, simulated dialectical instructions increased the gains from both Average and Inner-MCS. These results provide strong empirical support for the benefits of confidence-based choice in the inner-crowd in the US county domain.

Boundary conditions for inner-MCS gains. However, we note that not all agents benefited from using Inner-MCS relative to Average. We found that 30% of agents had worse performance (i.e.; higher MAD values) by using Inner-MCS than by using Average. To see what variables contributed to this effect, we conducted a Bayesian mixed logistic regression where we regressed the binary variable $MAD_{Average} < MAD_{InnerMCS}$ indicating when an agent had lower error from Average versus Inner-MCS on two fixed factors: the agent's bracketing percentage (BP), and the agent's confidence accuracy ratio (A_{conf}). We found a credible positive effect of bracketing percentage (95%HDI: 2.71, 3.30,) and a credible negative effect of confidence accuracy ratio (95% HDI: -3.09, -2.88). The higher an agent's bracketing rate and the lower its confidence accuracy ratio, the less likely it benefited from Inner-MCS over Average.

Simulation Discussion

Our simulation analysis generated three key results that we use as predictions in our empirical study. First, confidence in phase 1 estimates predict both best estimate change between first estimates and second estimates as well as confidence in phase 2 estimates. This suggests that confidence measures provide us with a window into a person's estimation process that best estimate measures are silent to. Second, confidence predicts accuracy. As a result of this relationship, people may benefit from choosing their high-confidence estimates using Inner-MCS instead of averaging. Moreover the degree to which people benefit from Inner-MCS should be a function of their confidence accuracy ratio. Finally, simulated dialectical instructions that caused agents to search for new information (new exemplars with a different memory probe) increased best estimate changes, averaging gains, and Inner-MCS gains.

Empirical Study:

Testing Predictions On The Inner Crowd, Confidence and Dialectical Bootstrapping

To see if these results would carry over to human judges, we conducted a study using the same stimuli and estimation paradigm used in the simulation. We had three main experimental conditions in the study that changed how participants made their estimates in phase 2. In the *control* condition, participants were asked to make another estimate as if they were answering it for the first time. This condition aims to assess the natural variability in people's estimates (see also Herzog & Hertwig, 2009, 2014b). We also included two *dialectical* conditions that were designed to increase estimate diversity. In the *dialectical consider-the-opposite* condition, participants were given dialectical instructions identical to those used in Herzog and Hertwig (2009). The instructions implored participants to think of reasons why their first estimates were wrong, and—based on those reasons—to apply a new estimation

strategy to their subsequent estimates (see Method section for the verbatim instructions). A benefit of this dialectical technique is that is purposefully generic and can thus be applied to different kinds of estimation strategies (e.g., from non-compensatory rule-based to exemplar-based strategies; von Helversen & Rieskamp, 2009). However, because the technique does not provide judges with specific instructions on how they could change their estimates, we anticipated that some participants might have difficulty deriving an alternative strategy. To help participants derive a new strategy, we created a new estimation strategy tailored to exemplar-based estimation processing that we call *consider-other-exemplars*. In the *dialectical consider-other-exemplars* condition, participants were asked to think of additional exemplars that matched the target cue profile, but whose population was likely to be different from their estimates in phase 1.

After completing their first two sets of estimates in phases 1 and 2, participants in all conditions were presented with each question again, along with their first two sets of judgments, and with all county cue profile information removed. They were then asked to make a third and final judgment based entirely on their previous confidence and best estimate judgments. This phase was designed to test the extent to which people use their confidence judgments in aggregating their inner-crowd.

Method

Participants

300 US-based adults (166 men and 134 women) were recruited online from the Amazon Mechanical Turk[11] (Mason & Suri, 2012). Participants were compensated $4.00 for their

[11] For those not familiar with the Amazon Mechanical Turk (mTurk), the mTurk is an online recruitment tool where "Requesters" (i.e.; employers) post "HITs" (an acronym for a "human intelligence task" representing a one-time job) which can be completed by "Workers" (i.e.; participants). Anyone in the general public can complete a

participation. Additionally, they were given a performance-based bonus based on the accuracy of the their first, second and third estimates[12]. The median bonus was $3.70 (IQR: $3.40, $3.90). Participants worked a median of 45 minutes (IQR [34, 61]) on the study.

Materials and Procedure

After giving consent to participate in the study, participants were told that the study would take place in several phases and that each phase would take approximately 10 minutes. They were instructed to not use any outside help in the form of an Internet search or calculation tool in any of the phases. They then read the following instructions for phase 1 of the study:

> "In Phase 1 of the study, you will be estimating the populations of US counties based on statistics about those counties. Here's how the task will work: We created 16 different groups of US counties based on their values on 4 different statistics (you will learn what they are shortly). Within each of the 16 groups, all the counties have similar values on the statistics. From each of the 16 groups, we selected a typical county. Your task is to estimate the population of each of these 16 typical counties. We will not tell you the name of each county. Instead, we will show you the statistical information about the county, and then ask you to make a population estimate based on those statistics. Your goal is to come up with a population estimate that is as close as possible to each county's true population. The closer your estimates are to the true populations, the higher your bonus will be! The highest bonus you can earn for Phase 1 is $2.00. You

---

HIT in exchange for payment. We required that participants in our study had successfully completed 50 HITs in the past with at least a 95% acceptance rate from previous requesters.

[12] We awarded participants with two separate bonuses. The first bonus was determined by the most accurate of their phase 1 and phase 2 estimates for each stimulus. The second bonus was determined by the accuracy of their phase 3 estimate for each stimulus. The maximum possible bonus for phase 1 and phase 2 estimates was $2.00, and the maximum possible bonus for phase 3 estimates was $2.00.

will see 4 different statistics for each of the counties. For each county, you will see whether or not the county has a HIGH or a LOW value on each of the 4 statistics." Participants then read descriptions of the four different cues. Next, participants viewed a sample version of the experimental stimuli with "??" replacing the cue values. They were then instructed to give a best estimate and upper and lower bounds of confidence intervals such that each interval had a 90% chance of containing the true county population (Russo & Shoemaker, 1992).

In phase 1, participants gave population estimates for each of the 16 stimuli. Stimuli were presented in the form of a 4 x 2 matrix, where each cue and its value was presented on a row (for a screenshot, see Appendix C). Participants were forced by the questionnaire to provide best estimates that were between the limits defined by their confidence intervals. Participants did not receive any feedback on their accuracy of their confidence intervals or their best estimates. We created two stimuli orders, one randomly generated and its reverse. Additionally, we created four different cue orders such that each cue was presented in each row across participants. Each participant was assigned to one of the stimuli and cue orders and viewed the stimuli in the same order in each of the three estimate phases.

After making their initial estimates for the 16 stimuli in phase 1, participants received instructions for phase 2. All were told that they would be shown the questions again and would be asked to give a second round of estimates. They were told that their bonus would be based on the *better* of their two estimates for each problem. This was meant to encourage participants to make different estimates in the second round (see Herzog & Hertwig, 2009, 2014b). Participants assigned to the control condition were not given explicit instructions on how to answer the questions but were simply told to answer them as if they had seen them again for the

first time (see Herzog & Hertwig, 2009, 2014b). Participants in the dialectical-consider the opposite (D-CTO) condition were given the following instructions (Herzog & Hertwig, 2009, p. 234):

> First, assume that your first estimate was off the mark. Second, think about a few reasons why that could be. Which assumptions and considerations could have been wrong? Third, what do these new considerations imply? Was your first estimate too high or too low? Fourth, based on this new perspective, make a second, alternative estimate.

Participants in the dialectical-consider other exemplars (D-COE) condition were instructed to think of example counties that matched the cue profile for each question, but whose size was substantially different from their first estimate as follows:

> For each question, look at the information you were provided with and the estimate you gave. Try to think of examples of counties that match the information you were given, but are likely to have a very different size than the estimate you gave. Then, using this information, try to come up with a new alternative estimate.

After reading phase 2 instructions, participants viewed the questions again along with their best estimates and confidence intervals for that stimulus from phase 1. Participants then gave confidence intervals and best estimates in the same manner as Phase 1.

After giving their second set of estimates, participants were told (again without warning) that they would be giving a third and final set of estimates for all problems. They were told that they would receive an additional bonus depending on the accuracy of their third set of estimates that was independent of the bonus they received in phases 1 and 2. For these final estimates, participants were not shown the stimuli, but instead were asked to make

estimates based purely on their previous estimates as follows: "County XXX. You first best estimate was ___, Your first confidence interval was ___ and ___. Your second best estimate was ___. Your second confidence interval was ___ and ___]. What is your current confidence interval for this county: ___ and ____. What is your current best estimate for the population of this county?" After completing the third estimation phase, participants completed a brief comprehension survey.

Results

As in our simulation results, we used Bayesian mixed-level regression analyses to test our hypotheses (Baayen, Davidson & Bates, 2008) using the MCMCglmm package in R (Hadfield, 2010). Except when specified otherwise, we included random intercepts assigned for both participants and stimuli. All raw data and code are available in our supplementary materials.

Estimate change. Dialectical estimates increased estimate change from phases 1 and 2, and phase 1 confidence was negatively related to estimate change: We regressed estimate change (Δb_{12}) on three fixed factors: two indicator variables indicating the agent's phase 2 estimate procedure condition (with the repeated condition coded as 0 and the dialectical conditions coded as 1 in the two separate variables) and phase 1 confidence (C_1). We found credible positive effects for both the D-CTO (95% HDI: 0.00, 0.28) and D-COE (95%HDI: 0.08, 0.35) conditions, and a credible negative effect of phase 1 confidence (95% HDI: -0.74, -0.70). These effects are consistent with our simulation results.

Phase 2 confidence. The more confident participants were in their phase 1 estimates, the more confident they were in their phase 2 estimates. Additionally, there was no credible effect of dialectical instructions on phase 2 confidence: We regressed phase 2 confidence (C_2)

on three fixed factors: two indicator variables for phase 2 estimate condition and phase 1 confidence (C_1). We found a credible positive effect for phase 1 confidence (95% HDI: 0.64, .68) suggesting that the more confident a participant was for a question in phase 1, the more confident the participant was that question in phase 2. We did not find a credible effect for the dialectical condition (D-CTO: 95% HDI: -0.05, 0.14; D-COE: 95% HDI: -.05, 0.14). These effects are consistent with our simulation results.

PAR parameters. Consider-other-exemplars instructions consistently increased phase accuracy ratios, confidence accuracy ratios, and bracketing percentages. Consider-the-opposite instructions had less consistent effects: We conducted three separate mixed-level regression analyses on phase accuracy ratios, confidence accuracy ratios, bracketing percentages. For each regression, we entered two indicator variables for phase 2 condition as fixed factors and random intercepts for each agent. We present posterior means and 95% highest density intervals for phase 2 condition in Table 3:

	Consider-the-opposite	Consider-other-exemplars
Phase Accuracy Ratio (A_{phase})	0.13 [-0.14, 0.39]	0.32 [0.06, 0.60]
Confidence Accuracy Ratio (A_{conf})	0.56 [0.12, 0.99]	0.85 [0.40, 1.28]
Bracketing Percentage (BP)	0.01[-0.03, 0.05]	0.05 [0.01, 0.09]

Table 3: Effects of dialectical instructions on PAR parameters relative to the control condition.

The sign of the effects of the consider-other-exemplars instructions mirrored those of our simulation.

Confidence calibration and resolution. Participants' confidence intervals were poorly calibrated and even less so than agents in our simulation. Across both estimate phases, the average participant produced confidence intervals that captured the criterion in only 34.4% [IQR: 12.5%, 50.0%] of cases. These values are consistent with prior research showing that confidence interval ranges tend to be far too narrow (e.g., Lichtenstein et al., 1982; Soll & Klayman, 2004; Yaniv & Foster, 1997; Yates, 1990). However, and most importantly for confidence-based aggregation, participants' confidence intervals had credibly positive resolution. For each estimate phase, we regressed absolute deviation (AD) on confidence (C). In both phases 1 and 2 we found a credible negative effect of confidence on absolute deviation (Phase 1: 95% HDI: [0.70, 0.74]; Phase 2 95% HDI: [0.71, 0.74]). Consistent with our simulation results, this suggests that confidence is indeed a valid cue to accuracy.

Comparing Inner-MCS to Average. Next, we compared the accuracy of Inner-MCS to Average. As in our simulation analyses, we calculated accuracy gains relative to two reference strategies: Choose randomly, and Choose first. A distribution of these gains separated by experimental condition is presented in Figure 3:

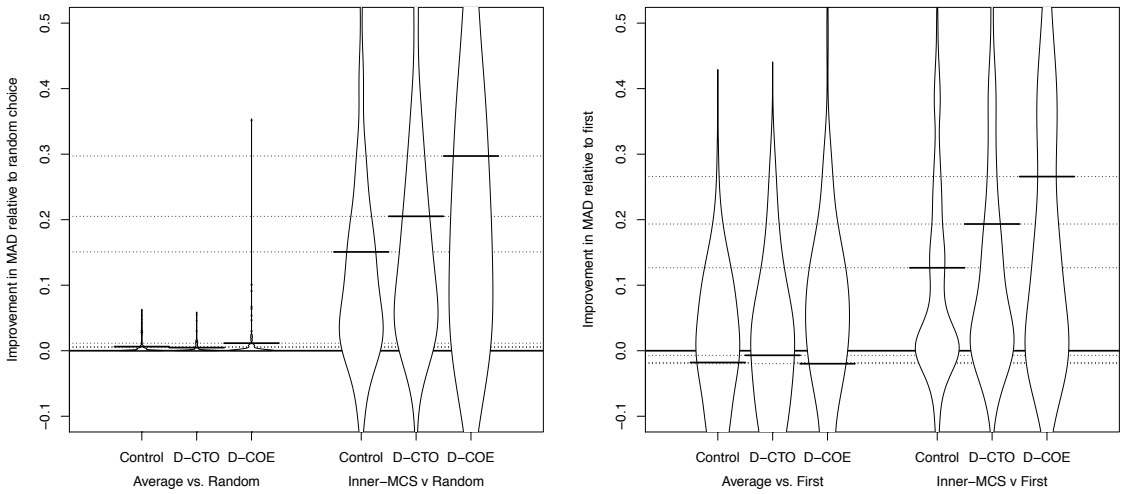


Figure 3: Distribution of accuracy gains for Average and Inner-MCS strategies in the study. The left and right panels show gains relative to *Choose randomly* and *Choose first* reference strategies respectively. Positive values indicate higher accuracy for Average and Inner-MCS. The horizontal lines show median values.

As Figure 3 shows, Inner-MCS outperformed Average relative to both reference strategies and in all experimental conditions. Across all phase 2 conditions, the mean accuracy gain from Average over Choose randomly was 0.007 [IQR: 0.000, 0.005], while the mean accuracy gain from Inner-MCS was much larger at 0.217 [IQR: 0.034, 0.290]. Gains were similar relative to Choose first. Additionally, relative to both reference strategies, dialectical

instructions increased the gains from both Average and Inner-MCS, with the highest gains in the dialectical-consider other exemplars condition.

While the benefits of high-confidence choosing in the study largely replicated those from our simulation (Figure 2), the benefits of Average in our study were much smaller than our simulation results. We suspect this is due at least in part to the small bracketing rates in the study. In the simulation, agents had a mean bracketing rate of 33% (IQR: 25%, 44%). In the study, this dropped to 16% (IQR: 6%, 25%). Because bracketing rates (an indicator of error correlation) drive averaging gains, our participants with small bracketing rates did not reap large averaging gains.

Boundary conditions for inner-MCS gains. As in our simulation, not all participants benefited from Inner-MCS relative to Average. We found that 14% of participants had worse performance by using high-confidence choosing relative to a simple average[13]. To see what variables contributed to this effect, we replicated the Bayesian logistic regression analysis from the simulation where we regressed the binary variable $MAD_{Average} < MAD_{InnerMCS}$ (indicating when a participant had lower error from Average versus Inner-MCS) on two fixed factors: the participant's bracketing percentage (BP), and the agent's confidence accuracy ratio (A_{conf}). We found a credible positive effect of bracketing percentage (95%HDI: 0.31, 2.72,) and a credible negative effect of confidence accuracy ratio (95% HDI: -0.50, 0.37). The higher a participant's bracketing percentage and the lower its confidence accuracy ratio, the less likely she benefited from Inner-MCS over Average.

[13] This value was lower than the percentage we observed in our simulation, where 30% of agents had higher accuracy from Average versus Inner-MCS. The lower percentage in our study could at least partially be due to the smaller bracketing rates in the study (16%) compared to the simulation (33%).

Confidence choosing (Inner-MCS) versus confidence weighting. Inner-MCS represents an extreme form of confidence-based estimation; namely, choosing the high confidence estimate and ignoring the low confidence estimate. How does confidence-based choice compare to confidence-based *weighting* – a strategy that can be normative in advice taking (Yaniv, 1997)? To test this, we calculated the optimal w_H value for each participant that minimized their MAD_{HC} value. If Inner-MCS outperforms confidence weighting, then most participants' optimal w_H value will be close to 1 (otherwise optimal w_H would be substantially less than 1). In Figure 4, we plot the cumulative distribution of our study participants' optimal w_H values next to the same values for agents from our simulation:

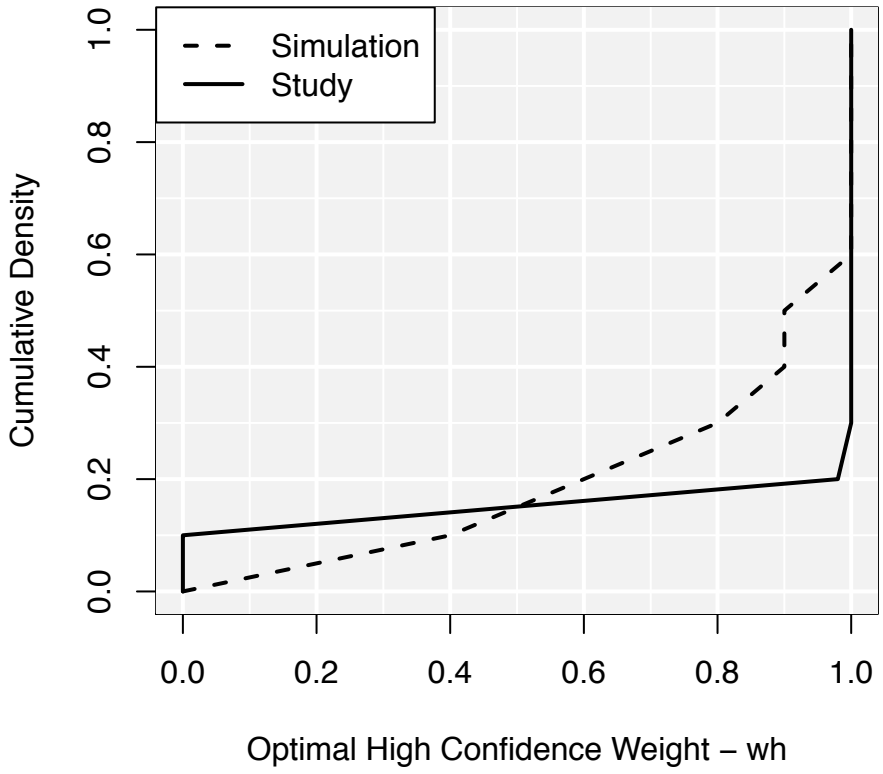


Figure 4: Cumulative density of optimal high confidence weights (w_H) for study participants and simulation agents.

As Figure 4 shows, the vast majority (80%) of our study participants had optimal w_H values greater than 0.90, while a full 70% of participants had optimal w_H values at 1.0, equivalent to Inner-MCS. For simulated agents, the magnitude of optimal w_H values was also quite high: a full 60% of agents had optimal w_H values greater than 90%. These results suggest that Inner-MCS either matches, or closely approximates the optimal confidence-based aggregation strategy for most participants and agents.

How did participants aggregate their first and second estimates?

In the next section, we analyze how participants aggregated their inner-crowd in phase 3. In this phase, participants were presented with their phase 1 and phase 2 estimates (both best estimates and confidence intervals) and were asked to make a single best estimate solely on their previous estimates. Our goal in this analysis is to see to what extent our participants actually used confidence as a cue in aggregating their phase 1 and phase 2 estimates.

Before describing how we modeled phase 3 estimates, let us briefly define a few terms. Again, we label b_i, l_i and u_i as the best-estimate, confidence-interval lower-bound estimate, and confidence-interval upper bound estimate for phase *i*. We then model phase 3 estimates using equation 1 (e.g., Soll & Larrick, 2009; Herzog & Hertwig, 2014b; Müller-Trede, 2011):

$$w_1 \times b_1 + (1-w_1) \times b_2 = b_3 \qquad \text{EQ 1.}$$

Rearranging terms allows us to define w_1 as $(b_3 - b_2) / (b_1 - b_2)$. The w_1 parameter defines how much weight is given to b_1, with $w_1 = 1$ meaning that the third estimate is equal to b_1 and $w_1 = 0$ meaning that it equals b_2. When b_3 estimates fall outside of the range of b_1 and b_2, w_1 is either less than 0 or greater than 1. For our phase 3 analyses, we ignore these estimates that fall outside of the range.

We use the term *aggregation* strategy as an umbrella term for *all* possible strategies that return w_1 values in the (inclusive) range [0, 1]. We define *combining* strategies as any strategy that returns a w_1 value in the (exclusive) range (0, 1). We further distinguish three types of aggregation strategies: *weighted-averaging* strategies as those where w_1 is in the interval (0, .4] or [.6, 1). When participants use a weighted-averaging strategy, they weight both estimates but give more weight to one estimate than the other. We define *averaging* strategies are those that give relatively equal weight to both estimates and return w_1 in the interval (.4, .6). Finally, we define *choosing* strategies as those where w_1 is equal to 0 or 1. When participants use a choosing strategy, they elect to choose between their previous estimates and in the process completely reject one of them. Early research in advice-taking tasks concluded that people tend to use weighted-averaging strategies (e.g., Budescu, Rantilla, Yu & Karelitz, 2003; Yaniv, 2004a), while more recent analyses performed at the individual level suggests that people mostly choose and only occasionally average (Soll & Larrick, 2009). When aggregating estimates in their inner-crowd, people tend to use an averaging strategy, but not consistently (Herzog & Hertwig, 2014b; Fraundorf & Benjamin, 2014).

In addition to distinguishing between combining and choosing strategies, we also distinguish strategies on how they represent estimates: *order-based* versus *confidence-based*. Strategies that are *order-based* represent estimates as a function of the order (or "phase" in our paradigm) they were generated. Several order-based strategies have been studied in the judgment and decision making literature, with several such as the "Take the First" (Johnson & Raab, 2003) and "First Instinct Fallacy" (Kruger, Wirtz & Miller, 2005). However, alternative judgment and decision-making models such as belief-updating models (Hogarth & Einhorn, 1992) and sequential sampling models (Hourihan & Benjamin, 2010) predict that people can

have a tendency to give more weight to later, more recent estimates. Strategies that are *confidence-based* represent estimates as a function of their associated confidence, where high-confidence estimates are assumed to be weighted more than low-confidence estimates. While confidence-based strategies are common in advice-taking tasks (e.g.; Sniezek & Van Swol, 2001; Yaniv, 1997; Moussaïd et al., 2013; Soll & Larrick, 2009), it is unclear whether people use confidence within one mind.

To represent confidence-based strategies using equation 1, we replace b_1 with b_H, and b_2 with b_L. This means that for confidence-based strategies, *w* represented the weight given to high-confidence estimates. We use the index h (w_H) to indicate weights for confidence-based strategies.

Modeling strategy use. We applied equation 1 to estimates to obtain a vector of w_1 (for order-based strategies) and w_H (for confidence-based strategies) values for each participant across stimuli. We used these vectors of *w* weights to compare strategies by fitting this vector of *w* values to each strategy. As has been noted by other researchers on advice-taking (Soll & Larrick, 2009), people rarely use one strategy exclusively across problems. To accommodate this intra-individual variability in strategy use, we model each strategy with a probability distribution that captures variability in strategy use while simultaneously measuring general tendencies. We model combining strategies (strategies 1, 3, and 5) using t-distributions bounded on the range [0, 1] with one or two free parameters[14] (see Table 4) corresponding to the mean and standard deviation. We model choosing strategies (strategies 2, 4 and 6) using beta distributions with either one or two free parameters corresponding to α and β (see Table 4) We compared 6 different combination strategies for each participant in addition to a random, null model. A description of each strategy is presented in Table 4 and a visual representation of the strategies is presented in Figure 5:

Strategy	Estimate representation	Aggregation type	Distribution	Parameters
0 - Random	Symmetric	NA	Unif(0, 1)	0
1 – Equal weighting	Symmetric	Combining	t(μ = .5, σ, df = 1)	1 (σ)
2 – Random Choosing	Symmetric	Choosing	beta(α, α)	1 (α < 1)
3 – Order-Based weighting	Phase	Combining	t(μ, σ, df = 1)	2 (μ, σ)

[14] For symmetric strategies, we set the mean of the t-distributions (for combining strategies) to 0.50, and forced the alpha and beta parameters of the beta distribution (for choosing strategies) to be equal. In contrast, for the non-symmetric strategies, we allowed the means of the t-distributions (for combining strategies) to be free and both alpha and beta parameters of the beta distributions (for choosing strategies) to be free.

4 – Order-Based choosing	Phase	Choosing	beta(α, β)	2 (α < 1, β <1)
5 – Confidence – Based weighting	Confidence	Combining	t(μ, σ, df = 1)	2 (μ, σ)
6 – Confidence Based choosing	Confidence	Choosing	beta(α, β)	2 (α < 1, β < 1)

Table 4: Description of six different phase 3 aggregation strategies

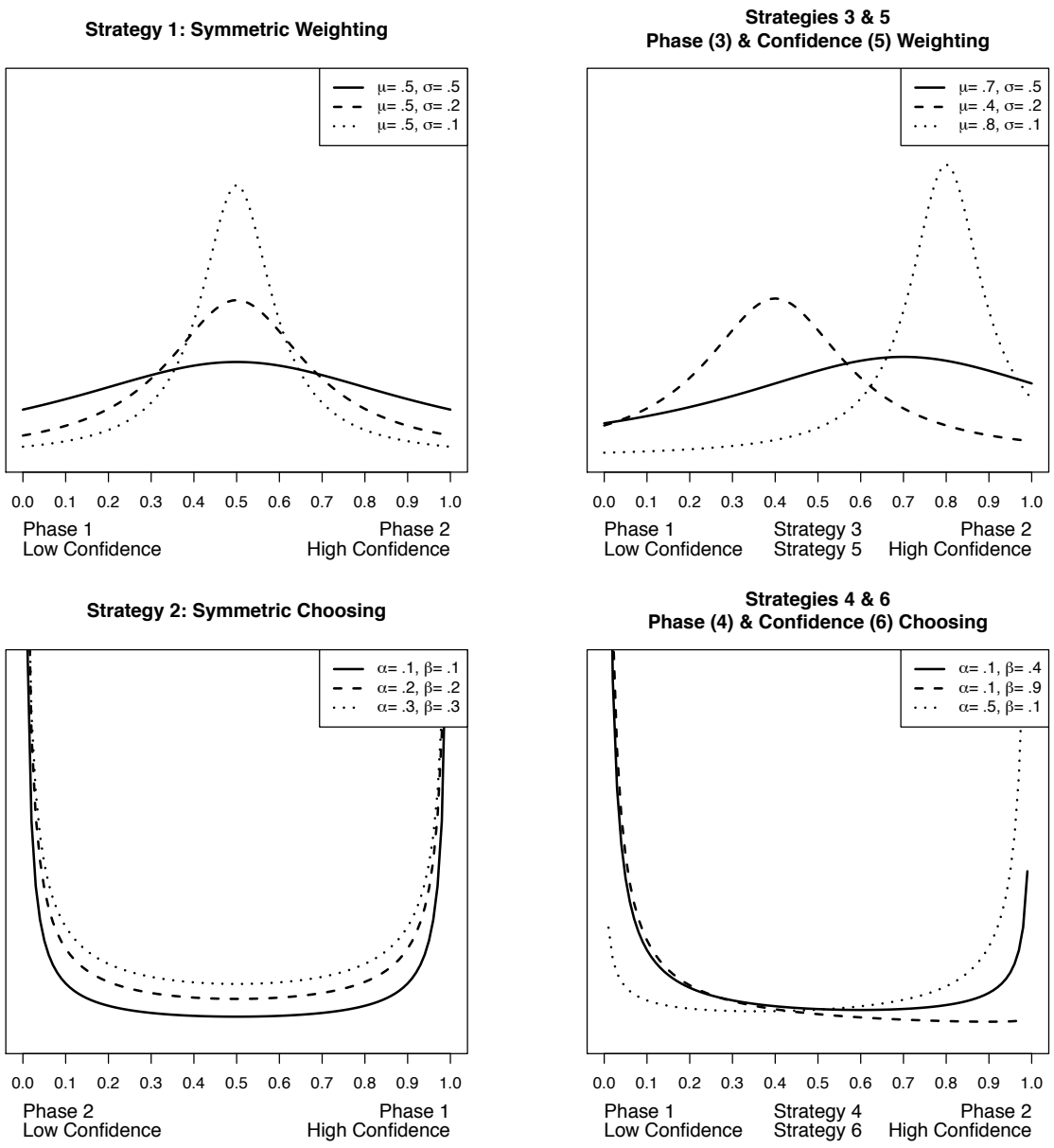


Figure 5: Examples of the 6 different aggregation strategies. The top-left panel shows examples of strategy 1, a symmetric weighting strategy. The bottom-left panel shows examples of strategies 2, a symmetric choosing strategy. The top-right panel shows strategies 3 and 5. These are non-symmetric weighting strategies, where strategy 3 is order-based and strategy 5 is confidence-based. The bottom-right panel shows strategies

4 and 6. These are symmetric choosing strategies, where strategy 4 chooses according to phase, and strategy 4 chooses according to confidence.

Strategy classification. For each participant, we calculated maximum likelihood estimates for each parameter in each strategy. Next, we calculated the Bayesian Information Criterion (BIC; Wagenmakers, 2007) for each model using the following equation, with the index *m* referring to a model (m in {0, 1, … 6}), *k* referring to a stimulus (k in {1, 2, … 16}), and p_m referring to the number of parameters in model *m* (p_m in {0, 1, 2}):

$$BIC_m = -2\sum_{k=1}^{16} ln\big(lik_m(b_k)\big) + p_m \ln(N)$$

Where lik_m is the likelihood of the data given model *m* using the maximum likelihood estimates for each parameter p_m in model m, and N is the number of data points[15]. The BIC measure rewards models that give high maximum-likelihoods to the data while simultaneously punishing models with many free parameters. We then calculated ΔBIC values for each model by subtracting the minimum BIC value from each model's BIC value. Finally, we calculated posterior probabilities of each model m using the equation

$$Post_m = \frac{e^{-.5*\Delta BIC_m}}{\sum_{i=0}^{6} e^{-.5*\Delta BIC_i}}$$

We classified each participant as using the model with the highest posterior model probability.

[15] For order-based models, N is always 16. However, for confidence-based models, when a participant gives the same confidence (i.e.; the same interval range) for both phase 1 and phase 2 estimates, w_H is undefined and cannot be fit by confidence-based models. Thus, for confidence-based models, N is the number of questions where a participant gave different confidence to both estimates. To keep the comparison between order-based and confidence-based models fair, we fit both models to the same questions for each participant. We also only modeled participants with at least ten valid (i.e.; possible to fit) data points.

What strategies did participants use? A table of the distribution of best fitting models across participants is presented in Table 5.

Strategy	N	Percent	Posterior Model probability (Median and IQR)
Insufficient Data	119	NA	NA
S0: Random weights / Unclassified	33	19.29%	48.9% [40.2%, 55.5%]
S1: Symmetric weighting	67	39.18%	57.9% [47.3%, 62.7%]
S2: Symmetric choosing	6	3.51%	51.1% [47.3%, 51.7%]
S3: Phase weighting	22	12.87%	56.2% [45.8%, 75.0%]
S4: Phase choosing	13	7.60%	80.0% [61.4%, 98.1%]
S5: Confidence weighting	26	15.2%	67.8% [50.4%, 81.2%]
S6: Confidence choosing	4	2.34%	52.8% [50.7%, 64.7%]

Table 5: Frequencies of phase 3 aggregation strategy classifications.

Aggregating over strategy types, we find two main results. First, weighting strategies were more common than choosing strategies: Of those participants who were not classified to random weight/unclassified strategy, 83% used a weighting strategy while 17% used a choosing strategy. Second, phase combination strategies were about as common as confidence combination strategies: Of those participants who used a non-symmetric strategy, 46% used a confidence-based strategy, while the remaining 54% used a phase-based strategy.

Next, we look at the distribution of parameters in non-symmetric strategies (strategies 3-6). The distribution of parameters in these strategies tells us whether or not participants give more weight to first or second estimates (for phase-based strategies), or high confidence or low

confidence estimates (for confidence -based strategies). For combining strategies (strategies 3 and 5), we focus on the distributions of μ parameters (strategies 3 and 5). Values of μ greater than .50 indicate higher weight on phase 1 or high confidence estimates for phase and confidence based strategies respectively. For choosing strategies (strategies 4 and 6), we look at the difference in estimated β and α values. When α is greater than β (positive difference), this suggests a higher rate of choosing phase 2 estimates or high confidence estimates for phase and confidence based strategies respectively. Summary statistics are presented in Table 6.

	Parameter	N	Median	Proportion preferring second estimate (Strategies 3 and 4) or high confidence estimate (Strategies 5 and 6)
Strategy 3: Phase Weight (Mu)	μ	22	.29	.91
Strategy 5: Confidence Weight (mu)	μ	26	.65	.88
Strategy 4: Phase Choosing	α - β	13	-.22	1.00
Strategy 6: Confidence Choosing	α - β	4	.62	1.00

Table 6: Summary statistics of phase 3 aggregation model parameters.

We begin by looking at the distribution of parameters for phase-based strategies (3 and 4). Collapsed over combining and choosing strategies, 88% had parameter values that favored phase 2 estimates over phase 1 estimates ($\mu < .5$ for strategy 3, and $\alpha > \beta$ for strategy 4). Therefore, participants who weighted or chose estimates based on the order they were made tended to prefer their second estimates to their first estimates. Next, we look at the distribution of parameters for confidence-based strategies (4 and 6). Collapsed over combining and choosing strategies, 24 out of 29 (83%) had parameter values that favored high-confidence

estimates over low-confidence estimates ($\mu < .5$ for strategy 5 and $\alpha > \beta$ for strategy 6). Therefore, participants who weighted or chose estimates based on confidence preferred their high- confidence estimates.

General Discussion

Extant research on the inner-crowd has explored how people can improve their judgments by generating multiple estimates and taking the simple average (Herzog & Hertwig, 2014a). In the current paper, we explored how people can, and should, use confidence associated with their estimates in deciding how to aggregate them. In an agent-based simulation, we found that confidence should be correlated with accuracy in our cue-based estimation task and that confidence-based choosing outperformed averaging. Further, simulated dialectical instructions that made agents diversity their estimates within the framework of the Naïve Sampling Model (Juslin et al., 2007) increased high-confidence choosing gains. In an empirical study, we replicated these effects and found that people could outperform the average of their inner-crowd with high-confidence choosing (as implemented by Inner-MCS).

Confidence predicts accuracy in the inner-crowd.

Given the large number of studies finding that people's confidence judgments are too high given their empirical accuracy (e.g.; Griffin & Brenner, 2004; Soll & Klayman, 2004), one might conclude that confidence is the result of a biased information processing system and thus should be ignored during an aggregation procedure. Our results do not support this view. While people's individual confidence intervals may be poorly calibrated, we find that they exhibit sufficiently high resolution to allow for confidence-based aggregation strategies (such as Inner-MCS) to out-perform others that ignore confidence (such as Average).

These results mirror previous research that found benefits of confidence-based aggregation between minds (e.g.; Koriat, 2012b, Yaniv, 1997) and extends them to the inner-crowd. Moreover, because our results were predicted by the NSM (Juslin et al., 2007), they are consistent with the idea that people act as naïve intuitive statisticians who process sample information (i.e.; samples in working memory) in an unbiased manner, but who are naïve with respect to biases that can occur in the sampling process (Fiedler & Juslin, 2000). Unfortunately, much (if not most) research on confidence in judgment seems to have focused on when confidence goes wrong with respect to calibration, and has ignored when it can go right with respect to resolution.

Choosing vs. Averaging Environments In the Inner Crowd

Our finding that confidence can improve inner-crowd judgments provides a new perspective on the ecological rationality of averaging versus choosing between inner-crowd estimates. To show this, we refer back to the PAR (Soll & Larrick, 2009) model. Again, the PAR model states how one should combine estimates from two advisers (call them A and B) each of whom are providing estimates across a number of problems. Under the PAR, the benefits of choosing estimates from one adviser (and ignoring the other) increase when the relative accuracy of one adviser to the other increases and the probability of detecting the better (i.e.; more accurate) of the two advisers increases.

We can reframe the PAR model in the inner-crowd by thinking of the inner-adviser A as “Phase 1” and inner-adviser B as “Phase 2.” In previous inner-crowd research, accuracy ratios (based on phases) in the inner-crowd were fairly low (condition means of 1.12, 1.11 and 1.09), Herzog & Hertwig, 2014b) which benefited averaging strategies. In our study, phase-based accuracy ratios were higher than previous studies at a median of 1.47 which favored

choosing strategies. Importantly, confidence-based accuracy ratios were even higher at a median of 1.78. In PAR terms, confidence allowed our participants to generate a new set of virtual inner-advisers "High Confidence" and "Low Confidence" from their inner-crowd that had a higher accuracy ratio than their original inner-advisers "Phase 1" and "Phase 2." In addition, using confidence to generate new advisers has the (likely) added benefit of increasing the probability of detecting the more accurate adviser. Previous research suggests that people are not very good at detecting whether or not their phase 1 or phase 2 estimates are more accurate (Fraundorf & Benjamin, 2014), which makes choosing strategies less accurate. A reason for this is that people differ dramatically in whether their first or second estimates are more accurate: in our study, 49%[16] of participants' phase 1 estimates were more accurate than their phase 2 estimates. Thus, unless individual participants have valid introspective insight into whether their first or second estimates are better, they can do no better than a coin-flip at detecting which phase is more accurate. In contrast, a full 87% of our participants' high confidence estimates were, on average, more accurate than their low confidence estimates. This means that people should be much better at detecting their more accurate inner confidence-based adviser than their inner phase-based adviser. These two effects of confidence, increased accuracy ratios and increased probability of detecting the better adviser, both increase the benefits of choosing over averaging.

However, not all participants had accuracy ratios sufficiently high enough to beat averaging. For them, Inner-MCS heuristic could not beat Average. In future research, it will be important to determine which internal (i.e.; personality, expertise) and external (e.g.; content domains) factors predict confidence accuracy ratios. Indeed, our simulation results suggest that

---

[16] We calculated this by comparing each participants' MAD_1 and MAD_2 values.

working memory capacity could be one important factor: in a regression analysis conducted on our simulation, we found that agents with higher working memory spans had smaller confidence based accuracy ratios than those with lower working memory spans (95% HDI: -0.72, -0.56). This result parallels that of another simulation and empirical study that found that smaller working memory spans are associated with larger averaging gains in the inner-crowd (Hourihan & Benjamin, 2010). However, our results are speculative and should be confirmed in a future behavioral study.

Thus, we are left with an important question: After increasing one's estimate diversity using dialectical bootstrapping (Herzog & Hertwig, 2009), should one average or choose? The answer depends on the statistical environment you find yourself (Fraundorf & Benjamin, 2014; Herzog & Hertwig, 2014b; Soll & Larrick, 2009). Do you believe that your high-confidence estimates are likely to be substantially more accurate than your low-confidence estimates? Do you believe the rate at which your estimates bracket the true value are low? If the answer is "yes" to both of these questions, then you should probably choose your high- confidence estimate. If your answers to the questions are "no" and "no", you should probably take the average. Either way, dialectical bootstrapping is likely to improve your estimates.

Inner-MCS versus between-MCS

In this paper we have focused on how an individual can benefit from high- confidence choosing within one mind using Inner-MCS, but how do the results compare to the performance of high confidence choosing between separate individuals (Koriat, 2012b)? Pairs of judges can benefit from using the MCS heuristic (which we label "Between-MCS"), wherein judges exchange estimates, and choose the estimate of the more confident judge (Koriat, 2012b). However, before applying the Between-MCS heuristic, one must first normalize each

judge's confidence estimates (Koriat, 2012b). Normalizing confidence judgments is data intensive because it requires knowledge of the mean and standard deviation of each judge's confidence ratings. When applying Inner-MCS, this normalization procedure is unnecessary because normalization does not change the ordinal pattern of a person's confidence judgments[17]. If we assume that people do not normalize confidence judgments, how does Inner-MCS compare to Between-MCS? To test this, we simulated the performance of the Between-MCS heuristic (without normalization) using our study data and compared it to Inner-MCS (see Appendix B for details).

Accuracy gains were generally higher for Between-MCS than for Inner-MCS. Across conditions, the median participant stood to have a 53.52% decrease in MAD relative to their first estimates using Between-MCS and 18.03% decrease by using Inner-MCS. However, we did find an effect of experimental condition on the difference in Between-MCS and Inner-MCS. The median difference in improvement between Inner-MCS and Between-MCS was smaller for the consider-other-exemplars condition than the control condition (% MAD change of 4.76% vs. 30.19%). Moreover, a full 41% of participants in the dialectical-consider other exemplars condition had *lower* MAD values by using Inner-MCS than Between-MCS. In other words, 41% of participants given consider-other-exemplars instructions would have performed better by themselves than by conferring with another random participant. In the control condition, this percentage dropped to 27%.

In summary, when applying MCS, while most participants would perform better by consulting another random participant than by generating an inner-crowd, dialectical instructions dramatically narrow the gap between high-confidence choosing gains in the inner-

[17] Applying a z-score transformation to a single person's confidence estimates will not change their order.

crowd and two separate individuals. These results parallel those of Herzog and Hertwig (2009) who found the same effect for dialectical instructions on averaging gains.

Conclusion

Previous research on how people manage both their inner-crowd and external advice suggested that people do not average as much as they should (Fraundorf & Benjamin, 2014; Herzog & Hertwig, 2014b; Müller-Trede, 2011; Larrick et al., 2012). In the current study, we simultaneously modeled estimation and confidence judgments on a real-world dataset and found that people can outperform averaging by choosing their most confident estimates (i.e.; an Inner-MCS heuristic, Koriat, 2012b). Further, we found that dialectical bootstrapping can increase the benefits associated with high-confidence choosing.

References

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., ... & Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*(2), 130-147. doi:10.1037/1076-898X.6.2.130

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412. doi:10.1016/j.jml.2007.12.005

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y. F., Roepstorff, A., et al. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition, 26*, 13–23. doi:10.1016/j.concog.2014.02.002

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081-1085. doi:10.1126/science.1185718

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, *49*(2), 188-207. doi:10.1016/0749-5978(91)90048-X

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*(2), 127–151. doi:10.1016/j.obhdp.2006.07.001

Budescu, D. V., Rantilla, A. K., Yu, H. T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, *90*(1), 178-194. doi:10.1016/S0749-5978(02)00516-2

Christensen-Szalanski, J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 928-935. doi:10.1037//0096-1523.7.4.928

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, *7*(1), 25-47. doi:xxx

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87-185. doi:10.1017/S0140525X01003922

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision, 1*, 79–101. doi:xxx

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519-527. doi:10.1037/0033-295X.101.3.519

Fiedler, K., & Juslin, P. (Eds.). (2006). *Information sampling and adaptive cognition*. Cambridge University Press.

Fraundorf, S. H., & Benjamin, A. S. (2014). Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language*, *71*(1), 17-38. doi: 10.1016/j.jml.2013.10.002

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528. doi:10.1037/0033-295X.98.4.506

Glaser, M., Langer, T., & Weber, M. (2013). True overconfidence in interval estimates: Evidence based on a new measure of miscalibration. *Journal of Behavioral Decision Making, 26*, 405–417. doi:10.1002/bdm.1773

Griffin, D., & Brenner, L. A. (2004). Perspectives on probability judgment calibration. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 177–199). Oxford, England: Blackwell.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software, 33*, 1–22. Retrieved from http://www.jstatsoft.org/v33/i02/paper. doi:xxx

Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes*, *103*(2), 277-290. doi:10.1016/j.obhdp.2007.01.003

Hansson, P., Juslin, P., & Winman, A. (2008). The role of short-term memory capacity and task experience for overconfidence in judgment under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1027–1042. doi:10.1037/a0012638

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*, 231-237. doi:10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2014a). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*. doi:10.1016/j.tics.2014.06.009

Herzog, S. M., & Hertwig, R. (2014b). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 218-232. doi:10.1037/a0034054

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1-55. doi:10.1016/0010-0285(92)90002-J

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 1068-1074. doi:10.1037/a0019694

Johnson, J. G. & Raab, M. (2003). Take the first: Option-generation and resulting choices. *Organizational Behavior and Human Decision Processes*, *91*(2), 215-229. doi:10.1016/S0749-5978(03)00027-X

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132*, 133–156. doi:10.1037/0096-3445.132.1.133

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114*(3), 678–703. doi:10.1037/0033-295X.114.3.678

Keren, G. (1997). On the calibration of probability judgments: Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making, 10*, 269–278. doi:10.1002/(SICI)1099-0771(199709)10:3<269::AID-BDM281>3.0.CO;2-L

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216-247. doi:10.1006/obhd.1999.2847

Koriat, A. (1980). Reasons for Confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 1–12. doi:10.1037//0278-7393.6.2.107

Koriat, A. (2012a). The self-consistency model of subjective confidence. *Psychological Review, 119*, 80–113. doi:10.1037/a0025648

Koriat, A. (2012b). When are two heads better than one and why? *Science, 336*, 360–362. doi:10.1126/science.1216549

Kruger, J., Wirtz, D., & Miller, D. T. (2005). Counterfactual thinking and the first instinct fallacy. *Journal of Personality and Social Psychology, 88*, 725–735. doi:10.1037/0022-3514.88.5.725

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299-312. doi:10.1177/1745691611406925

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52*, 111–127. doi:10.1287/mnsc.1050.0459

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers in social psychology: Social judgment and decision making* (pp. 227–242). New York, NY: Psychology Press.

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin, 114*(1), 162–173. doi:xxx

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Decision Processes, 20*(2), 159–183. doi:10.1016/0030-5073(77)90001-0

Lindskog, M., & Winman, A. (2014). Are all data created equal? Exploring some boundary conditions for a lazy intuitive statistician. *PLoS ONE, 9*, e97686. doi:10.1371/journal.pone.0097686

Lindskog, M., Winman, A., & Juslin, P. (2013a). Calculate or wait: Is man an eager or a lazy intuitive statistician? *Journal of Cognitive Psychology, 25*, 994–1014. doi:10.1080/20445911.2013.841170

Lindskog, M., Winman, A., & Juslin, P. (2013b). Naïve point estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 782–800. doi:doi:10.1037/a0029670

Mannes, A. E., Soll, J. B., & Larrick, R. P. (in press). The wisdom of select crowds. *Journal of Personality and Social Psychology*.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1-23. doi:10.3758/s13428-011-0124-6

Merkle, E. C., Van Zandt, T., & Sieck, W. (2008a). Error in confidence judgments. *Journal of Behavioral Decision Making, 21*, 453–456. doi:10.1002/bdm.605

Merkle, E. C., Sieck, W., & Van Zandt, T. (2008b). Response error and processing biases in confidence judgment. *Journal of Behavioral Decision Making, 21*, 428–448. doi:10.1002/bdm.597

Miller, G. A., (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63(2)*, 81-97. doi:10.1037//0033-295X.101.2.343

Moore, D. A., Tenney, E. R., & Haran, U. (in press). Overprecision in judgment. In G. Wu & G. Keren (Eds.), *Handbook of judgment and decision making*. New York, NY: Wiley.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*, 502–517. doi:10.1037/0033-295X.115.2.502

Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social influence and the collective dynamics of opinion formation. *PloS One*, *8*(11), e78433. doi:10.1371/journal.pone.0078433

Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. *Judgment and Decision Making*, *6*(4), 283-294. doi:xxx

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

Olsson, H., Juslin, P., & Winman, A. (2008). The role of random error in confidence judgment: Reply to Merkle, Sieck, and Van Zandt (2008). *Journal of Behavioral Decision Making, 21*, 449–452. doi:10.1002/bdm.604

Peterson, D. K., & Pitz, G. F. (1986). Effects of amount of information on predictions of uncertain quantities, *Acta Psychologica, 61*(3), 229–241. doi:10.1016/0001-6918(86)90083-1

Phillips, N. D., Herzog, S. M., & Hertwig, R. (2014). *How The Inner Crowd Can Help Non-Bayesians Become More Bayesian*. Manuscript in preparation.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic, *Journal of Behavioral Decision Making, 17*(1), 39–57. doi:10.1002/bdm.460

Rauhut, H., & Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology, 55*, 191–197. doi:10.1016/j.jmp.2010.10.002

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, *84*(2), 288-307. doi:10.1006/obhd.2000.2926

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 299–314. doi:10.1037/0278-7393.30.2.299

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 780–805. doi:10.1037/a0015145

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Garden City, NY: Doubleday.

von Helversen, B., & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 867–889. doi:10.1037/a0015501

Vul, E., & Pashler, H. (2008). Measuring the Crowd Within: Probabilistic Representations Within Individuals. *Psychological Science*, *19*(7), 645–647. doi:10.1111/j.1467-9280.2008.02136.x

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review*, *101*(3), 490-504. doi:10.1037/0033-295X.101.3.490

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*, 779–804. doi:10.3758/BF03194105

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*, *66*(336), 675-685. doi:10.2307/2284212

Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: How to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 1167–1175. doi:10.1037/0278-7393.30.6.1167

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes, 69*, 237–249. doi:10.1006/obhd.1997.2685

Yaniv, I. (2004a). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93*, 1–13. doi:10.1016/j.obhdp.2003.08.002

Yaniv, I. (2004b). The benefit of additional opinions. *Current Directions In Psychological Science, 13*, 75–78. doi:10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, *10*(1), 21-32. doi:10.1002/(SICI)1099-0771(199703)10:1<21::AID-BDM243>3.0.CO;2-G

Yates, J. F. (1990). *Judgment and decision making*. Prentice-Hall, Inc.

Appendix A: Simulation Details

The following is a description of the agent-based simulation. Full code are available in our online supplementary materials.

Modeling agents' knowledge base and working memory (WM) capacity

We began by specifying long-term memory (LTM) and working memory (WM) parameters for each agent. County knowledge in long-term memory was operationalized as the number of exemplars from the total county database that was stored in an agent's long-term memory. Specifically we assigned to each agent a knowledge-base size drawn from a normal distribution with mean 100 and standard deviation 20. Further, each agent was assigned a working memory capacity randomly drawn from the set {3, 4, 5} (Cowan, 2001).

We did not assume that agents would have perfect knowledge of county populations in long-term memory. Instead, we assigned a population *bias* and *random error* to each agent. Agents with positive population biases had population values that were too high stored in LTM, while those with negative population biases had values that were too low in LTM. Agents with high random errors hard larger noise in population LTM values, while those with low random errors had less noise. We assigned each agent a population bias value drawn from a normal distribution with mean 0 and standard deviation 1000, and a population random error value drawn from a normal distribution (truncated to be greater than 0) with mean 1,000 and standard deviation 300.

Additionally, we did not assume that all agents had perfect memory of population cue values. Instead, we assigned each agent a *cue noise* value between 0 and .5. Agents with a cue noise value of 0 had perfect knowledge of all county cues, while agents with a cue noise value of .5 had incorrect knowledge of half of all county cue values. Thus, agents with a cue noise

value of .5 could only remember county cue values at chance level. We assigned each agent a cue noise value randomly drawn from the set {0, .1, .2, .3, .4, .5}.

To summarize, the six agent-level parameters we varied in the simulation are presented in Table A1:

Agent level parameters	Parent Distribution
Phase 2 Estimation procedure (Condition)	{Repeated, New-Cue, New-Exemplar}
Knowledge in LTM (N.LTM)	Normal(mu = 100, sd = 20)
WM Capacity (N.WM)	{3, 4, 5}
Population bias (BIAS)	Normal(mu = 0, sd = 1000)
Population random error (RANDOM)	Normal(mu = 1000, sd = 300)
Cue noise (C.NOISE)	{0, .1, .2, .3, .4, .5}

Table A1: Six agent level parameters varied in the simulation.

We generated 5,000 agents and assigned to each of them a random parameter value for each of the six parameters in Table A1. Next, we had each agent generate phase 1 and phase 2 best estimates and confidence intervals for all 16 cue profiles in Table 1 using Steps 1 – 6 outlined below:

Generating first estimates

Agents gave best estimates and confidence intervals for each of the 16 stimuli in Table 1. For each stimuli, agents were given the stimuli cue profile and followed the following six sequential steps. Note that we use the index i to represent the ith agent:

- *Step 1 (Phase 1 - Cue Selection)*: The agent selected one of the four cues at random and looked up its value. The cue and its value represents the agent's memory *probe.* For example, when presented with the cue-profile [0, 0, 1, 1], an agent could use "Bachelor's degree cue is LOW" as the probe.
- *Step 2 (Phase 1 - SSD generation)*: The agent then searched its long-term memory LTM_i for WM_i counties in long-term memory that matched the probe cue, where WM_i was the agent's working memory capacity. The set of populations for those counties represented the agent's SSD. For example, an agent *i* with a working memory capacity $WM_i = 4$ and the probe cue "Bachelor's degree is LOW" would select randomly four counties from its LTM_i that were tagged as having a low bachelor's degree. The four population values tagged to those four counties would constitute the agent's phase 1 SSD.
- *Step 3 (Phase 1 – Responses)*: The agent generated its best estimate for the question as the median population in its phase 1 SSD generated in Step 2, and its confidence interval as the minimum and maximum values in its SSD. For example, an agent with a phase 1 SSD of {5010, 10450, 8760, 12332} would give a phase 1 best estimate of 9,605 and a phase 1 confidence interval of {5010, 12332}.

Generating second estimates

After generating its best estimates and confidence intervals for phase 1, agents proceeded to generate their phase 2 responses to the question. Unlike phase 1, we implemented two different *phase 2 estimation procedures* that agents could generate their phase 2 estimates. We implemented these different procedures to model the effects of different levels of estimate diversity as estimate diversity is known to drive much of the inner-crowd effect (Herzog &

Hertwig, 2014a). Agents using a "Repeated" procedure generated new exemplars in phase 2 using the same probe cue as phase 1. Because these agents did not consider a new cue relative to phase 1, this procedure was meant to generate a relatively low level of exemplar diversity between phases 1 and 2 and could represent the control conditions from previous studies on the inner-crowd (see Herzog & Hertwig, 2009, 2014b).

We implemented a estimation procedures called "New-Cue" that were designed to increase agents' estimate diversity. Conceptually, this procedure could represent how *dialectical* participants may increase estimate diversity in their inner-crowd. Agents using the "New-Cue" procedure selected a new probe-cue that was different from the one they used for that question in Phase 1 and generated a new set of exemplars from that new probe-cue

Next, agents generated estimates for phase 2 in the following three steps:

- *Step 4 (Phase 2 –Cue Selection)*: The probe-cue agents used in phase 2 depended on their estimation condition:
    1. Repeated: Use the same cue as in phase 1
    2. New-Cue: Select a new random probe cue that is different from the Phase 1 probe-cue.
- *Step 5: (Phase 2 – SSD Generation)*: The Phase 2 SSD generation procedure agents used depended on their estimation condition:
    1. Repeated & New-Cue – Select WM_i random exemplars from LTM that match the probe-cue selected in Step 4. For agents using the repeated and new-cue procedure, these exemplars need not be different from those selected in phase 2.

The set of WM_i exemplars generated in Step 5 constituted the agent's Phase 2 SSD.

- *Step 6: (Phase 2 – Responses)*: Generate best estimates and confidence intervals using the median and range of the set of population values in Step 5 (as in step 3).

Appendix B: Between-MCS vs. Inner-MCS

To test whether the MCS algorithm succeeds in our estimation task, we simulated the expected estimation accuracy of inner-MCS with group-MCS as follows. First, we created virtual pairs of all participants *i* and *j* within each condition[18]. We looked at each participant's estimate in both phases 1 and 2, giving us 4 total estimates for each stimulus. Next, for each stimulus k, we took the estimate with the highest confidence as the group-MCS estimate for participants i and j. We calculated the absolute deviation of this estimate and the stimuli criterion. Next we took the mean absolute deviation of group-MCS estimates across all stimuli to calculate a group-MCS mean absolute deviation for participants i and j. For each participant i, we then calculated the median group-MCS MAD value across all possible pairings. We compared this median group-MCS MAD value to each participant's first estimate MAD values to calculate a % improvement due to the two heads are better than one effect. Results are presented in Table C1:

	Inner Crowd Median / IQR	Between Median / IQR	Between – Inner Median / IQR
Control	8.90% [0.00%, 37.03%]	52.53% [27.13%, 71.70%]	30.19% [2.04%, 51.05%]
CTO	15.64% [0.87%, 47.07%]	54.67% [31.90%, 70.36%]	25.80% [1.59%, 47.57%]
COE	31.38% [0.02%, 60.29%	57.18% [18.96%, 75.63%]	4.76% [-11.4%, 31.80%]
All	18.03% [0.03%, 49.72%]	53.52% [24.58%, 72.05%]	21.89% [-3.24%, 45.64%]

Table B1: Percent decrease in MAD by using Inner-MCS versus Between-MCS.

[18] In Koriat's (2012b) analysis, Koriat only paired participants with similar accuracy levels and z-transformed participants' confidence levels before applying the MCS algorithm. We elected to pair all participants and did not transform confidence levels prior to applying the algorithm. We did this because in real world advice-taking, one receives advice from people with varying levels of accuracy and without z-transformed confidence ratings. We did run alternative simulations by matching participants with similar accuracy levels and found very similar effects.

Appendix C: Study Stimuli

The following is a screenshot of the an example stimuli from the study

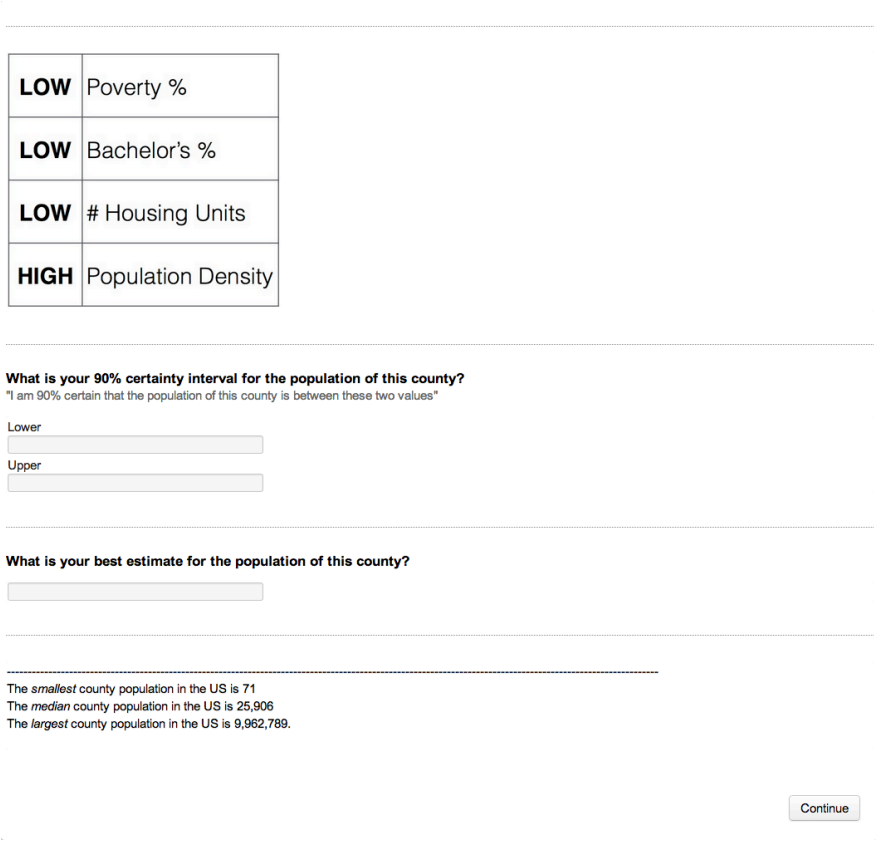


Figure C1: Screenshot of the experimental stimuli

Curriculum Vitae

CONTACT

Nathaniel David Phillips
Kolonnenstraße 58/59
10829 Berlin
Germany

email: nathaniel.phillips.is@gmail.com
website: www.nathanieldphillips.info

EDUCATION

2005 B.A., Mathematics, Grinnell College, Grinnell, IA
2010 M.S., Experimental Psychology, Ohio University, Athens, OH

EMPLOYMENT

2005–2006 Marketing Statistician, Musician's Friend, Medford, OR
2012–2014 Predoctoral Research Fellow, Max Planck Institute for Human Development, Berlin, Germany

PUBLICATIONS

Phillips, N. D., Hertwig, R., Kareev, Y., & Avrahami, J. (2014). Rivals in the dark: How competition influences search in decisions under uncertainty. *Cognition, 133(1),* 104-119.

González Vallejo, C., Cheng, J., Phillips, N. D., Chimeli, J., Bellezza, F., Harman, J., & Lindberg, M. J. (2013). Early positive information impacts final evaluations: No deliberation−without−attention effect and a test of a dynamic judgment model. *Journal of Behavioral Decision Making, 27(3)*, 209-225.

González-Vallejo, C., & Phillips, N. D. (2010). Predicting soccer matches: A reassessment of the benefit of unconscious thinking. *Judgment and Decision Making, 5(3),* 200-206.

Lassiter, G. D., Lindberg, M. J, Gonzalez-Vallejo, C., Bellezza, F. S., & Phillips, N. D. (2009). The deliberation-without-attention effect: Evidence for an artifactual interpretation. *Psychological Science, 20(6)*, 671-675.

MANUSCRIPTS

Phillips, N. D., Hertwig, R., & Kareev, Y. (in prep.). *Exploring the unknown: Adaptive information search in decisions from experience*.

Phillips, N. D., Herzog, S., & Hertwig, R. (in prep.). *Harnessing the Bayesian crowd within: How contradicting yourself improves Bayesian reasoning judgments*.

Phillips, N. D., Herzog, S., Kämmer, J., & Hertwig, R. (in prep.). *Confidence and Dialectical Bootstrapping Facilitates Choosing in The Inner-Crowd*.

CONFERENCE PRESENTATIONS

Phillips, N. D., Hertwig, R., Kareev, Y., & Avrahami, J. (2013). *Competing in the dark: How competition affects information search and decisions under uncertainty.* Paper presented at the annual meeting of the Association for Mathematical Psychology, Potsdam, Germany.

Phillips, N. D., Hertwig, R., Kareev, Y., & Avrahami, J. (2013). *Competing in the dark: How competition affects information search and decisions under uncertainty.* Paper presented at the JDM Workshop for Young Researchers, Berlin, Germany.

Phillips, N. D., Herzog, S., & Hertwig, R. (2013). *Modeling Bayesian inference judgments.* Paper presented at the Tagung experimentell arbeitender Psychologen, Vienna, Austria.

Phillips, N. D., Herzog, S., & Hertwig, R. (2012). *Contradicting yourself makes you more Bayesian: Averaging non-Bayesian judgments with dialectical bootstrapping improves judgments.* Paper presented at the annual meeting of the Society for Judgment and Decision Making, Minneapolis, MN.

Phillips, N. D., Herzog, S., & Hertwig, R. (2011). *Averaging multiple intuitive strategies improves performance in Bayesian estimation tasks: How groups can act (more) Bayesian without being Bayesian.* Poster presented at the annual meeting of the Society for Judgment and Decision Making, Seattle, WA.

Phillips, N. D., Hertwig, R., & Kareev, Y. (2011). *Exploring the unknown: Sample allocation in decisions from experience.* Paper presented at the Subjective Probability, Utility, and Decision Making conference, Kingston-upon-Thames, UK.

Phillips, N. D., Hertwig, R., & Kareev, Y. (2011). *Modeling search behavior in decisions-from-experience.* Paper presented at JDM Young Researchers Conference, Bonn, Germany.

Phillips, N. D., & Gonzalez-Vallejo, C. (2010). *Modeling the joint effects of experiences and descriptions on impressions and choices.* Paper presented at the annual meeting of the Society for Judgment and Decision Making, St. Louis, MO.

Phillips, N. D. & Gonzalez-Vallejo, C. (2010). *Incorporating recommendations and experiences: A mathematical model of impression formation.* Poster presented at the annual meeting of the Association for Psychological Science, Boston, MA.

Phillips, N. D., & Gonzalez-Vallejo, C. (2009). *Modeling the joint effects of description and experience on impression formation and decision-making.* Poster presented at the Summer Institute on Bounded Rationality, Berlin, Germany.

Lassiter, G. D., Lindberg, M. J., González-Vallejo, C., Bellezza, F., & Phillips, N. D. (2009). *Why inattention to complex decisions yields "optimal" judgment: It's not what you (unconsciously) think.* Paper presented at the Society for Personality and Social Psychology, Tampa, FL

Phillips, N. D., & Gonzalez-Vallejo, C. (2008). *Modeling the joint effects of description and experience on impression formation and decision-making.* Poster presented at the meeting of the Society for Judgment and Decision Making, Chicago, IL.

Phillips, N. D. (2008). *Decisions from description and experience: Modeling impression formation in the red bean task.* Paper presented at Every Area Talks, Ohio University.

Phillips, N. D., Gonzalez-Vallejo, C., Bellezza, F. S., Chimeli, J., Harman, J., Lassiter, G. D., & Lindberg, M. J. (2007). *Testing unconscious thought: Is distraction really a panacea for difficult decisions?* Paper presented at the meeting of the Society for Judgment and Decision Making, Long Beach, CA.

Phillips, N. D. (2006). *Order effects in online impression formation.* Paper presented at Every Area Talks, Ohio University.

WORKSHOPS AND SUMMER SCHOOLS ATTENDED

Bayesian Modeling for Cognitive Science: A WinBUGS Workshop. Amsterdam, Netherlands, August 12-16, 2013.

JDM Workshop for Young Researchers. Max Planck Institute for Human Development, Berlin, Germany, July 17-19, 2013.

Summer School on Cognitive Modeling. Bergün, Switzerland, June to July 2012.

Fate of the Memory Trace: Summer school of the European Campus of Excellence. Ruhr-University Bochum, September 2011.

Summer Institute on Bounded Rationality. Max Planck Institute, Berlin, Germany, July 15-22, 2009.

JDM Workshop for Young Researchers. Max Planck Institute for Collective Goods, Bonn, Germany, August 5-8, 2011.