

A Sharp Test of the Portability of Expertise*

Etan A. Green^{†1}, Justin M. Rao², and David Rothschild²

¹The Wharton School, University of Pennsylvania

²Microsoft Research

September 5, 2016

Abstract

As-if justifications of rationality contend that economic actors approximate optimal behavior through heuristics. We evaluate this contention by observing experts perform a task that is logically isomorphic to—but contextually distinct from—a familiar task in which they are skilled. We find that performance plummets when contextual cues disappear, implying that the expertise we observe on the familiar task is more heuristic than conceptual and does not travel far. Our results provide support for as-if justifications: with experience in a domain, actors develop heuristics that are adaptive in that setting. Our results also delineate bounds on such justifications, showing that heuristics can fail even in contexts with the same logical structure. This observation entails a normative implication for experimental design. If economic actors approximate rationality through context-dependent heuristics, then studies which abstract away contextual cues bias their findings against standard theories of rationality.

*We thank seminar participants at Cornell, Microsoft Research, and the 2016 Behavioral Decision Research in Management conference, in particular Ted O’Donoghue, Jesse Shapiro, Glen Weyl, and Jenn Wortman Vaughan. We also thank Jim Andreoni, Hengchen Dai, Dorothy Kronick, David Reiley, Joe Simmons, and Charlie Sprenger for helpful comments and suggestions on previous drafts. We are grateful to Henry Abbott, Kevin Arnovitz, and Royce Webb for inviting us to work with them on their polling, and to Microsoft Research for generous financial support.

[†]Corresponding author: etangr@wharton.upenn.edu

1 Introduction

In many markets, individually rational behavior requires considerable sophistication. For example, the optimal bid in a “simple” sealed-bid, first-price auction with symmetric, risk-averse players is a non-additively separable function of the private valuation, the number of players, and the risk tolerance (cf. [Harrison, 1989](#)). In cases like these, it is unlikely that market actors are perfect logicians who derive mathematical formulas to guide behavior. Rather, experienced actors are thought to develop heuristics, or mental shortcuts, that allow them to behave *as if* they were logicians.¹ This “as if” justification provides a convenient way to bridge abstract models and real-world behavior without descending into the messy details of human cognition. With as-if actors, the goodness of a model depends on the accuracy of its predictions, rather than on the realism of its behavioral assumptions ([Friedman, 1953](#)).

One issue with the notion that heuristics undergird as-if rational behavior is that these mental shortcuts are often context dependent. Whereas the logician discerns logical rules, the boundedly rational actor develops heuristics through experience. As a result, mental shortcuts honed in one context may leave actors ill-equipped in another. [Neelin, Sonnenschein and Spiegel \(1988\)](#) illustrate this concern through a sequential bargaining experiment. They find that with practice, novice subjects learn to approximate the subgame-perfect first offer in the initial two-round game, but they continue to make similar first offers when the number of rounds increases, even though doing so is off the equilibrium path.² As this work demonstrates, lessons learned from experience, at least by novices, may not be robust to even minor changes in the formulation of the problem.

We test whether small changes in the contextual representation of a task affect the

¹For instance, [Harrison and List \(2004\)](#) hypothesize that “naturally occurring markets are efficient because [experienced] traders use heuristics to avoid the inferential error that underlies the winner’s curse.”

²[Kagel and Levin \(1986\)](#) document a similar pattern in an experimental study of repeated common value auctions. They find that laboratory subjects learn to avoid the winner’s curse by shading their bids, but when the number of bidders increases, players shade insufficiently and experience the winner’s curse again.

performance of experts. A handful of studies, which we review in Section 2, document the performance of non-standard subjects on conventional laboratory games that arguably resemble tasks in which they are experienced. But as many of these papers note, experience does not always imply expertise, and games played in the lab often differ meaningfully from tasks performed in the field. Perhaps as a consequence, the findings are equivocal. In this paper, we conduct a sharper test of the portability of expertise. First, we demonstrate that the experts we study possess expertise in a familiar task. We then reduce this task to its logical structure by removing contextual cues, and we compare the performance of the same experts on the familiar and unfamiliar versions of this logically isomorphic task.

Our main finding is that performance declines sharply when contextual cues disappear, implying that the expertise we observe in the familiar domain is more heuristic than conceptual and does not travel far. In addition, we find that the experts improve with repetition in the unfamiliar setting, but in a manner that again suggests heuristic learning. Our results provide support for as-if justifications: with experience in a domain, actors develop heuristics that are adaptive in that setting. These findings also delineate bounds on such justifications, showing that heuristics can fail even in contexts with the same logical structure. This observation entails a normative implication for experimental design. If economics actors approximate rationality through context-dependent heuristics, then studies which abstract away contextual cues bias their findings against standard theories of rationality by making it difficult for subjects to apply their everyday expertise.

The experts we study are basketball writers, analysts, commentators, and executives at the ESPN sports network who make probabilistic predictions about 1) game-by-game outcomes for playoff series in the National Basketball Association, and 2) outcomes of those series. Formally, the task is similar to predicting outcomes for both individual events and an encompassing set—for instance, predicting the probability of default on individual mortgages in a mortgage-backed security and the probability that the security will be downgraded, or

predicting state-by-state electoral odds for a presidential candidate and the probability that she wins the election. We are interested in whether an expert’s predictions for the individual events are consistent with her predictions for the encompassing set—i.e., whether an expert’s predictions reflect the logic that links game outcomes and series outcomes.

Previous research has shown that individuals often report probabilities for sets of outcomes that cannot be rationalized by the probabilities of constituent events (Grether, 1980, 1992; Tversky and Kahneman, 1983; Charness, Karni and Levin, 2010). In contrast, we find a high degree of consistency between the individual and aggregate predictions made by these experts in their domain of expertise, both in absolute terms and relative to predictions made by novices. For instance, the median deviation between the stated probability of a series victory and the value implied by the game-by-game probabilities is just 5 percentage points for experts, compared to 10 percentage points for novices.

To measure the portability of this expertise in making consistent predictions, the experts were asked, months later, to perform a logically isomorphic but unfamiliar task “designed to improve ESPN’s NBA playoffs forecasting,” as the invitation from an ESPN editor read. The task described sequences of ordered jars, each sequence with 7 jars, and each jar with black and red marbles in specified quantities and with 100 marbles total. For the first 4 sequences, each expert observed marble proportions that were matched exactly to the 7 game-by-game probabilities that he or she had reported for an NBA playoff series. The fifth and final sequence, shown to all respondents, contained 95 black marbles (and 5 red marbles) in each jar. For each sequence, subjects were told that a single marble would be drawn from each jar in order. They were then asked to evaluate 1) the probability that at least 4 of the 7 drawn marbles would be black, and 2) the jar from which the fourth marble of the same color would most likely be drawn—i.e., summary statistics that correspond exactly to predictions made for NBA playoff series.

Despite this isomorphism, the experts exhibit significantly less consistency on the unfa-

miliar marbles exercise than on the familiar forecasting task, both in economic and statistical terms. The average level of inconsistency is twice as high in the unfamiliar domain as in the familiar one. Moreover, consistency is only weakly correlated across domains, suggesting that the experts approached the jars task as if facing a new problem, rather than one they had solved before. It is difficult to attribute this result to differences in motivation between the contexts, as the experts take their time on the marbles task and improve with repetition, suggesting attentiveness.

The contexts differ principally in their representations—games and teams in the NBA versus jars and marbles in the abstract domain. Research in cognitive science shows that individuals have difficulty transferring expertise from specific to abstract domains (Loewenstein, 1999)—e.g., solving an algebra problem after learning to solve a conceptually identical problem in physics (Bassok and Holyoak, 1989). The Wason selection task offers a complementary insight: when problems are rendered in familiar, rather than abstract terms, even inexperienced subjects prove more capable.³ Analogously, NBA experts have difficulty applying skills learned from games and teams to an abstract, though formally equivalent, task involving jars and marbles.

A reasonable interpretation is that the internal consistency displayed by the experts in the field is implicitly learned. While many of the experts we study have statistical training, it is unlikely that many explicitly evaluate the probabilistic logic of an NBA playoff series when making their predictions. Instead, NBA experts may possess an intuitive sense for game and series outcomes and it could be that by observing those outcomes, these beliefs become consistent. Under this interpretation, the inconsistency in the abstract domain reflects the

³In its abstract form, the Wason task presents subjects with four cards on a table marked ‘A’, ‘K’, ‘2’, and ‘7’, respectively. Subjects are told that each card has a letter on one side and a number on the other. A logical rule is stated: If there is an ‘A’ on one side, then there is a ‘2’ on the other side. Subjects are then asked to turn over those cards, and only those cards, that determine whether the rule is violated. In this rendering, few subjects recognize the “if P , then not Q ” logic and turn over ‘A’ and ‘7’. However, many more choose correctly when the logic is framed in familiar terms, such as “If a player wins a game, then he will have to treat the others to a round of drinks.” (Gigerenzer and Hug, 1992)

experts' inability to survey their experience without the help of contextual cues.

By necessity, the contexts we examine differ in two ways unrelated to abstraction. First, the probabilities of individual events were elicited from experts in the familiar domain but presented to experts in the unfamiliar setting. As a result, we cannot rule out the possibility that generating probabilities for individual events enhances consistency relative to evaluating probabilities that one previously provided. Nonetheless, this interpretation is consistent with our main conclusion that the expertise we document is not robust to even minor changes in the structure of the problem. Second, the marble draws in the unfamiliar domain are explicitly independent, whereas the experts may believe that game-by-game outcomes are sequentially dependent (though as we show, large deviations from independence are inconsistent with their series predictions). We think it is unlikely that the experts possess a logical mastery of sequentially dependent events but are feeble at assessing comparatively simpler distributions of sequentially *independent* events. Nevertheless, such an interpretation is likewise consistent with our main conclusion that expertise does not travel far.⁴

Our second set of results measures improvement in the unfamiliar environment by comparing performance across the 4 matched sequences—which are randomly ordered—finding that consistency increases with repetition. As with the experts' skill in making consistent NBA predictions, this improvement likely does not reflect a conceptual understanding of how to calculate summary statistics for complex distributions. For the fifth sequence, in which all jars have 95 black marbles, the majority of experts report a series probability of 95%, an apparently intuitive response that contrasts with the correct answer of greater than 99.98%. Given this conceptual failing, the observed improvement indicates that the experts devise a heuristic for the unfamiliar task.

⁴A third difference is that likelihoods for individual events in the familiar domain are represented as probabilities, whereas those likelihoods are represented as frequencies in the unfamiliar domain. However, this difference likely biases our results in favor of greater consistency in the unfamiliar domain, as various biases in judgment have been shown to attenuate or disappear when likelihoods are discussed in terms of frequencies rather than probabilities (Cosmides and Tooby, 1996).

These results speak to a debate about the generalizability of findings from the laboratory to field contexts of interest (e.g. [Levitt and List, 2007a,b, 2008](#); [Falk and Heckman, 2009](#); [Camerer, 2011](#); [Al-Ubaydli and List, 2013](#)). In particular, they raise questions about the extent to which behavior by experts in abstract laboratory settings (e.g. [Haigh and List, 2005](#); [List and Haigh, 2005](#)) predicts behavior by those actors in more familiar contexts. Our results also provide evidence pertaining to the influence of abstraction on conventional laboratory subjects, who may find the contexts in which they have experience more familiar. In contrast to many laboratory experiments in social psychology, in which elements of familiar environments are recognizable, the convention in experimental economics is to render laboratory experiments in abstract terms, in which theoretical concepts like incentives are readily identifiable, but contextual cues are not ([Hertwig and Ortmann, 2001](#)).⁵ If rationality is approximated by context-dependent heuristics, then abstraction reduces the external validity of laboratory findings and biases those findings against standard theories of rationality.

The remainder of the paper is organized as follows. Section 2 reviews studies of expert rationality outside of the domains of expertise. Section 3 describes the familiar context and documents the subjects' expertise in making internally consistent predictions. Section 4 describes the unfamiliar context and contrasts the consistency of experts across the domains. Section 5 concludes.

2 Related Literature

Three sets of studies on experts examine the portability of expertise in the field to traditional laboratory games. In each case, the fidelity between field and lab is inexact, and behavior in the lab is sometimes consistent with behavior in the field and other times not. Commonly,

⁵Of course, there are many exceptions (e.g. [Kessler and Roth, 2014](#)).

expertise is assumed rather than documented. As a result, measures in the lab are weighed against novice performance or theoretical baselines, rather than comparable measures from the field. Moreover, the laboratory games are often so abstract that subjects do not recognize the formal similarities between lab and field. In some cases, *formal* elements differ between lab and field—such that what is optimal in the field is not optimal in the lab. By contrast, our study compares the performance of the same experts across logically isomorphic settings, allowing us to more credibly measure the portability of expertise.

The first set of studies examines bidding behavior in the lab by subjects who compete in common-value auctions professionally. [Dyer et al. \(1989\)](#) show that construction executives who have experience bidding on contracts with uncertain costs insufficiently shade their bids in the lab as routinely as inexperienced undergraduates, leading the authors to speculate that “the executives have learned a set of situation specific rules of thumb which permit them to avoid the winner’s curse in the field but which could not be applied in the lab.” Interviews with construction executives offer another explanation: they do not avoid the winner’s curse in the field ([Dyer and Kagel, 1996](#)). Insufficient shading is a feature of construction bidding because unprofitably low bids can be amended or canceled after the contract has been awarded. [Harrison and List \(2008\)](#) conduct similar experiments with participants at a sportscard show, where bidding on an unopened pack of baseball cards resembles a common-value auction. They find that dealers, who the authors argue are experienced in avoiding the winner’s curse, make lower bids than non-dealers.

The second set of studies leverages the findings of [Chiappori, Levitt and Groseclose \(2002\)](#) and [Palacios-Huerta \(2003\)](#) that in professional soccer, penalty kicks appear consistent with minimax predictions: kickers and goalies randomize directions in proportion to expected payoffs. In order to measure the portability of this expertise in randomization, [Palacios-Huerta and Volij \(2008\)](#) and [Levitt, List and Reiley \(2010\)](#) invite professional soccer players to play stylized card games with simple mixed strategy equilibria. [Palacios-Huerta and Volij](#)

(2008) find that experts' card play is consistent with minimax predictions, and considerably more so than non-experts, though Wooders (2010) arrives at the opposite conclusion from the same data. By contrast, Levitt et al. (2010) find that professional soccer players playing the same card game randomize as poorly as non-experts, noting that these subjects do not recognize similarities between the card game and the soccer field. As Kovash and Levitt (2009) observe, soccer players may not expertly play mixed strategies in the field. Tests of optimal randomization in penalty kicks are underpowered, and a failure to reject the null of optimal play does not imply optimal play.

Similarly inconsistent results are found when chess grandmasters, who are presumably skilled in backward induction, play stylized games that purport to test that skill. Whereas Palacios-Huerta and Volij (2009) find that each of the 26 grandmasters in their study play the Nash equilibrium prediction—stopping at the first node when playing the centipede game against other grandmasters—Levitt, List and Sadoff (2011) find that each of their 16 grandmasters deviate from the backward induction prediction in the centipede game, choosing to cooperate rather than play the Nash equilibrium. One concern is that the centipede game is a poor conceptual replica of chess because it allows for cooperative equilibria. Indeed, it was designed to show the brittleness of Nash equilibria when collusive but individually irrational play yields far higher returns.

A separate literature measures the degree to which preferences, rather than expertise, correlate across lab and field. In some cases, preferences in the lab predict those in the field (e.g. Bonin, Dohmen, Falk, Huffman and Sunde, 2007; Benz and Meier, 2008; Burks, Carpenter, Goette and Rustichini, 2009; Meier and Sprenger, 2010; Sutter, Kocher, Glätzle-Rützler and Trautmann, 2013), and in other cases, they do not (e.g. List, 2006; Stoop, Noussair and Van Soest, 2012). For a review, see Camerer (2011).

3 NBA playoff predictions

The ESPN sports network regularly surveys a panel of basketball writers, sportscasters, analysts, and executives. From 2013-15, an ESPN editor asked this panel to predict outcomes of NBA playoff series. During the 2014 playoffs, ESPN also surveyed readers of the TrueHoop blog on ESPN.com, asking them an identical set of questions as the experts. All respondents were told that aggregated predictions would be published on ESPN.com; no further incentives were offered. Respondents were unaware that their responses would be evaluated for internal consistency or used for any other research purpose.

NBA playoff series follow a best-of-7 format: the first team to win 4 games wins the series, and games 5, 6, and 7 are only played if neither team has won 4 games up to that point. We refer to the *series home team* as the team that plays at home for game 1, and we restrict our analyses to the common format, in which the series home team also plays at home for games 2, 5, and 7. Each ballot asked first for the probability that the series home team will win each game of the series—7 probabilities total, with those for games 5, 6, and 7 conditioned on that game being played. Respondents chose these game-by-game probabilities from 11 options: every 10 percentage points from 5% to 95%, as well as 50%. Respondents were then asked to choose the most likely series outcome of the series from among 8 mutually exclusive and completely exhaustive options—i.e., series home team in 4, 5, 6, or 7 games, or series road team in 4, 5, 6, or 7 games.⁶ We denote an outcome as the pair of games won by the series home team and series road team; for example, 4-1 implies that the series home team wins the series 4 games to 1. Beginning in 2014, respondents were also asked to report the probability that the series home team will win the series. For this question, respondents typed in a number, which we round to the nearest percent. With the

⁶The editor phrased this question using a colloquial, rather than statistical, meaning of expectations, asking how many games the respondent expected each team to win. Consistent with the editor's intent of asking for the most likely series outcome, every expert response predicts that one team will win exactly 4 games and that the other team will win an integer number of games less than 4.

exception of the game 1 probability, none of these quantities are reliably estimated by Vegas betting lines or prediction markets prior to game 1.⁷

Each series was presented on a separate ballot, and subjects chose which, if any, series to predict.⁸ Our sample comprises 165 experts, who each complete between 1 and 32 surveys (with a median of 6 and mean of 9) for a total of 1480 responses (1010 of which have series probability predictions), as well as 410 responses from 357 readers over 2 series.⁹ Although our primary interest is in internal consistency, we note that the experts appear to make more accurate predictions than the readers.¹⁰

We first evaluate the internal consistency of each response under the assumption that the reported game-by-game probabilities are sequentially independent; later, we relax this assumption. Specifically, we compute the probability of each potential series outcome—e.g., series home team winning in 5 games, 4 games to 1—by multiplying the associated game-by-game probabilities.¹¹ The implied *series probability* is the cumulative probability among series outcomes of 4-0, 4-1, 4-2, and 4-3—i.e., in which the series home team wins the series. The implied *most likely series outcome* is the series outcome with the highest probability.

We measure the internal consistency of a prediction by its absolute error. For the series probability, the absolute error is the difference between the reported and implied values; for the most likely series outcome, the absolute error is the difference between the probability of the reported most likely series outcome and the probability of the implied most likely

⁷Betting lines and prediction market securities for individual games are only traded immediately prior to that game. In some cases, markets exist for series outcomes prior to the start of the series, but these are frequently illiquid.

⁸Appendix A shows examples of the invitation email and the online ballot.

⁹This count reflects removal of 7 reader responses which report impossible most likely series outcomes—i.e., fewer than 4 wins for both teams or at least 4 wins for both teams.

¹⁰We compare error rates for the 2 series that experts and readers both evaluated, considering the mean-squared error (MSE) of probabilistic predictions for the 11 games played (5 in one series, 6 in the other). For game-by-game predictions made before the series, experts averaged a mean-squared error of 0.318, compared to 0.341 for readers ($p = 0.17$; standard errors clustered by game). We attribute these high error rates to the coincidence that both series were won by the underdog.

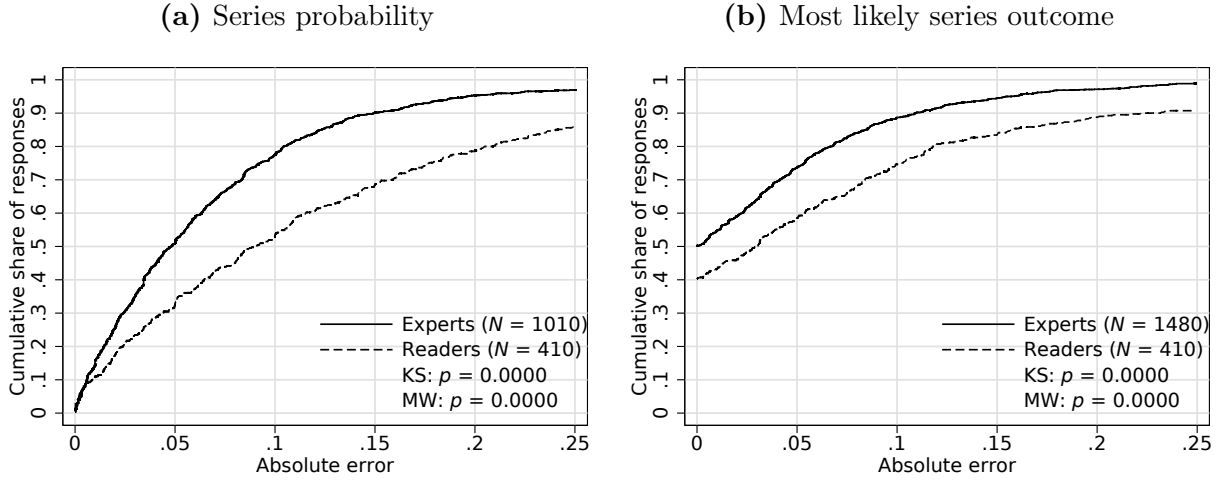
¹¹For example, $P(4-1) = p_1 \cdot p_2 \cdot p_3 \cdot (1-p_4) \cdot p_5 + p_1 \cdot p_2 \cdot (1-p_3) \cdot p_4 \cdot p_5 + p_1 \cdot (1-p_2) \cdot p_3 \cdot p_4 \cdot p_5 + (1-p_1) \cdot p_2 \cdot p_3 \cdot p_4 \cdot p_5$, where p_n is the series home team's probability of winning game n .

series outcome. For example, a commonly reported sequence of game-by-game probabilities assigns the series home team a 65% chance of winning games 1, 2, 5, and 7 (which the series home team plays at home) and a 45% chance of winning games 3, 4, and 6 (which the series home team plays on the road). Assuming sequential independence, these game-by-game probabilities imply that 1) the series home team has a 64.1% chance of winning the series, and 2) the most likely series outcome is the series home team winning 4 games to 3, which occurs with probability 20.2%. Hence, a reported series probability of, say, 60% has an error of 4.1 percentage points, and a reported most likely series outcome of 4-1, which occurs with 19.6% probability, has an error of 0.6 percentage points. Reported game-by-game probabilities typically imply series probabilities and most likely series outcome probabilities that are distant from 0 or 1, alleviating boundary concerns in our measure of absolute error.¹² All p -values for mean comparisons are from two-way tests with standard errors clustered by respondent. For distributional tests, standard errors are unclustered.

Figure 1 shows the empirical cumulative distribution function of the absolute error—i.e., the share of responses for which absolute errors are less than or equal to a given value—for the series probability (1a) and the most series likely outcome (1b). Experts exhibit high levels of internal consistency, both in absolute terms and relative to readers. In Figure 1a, 52% of expert responses report a series probability within 5 percentage points of its implied value, compared to 34% of readers; 78% of expert responses are within 10 percentage points, compared to 54% of readers. The mean error for experts is 6.8 percentage points, compared to 12.9 percentage points for readers ($p < 10^{-4}$). In Figure 1b, 50% of expert responses report the most likely series outcome implied by their game-by-game predictions, compared to 40% of reader responses ($p = 0.001$). For both the series probability and most likely series

¹²16% of game-by-game probabilities reported by experts and 24% of game-by-game probabilities reported by readers imply series probabilities either greater than 90% or less than 10%. Every reported sequence of game-by-game probabilities by either experts or readers implies a most likely series outcome that occurs with probability between 15% and 81%.

Figure 1: NBA playoffs: experts vs. readers.



Note: Experts exhibit greater internal consistency than readers. The distribution of expert responses stochastically dominates the distribution of reader responses in both graphs. p -values for the Kolmogorov-Smirnov (KS) and Mann-Whitney (MW) tests of distributional equivalence are reported in the figure legends.

outcome, the distribution of expert responses stochastically dominates the distribution of reader responses. Graphically, the cumulative distribution of expert responses lies above and to the left of the reader curve. For every level of inconsistency, proportionally more expert than reader responses exhibit that amount of inconsistency or less.¹³

Experts also outperform a set of heuristics when reporting the series probability. If each expert had reported a series probability equivalent to his or her reported game 1 probability, or to the mean, median, or mode of his or her reported game-by-game probabilities, the mean absolute error would have been 8.6, 9.6, 8.4, and 8.3 percentage points, respectively—in all cases higher than the observed mean error of 6.8 percentage points ($p < 10^{-4}$ for all

¹³We measure the statistical significance of the difference in consistency between the full distributions of expert and reader predictions using two-sample Kolmogorov-Smirnov (KS) and Mann-Whitney (MW) rank-sum tests, which test the null hypothesis that two empirical distributions are drawn from the sample population; p -values for these tests under homoscedasticity are reported in the legends of Figures 1a and 1b. For each outcome, both tests reject the null hypothesis, implying (given stochastic dominance) that experts demonstrate statistically greater internal consistency than readers.

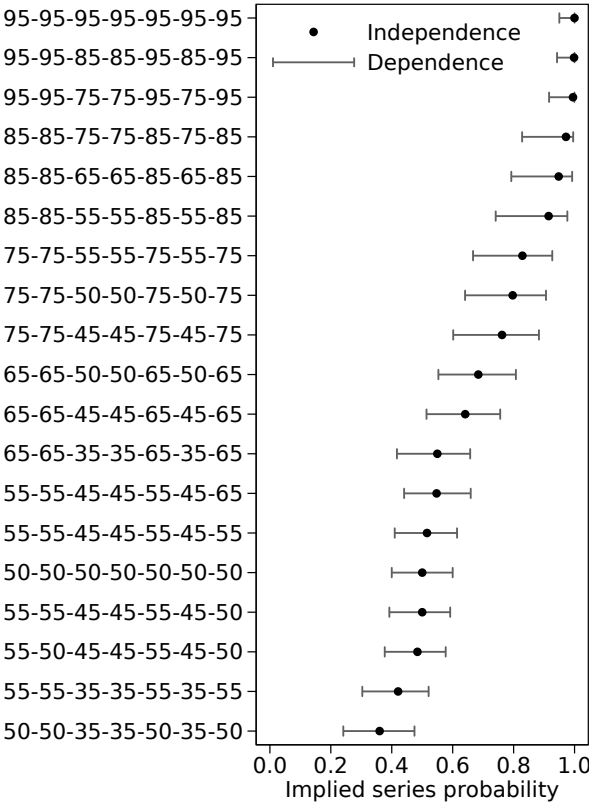
comparisons).¹⁴

One issue in assessing expert consistency is our assumption of sequential independence—i.e., that the outcome of a game, should it be played, does not depend on the outcomes of games earlier in the series. This concern is important for comparison across domains, as sequential independence is an explicit element of the problem in the unfamiliar context. Here, we relax the assumption of sequential independence, showing that reported series probabilities are inconsistent with dependence structures that diverge greatly from our sequence independence assumption. Specifically, we assume a flexible dependence framework in which game outcomes are conditional on the current series score (e.g., 2-1 in favor of the series home team), and reported game-by-game probabilities represent unconditional estimates. Under this sequential dependence structure, a sequence of game-by-game probabilities identifies a range of implied series probabilities, rather than the single implied value identified by the sequential independence assumption.

That range is typically quite large. Figure 2 shows bounds on the implied series probabilities for the most commonly reported game-by-game probabilities; Appendix B details the estimation. For many commonly reported sequences, the range of implied series probabilities is 20 percentage points or more. If experts possess dependence structures that deviate maximally from independence, and they evaluate series probabilities according to those structures, then they would report series probabilities that consistently differ from the implied value under independence by 10 percentage points or more. However, nearly 80% of reported series probabilities by experts are within 10 percentage points of their implied values under independence. Hence, for experts to believe in strong sequential dependence, they would either have to evaluate those beliefs precisely almost every time, or they would

¹⁴Reader responses underperform most of these heuristics. The mean error rate among reader responses of 12.9 percentage points outpaces the 10.9 percentage points ($p = 0.003$), 9.8 percentage points ($p < 10^{-4}$), and 9.9 percentage points ($p < 10^{-4}$) for the mean, median, and mode of the sequence of game-by-game probabilities; reporting the game 1 probability as the series probability yields greater inconsistency, with an average error of 14.5 percentage points ($p = 0.037$).

Figure 2: Bounds on the implied series probability for sequences of game-by-game probabilities reported at least 4 times by experts.



Note: Bounds were estimated by minimizing and maximizing the series probability over the unobserved conditional game-by-game probabilities subject to the observed unconditional game-by-game probabilities.

have to systematically misevaluate those beliefs in the direction of the implied value under independence. We contend that a more likely interpretation is that the experts report game-by-game probabilities as if game outcomes are more or less independent events.¹⁵

¹⁵Other concerns relate to the comparison between experts and novices, which we employ to highlight the expertise of our expert population. For instance, experts and readers may employ different beliefs about sequential dependence. However, this cannot explain observed differences in consistency between the populations, as the disparity persists even when allowing for sequential dependence. We find that 87% of reported series probabilities by experts fall within the estimated dependence bounds, compared to just 66% for readers ($p < 10^{-4}$). Another concern is that experts and readers may differ in the sequences of game-by-game probabilities they report. In particular, differences in internal consistency could merely reflect differences in the difficulty of evaluating summary statistics from reported game-by-game probabilities. In Appendix C, we show that differences in observed consistency cannot be explained by differential selection

Collectively, these results suggest that the experts we study possess expertise in a task that is familiar to them. The next study asks whether that expertise can be applied to an unfamiliar but logically isomorphic task, or equivalently, whether that expertise is in evaluating summary statistics of complex distributions or is more heuristic and context dependent.

4 Jars and marbles

Six months after the conclusion of the 2015 NBA playoffs, the same ESPN editor asked the expert panel to participate in a task “designed to improve ESPN’s NBA playoffs forecasting...[and] to contribute to academic research.”¹⁶ We designed the task to conceptually mimic the playoff prediction task in an unfamiliar environment. Specifically, the task described 7 ordered jars, each with 100 marbles total. Marbles were either black or red, and the proportions of black and red marbles in each jar were explicitly stated. Figure 3 shows an example problem, with 55 black marbles and 45 red marbles in jars 1, 2, 5, and 7, and 45 black marbles and 55 red marbles in jars 3, 4, and 6.

Subjects were told, “One marble will be drawn randomly from each jar in order, starting with Jar 1 and ending with Jar 7.” They were then asked to state 1) the probability that at least 4 of the 7 drawn marbles would be black, and 2) the color and jar combination from which they expect the 4th marble of the same color to be drawn from. (For the example in Figure 3, 4 or more black marbles are drawn with 51.6% probability, and the most likely outcome is that the 4th black marble will be drawn from the 7th jar.) For the first question, subjects typed in a percentage, which we round to the nearest percent. For the second question, subjects chose from among the 8 possible options—black in the 4th, 5th, 6th, 7th, red in the 4th, 5th, 6th, 7th. ¹⁶Appendix F contains the pre-registration document and an example of the invitation email.

Figure 3: Example survey page.

Imagine **7 jars** with **100 marbles each**. The marbles are either black or red.

Jar 1 has **55 black** marbles and 45 red marbles.

Jar 2 has **55 black** marbles and 45 red marbles.

Jar 3 has **45 black** marbles and 55 red marbles.

Jar 4 has **45 black** marbles and 55 red marbles.

Jar 5 has **55 black** marbles and 45 red marbles.

Jar 6 has **45 black** marbles and 55 red marbles.

Jar 7 has **55 black** marbles and 45 red marbles.

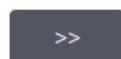
One marble will be drawn randomly from each jar in order, starting with Jar 1 and ending with Jar 7.

7 marbles will be drawn in total. **What is the probability that 4 or more will be black?**

Please enter a percentage between 0 and 100.

The 7 drawn marbles will either be majority black or majority red. **From which jar will the 4th marble of the same color most likely be drawn?**

- The 4th **black** marble will be drawn from **Jar 4**. The 4th **red** marble will be drawn from **Jar 4**.
- The 4th **black** marble will be drawn from **Jar 5**. The 4th **red** marble will be drawn from **Jar 5**.
- The 4th **black** marble will be drawn from **Jar 6**. The 4th **red** marble will be drawn from **Jar 6**.
- The 4th **black** marble will be drawn from **Jar 7**. The 4th **red** marble will be drawn from **Jar 7**.



or 7th jar, or red in the 4th, 5th, 6th, or 7th jar. This abstract problem is conceptually analogous to the structure of an NBA playoff series: the jars and marbles represent games and game-specific probabilities, respectively, and the questions ask for the series probability and the most likely series outcome. Contextual differences make the task unfamiliar: jars and marbles abstract away from teams and games, and participants observe sequences of proportions instead of reporting game-by-game probabilities.

Each respondent was directed to a personalized form comprising 5 sets of jars-and-marbles problems, with each set on a separate page. The first 4 pages showed sequences in which the proportions of black marbles replicated game-by-game probabilities reported by the same respondent for an NBA playoff series.¹⁷ Matching marble proportions to game-by-game probabilities allows us to compare performance by the same expert on a logically isomorphic problem. The fifth page showed the same sequence to all respondents: 7 jars containing 95 black marbles (and 5 red marbles) each. We designed this sequence to measure conceptual understanding at the end of the survey. The first 4 pages were randomly ordered, respondents could not navigate to previously completed pages, and no feedback was given.¹⁸

The ESPN editor sent personalized forms to the 119 experts who reported at least one non-trivial sequence of game-by-game probabilities during the 2014 or 2015 playoffs,¹⁹ and who was still affiliated the company at the time of this experiment. Of these 119 experts, 44 reported estimates for at least one matched sequence. This participation rate compares favorably to a coterminous ESPN poll, in which a superset of 382 basketball experts were asked to rank NBA players using an online form, and 98 participated in some capacity.

¹⁷When an expert reported more than 4 unique sequences of game-by-game probabilities for NBA playoff series, we choose the 4 sequences that maximize the minimum distance between any two sequences, where distance is defined as the sum of the absolute differences in game-by-game probabilities across the 7 games. When an expert reported fewer than 4 unique sequences, we include commonly reported (but unmatched) sequences to fill the difference. We analyze only the sequences that are matched by respondent across the familiar and unfamiliar contexts.

¹⁸Appendix E details the design of a pilot study and its results.

¹⁹We define trivial sequences as those for which the game-by-game probabilities are all 5%, 50%, or 95%, which occur 0, 14, and 4 times.

The experts who participate in the marbles survey show similar levels of consistency on the NBA prediction task as the experts who do not participate.²⁰ Among participants, however, the sequences that we sample for the marbles task appear easier to evaluate than those we do not sample, resulting in higher consistency in the NBA among matched sequences than in the NBA results as a whole. For our initial results, in which we compare performance on matched sequences across domains, we analyze the 143 matched sequences from all 44 participants. When we later compare performance across domains at the respondent level, we restrict our analyses to a subsample of 124 matched sequences from the 31 participants who each completed 4 matched sequences.

A natural concern in studies that compare across contexts or subject pools is differential motivation between the groups. We believe that this concern is unlikely to explain our results, for a number of reasons. First, participation in the marbles survey was optional. Those who participated did so out of interest, out of duty to the editor, or to help improve ESPN’s polling. Second, those who participate appear to take the survey seriously. The median completion time is 9 minutes and 7 seconds, or almost 2 minutes per page, and the minimum completion time is 3 minutes and 23 seconds.²¹ Error rates for those who finish the survey faster than the median completion time and slower than the median are comparable and statistically indistinguishable, suggesting that the fastest finishers do not greatly sacrifice accuracy.²² One subject appeared to solve the problems by simulation, taking more than 20 minutes and correctly answering each question to the nearest percentage point (but not to

²⁰For the series probability, the mean error for the series probability was 7.3 percentage points for survey participants, compared to 6.3 percentage points for non-participants ($p = 0.126$). For the most likely series outcome, 51% of survey participants predicted the implied most likely outcome, compared to 50% of non-participants ($p = 0.687$).

²¹Completion times are not comparable with the NBA context because the NBA playoff surveys elicited individual game probabilities in addition to series outcomes. We note that completion times may reflect time not spent on the survey while the survey window was open.

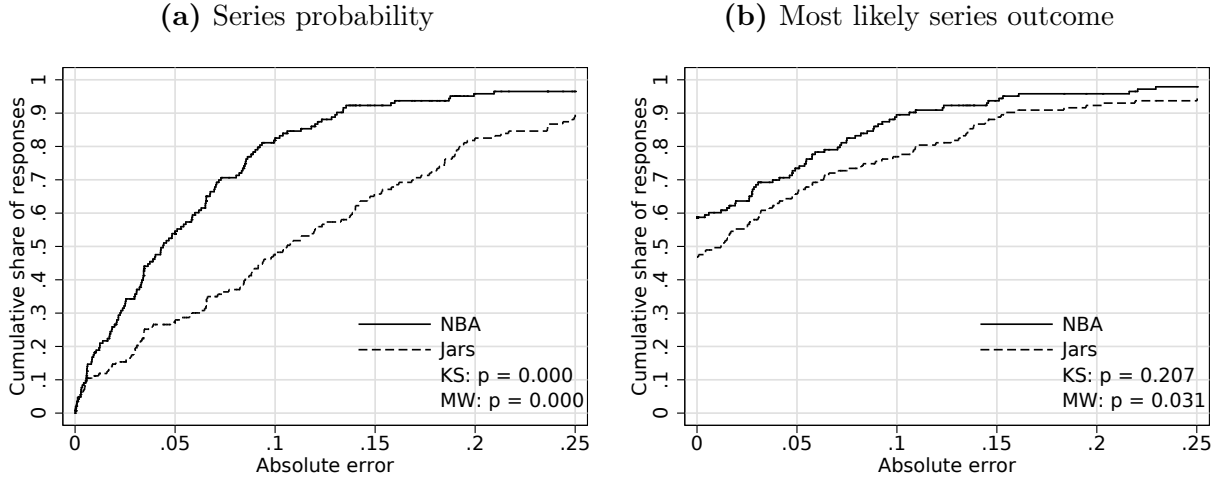
²²For the series probability, mean absolute error by respondent is 14% for the fastest 18 finishers and 13% for the slowest 18 ($p = 0.661$). For the series probability, the fastest 18 respondents choose the correct most likely series outcome 44% of the time, while the slowest 18 respondents make the correct choice 50% of the time ($p = 0.597$).

the nearest tenth of a percent, which the subject reported) or game. A second respondent, who sent a spreadsheet to the editor, correctly reported each series probability to the nearest percent, but chose the correct most likely series outcome for only two of the five sequences. Finally, performance improved with repetition, as we show later, suggesting attentiveness.

4.1 Performance across domains

Consistency plummets on the unfamiliar task. Figure 4 shows cumulative distributions of absolute error for the NBA and jars and marbles responses, separately for the series probability (4a) and most likely series outcome (4b). Error rates are considerably lower in the familiar domain. 54% of NBA responses report series probabilities within 5 percentage points of their implied values, whereas only 28% do so when the same problem is rendered abstractly. Similarly, 81% of NBA responses are within 10 percentage points of their implied values, compared to just 44% for the jars and marbles task. The mean error rate in the unfamiliar domain is more than twice as high as the corresponding rate in the NBA: 13.5 to 6.6 percentage points ($p < 0.001$). Comparable results pertain for the most likely series outcome: 59% of NBA responses predict the implied most likely series outcome, but only 47% do so in the unfamiliar domain ($p = 0.038$). For both the series probability and most likely series outcome, the error distribution for NBA responses stochastically dominates the error distribution for the jars and marbles task—i.e., there is no error level for which more marbles than NBA responses fall at or under that threshold. The probability that the error distributions are drawn from the same population is less than 10^{-7} for the series probability, regardless of whether the Kolmogorov-Smirnov or Mann-Whitney test is used. For the most likely series outcome, the error distributions are significantly different under the MW test ($p = 0.031$) but not under the KS test ($p = 0.207$). In sum, the experts prove largely unable to apply their expertise in an unfamiliar domain, failing to solve problems that are logically isomorphic to those they solved before.

Figure 4: NBA playoff predictions vs. jars and marbles.

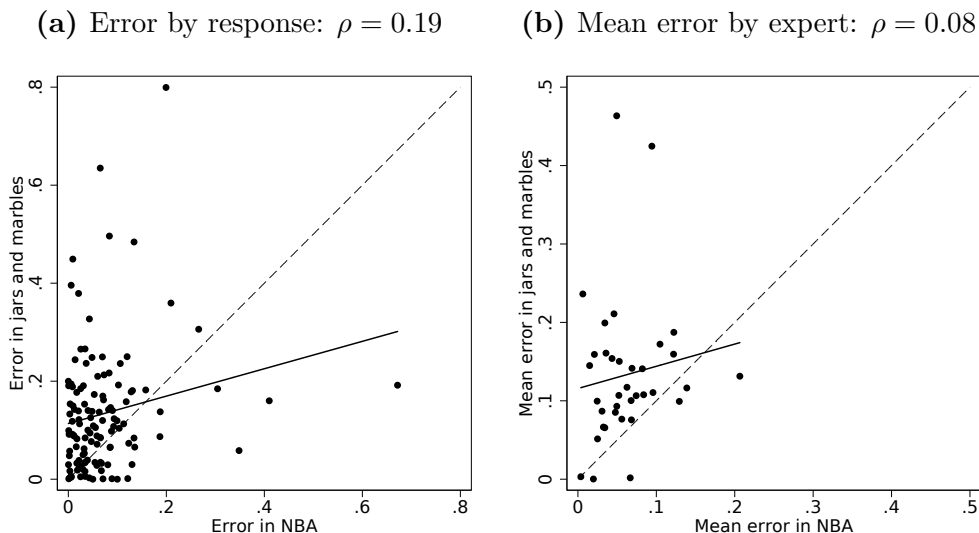


Note: The sample comprises 143 matched responses, in which the proportions of black marbles in each jar are identical to those reported by the expert for an NBA playoff series. The distribution of absolute error in the NBA stochastically dominates the distribution of absolute error on the matched jars-and-marbles problems, implying that experts fail to apply their expertise to the unfamiliar task.

Whereas the experts outperform a set of heuristics when making NBA predictions, their predictions for matched sequences in the jars task underperform those same heuristics. Among matched sequences, series probabilities equivalent to the reported game 1 probability, or to the mean, median, or mode of reported game-by-game probabilities each increase the mean NBA error of 6.6 percentage points by at least 2.3 percentage points ($p < 0.01$ for all heuristics). By contrast, the mean jars error of 13.5 would have decreased had each respondent instead reported a series probability equivalent to the proportion of black marbles in the first jar (by 4.3 percentage points, $p = 0.020$), or the mean (by 2.3 p.p., $p = 0.141$), median (by 4.5 p.p., $p = 0.006$), or mode (by 4.6 p.p., $p = 0.011$) of the proportions of black marbles among the 7 jars. The average expert spent minutes evaluating each sequence in the jars experiment. She would performed better by employing a quick and simple rule of thumb instead.

Performance in the jars task is not only poor, but it is only loosely correlated with performance in the NBA. Figure 5 shows scatter plots of absolute error in series probability judgments by context; Figure 5a compares errors for each matched response, and Figure 5b compares mean error rates by respondent. Errors in series probability judgments are correlated at 0.19 by response and 0.08 by respondent.²³ This decoupling is similarly severe for evaluations of most likely series outcomes. Correctness in these judgments is correlated across contexts at just 0.08, and mean correctness by respondent is correlated at 0.23. Performance in one domain poorly predicts performance in the other.

Figure 5: Series probability error rates across domains.



Note: The sample is restricted to 124 responses by the 31 experts who each complete 4 matched sequences. The solid line represents the best linear fit. Expert performance is weakly correlated across contexts.

²³Removing outliers does not meaningfully change these correlations. Removing the 3 observations in Figure 5a with an error in one domain in excess of 0.5 reduces the correlation to 0.16. Removing the 2 observations in Figure 5b with a mean error in the marbles survey in excess of 0.4 increases the correlation to 0.14.

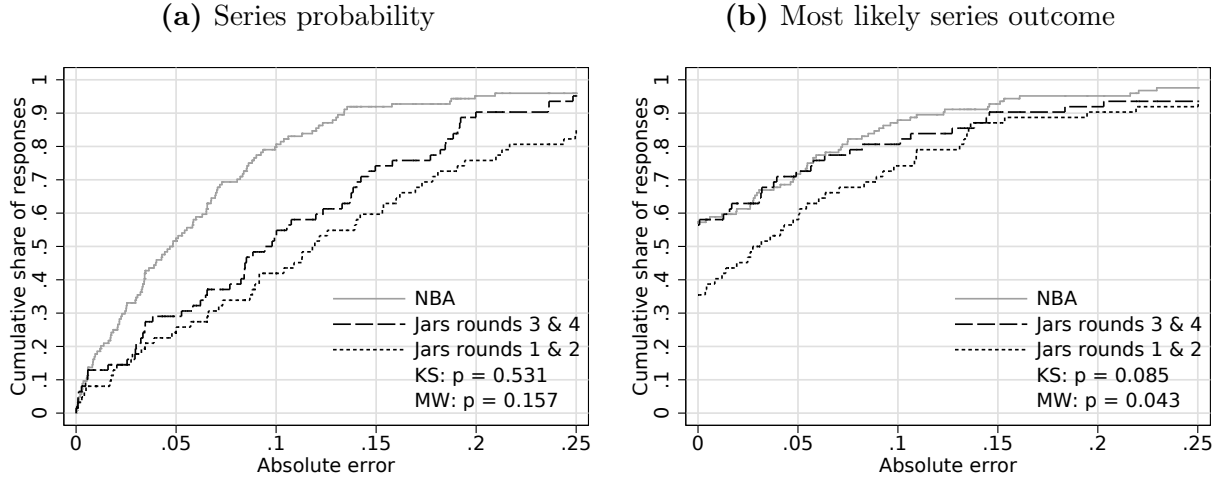
4.2 Learning in the unfamiliar domain

With practice, performance improves. Figure 6 shows cumulative distributions of absolute error for the first two matched sequences and the last two matched sequences, separately for the series probability (6a) and most likely series outcome (6b). The four matched sequences are randomly ordered, implying equivalence in average difficulty from round 1 to round 4. Yet in both figures, the error distributions for the third and fourth rounds stochastically dominate the error distributions for the first and second rounds—i.e., error rates decrease with repetition. For the series probability, neither distributional test can reject the null hypothesis that initial and subsequent performance are equivalent. For the most likely series outcome, however, the difference is more pronounced, and both the KS ($p = 0.085$) and MW ($p = 0.043$) tests reject equivalence at conventional significance levels. Mean comparisons show improvement on both questions, particularly for the most likely series outcome. The average error for the series probability declines from 15% in rounds 1 and 2 to 11% in rounds 3 and 4 ($p = 0.039$), though this improvement falls short of the 6.9% mean error for matched NBA responses ($p = 0.019$). For the most likely series outcome, just 35% of jars responses report the correct choice in the first two rounds, but 56% do so for the last two matched sequences ($p = 0.017$), which is comparable to the 57% rate for matched NBA responses ($p = 0.923$). With just a few repetitions, the experts more accurately judge series probabilities, and they evaluate most likely series outcomes as if predicting predicting playoff series rather than solving an abstract exercise.

These results show that the experts improve with repetition, but what exactly are they learning? We can rule out the possibility that they learn the underlying probabilistic structure. The fifth and final sequence, in which each of the 7 jars contains 95 black marbles and 5 red marbles, is likely to be unfamiliar to the experts, as it ignores home-court advantage.²⁴ This sequence has a series probability of 99.98%—i.e., drawing 4 or more black

²⁴Only 4 NBA responses report game-by-game probabilities of uniformly 95%, and all 4 report a series

Figure 6: Jars and marbles: rounds 1 & 2 vs. rounds 3 & 4.



Note: The sample is restricted to 124 responses—62 in rounds 1 and 2, and 62 in rounds 3 and 4—by the 31 experts who each complete 4 matched sequences. p -values for the KS and MW tests refer to comparisons between the first two and last two matched rounds of the jars study. Expert performance improves with repetition.

marbles is virtually guaranteed.²⁵ However, 60% of experts report a series probability of 95%, an intuitively appealing choice that reflects a conceptual misunderstanding of the binomial distribution; by contrast, only 27% report a series probability of 99 or 100%. Given this conceptual error, the observed improvement suggests that the experts develop a heuristic understanding of the marbles problem.

probability of 99% or 100%. However, these responses may reflect a censored view of the respondent’s true beliefs, as respondents could not choose a game-by-game probability higher than 95%. We attempted to circumvent this concern by asking the same set of questions again for two first-round series during the 2016 NBA playoffs in which the series home team was a historically prohibitive favorite—this time using sliders from 0% to 100%, incremented by 1%, to elicit the game-by-game probabilities. Unfortunately, only 2 of the 209 expert responses reported the sequence of interest (of which one reported a series probability of 99% and the other a series probability of 95%).

²⁵For this sequence, the true most likely series outcome coincides with perhaps the most intuitive response, and 90% of experts choose the correct most likely series outcome of “The 4th black marble will be drawn from Jar 4.”

4.3 Learning rates for experts and novices

An open question is whether the experts learn to apply their expertise to the marbles problem or instead develop an adaptive heuristic that is independent of their expertise. We arbitrate between these interpretations—learning to apply expertise, or learning an orthogonal heuristic—by recruiting a sample of non-experts to complete the same jars and marbles surveys.²⁶ We match initial performance on the marbles exercise between experts and non-experts by calibrating the incentives offered to non-experts.²⁷ This equivalence on initial performance indicates that the populations are comparable on ability and motivation. Presumably, they differ on expertise. We find that the non-experts improve more slowly than the experts, suggesting that expertise accelerates improvement on the unfamiliar task. However, we note that expertise is not randomly assigned, and as such, we cannot rule out the possibility that differences along dimensions other than expertise explain the observed differences in learning rates.

Specifically, we recruited a panel of respondents on Amazon Mechanical Turk.²⁸ We advertised a task about probabilistic judgment for academic research, comprising 5 problems of an identical format and 10 questions total, and with payment based on correctness—between \$0 and \$6, with an expected average of \$2.²⁹ Subjects were not given any further

²⁶Appendix F contains the pre-registration document for this study.

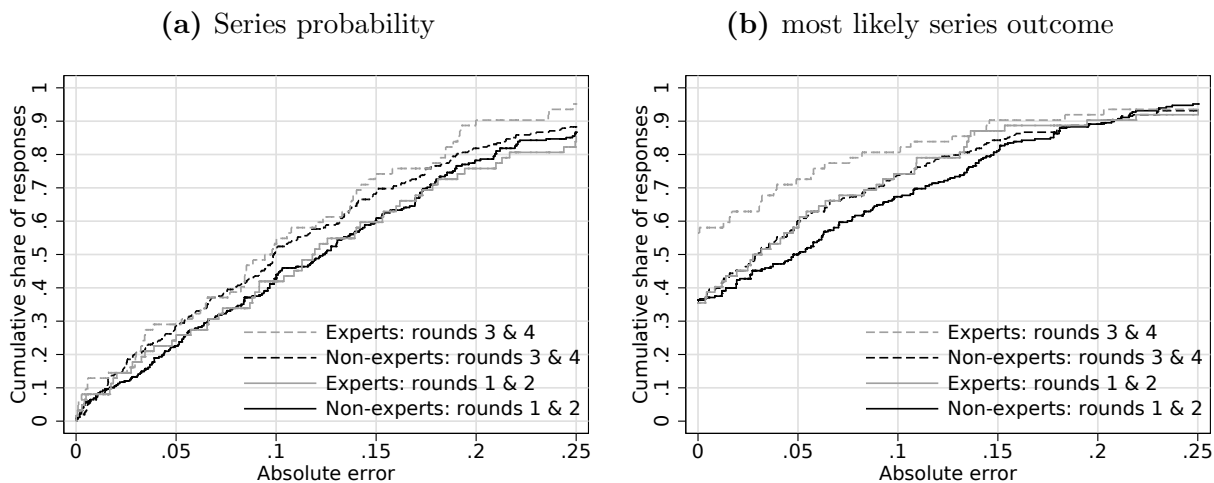
²⁷We calibrated the payment scheme so as to match the non-expert and expert samples on average performance over the first two sequences. Specifically, we ran four pilot studies with fewer subjects, and each with different incentives. In the first pilot, subjects were paid \$1 for completing the survey and no performance-based payment. In the second pilot, subjects were paid 50 cents for completion, along with performance incentives of up to \$3. In the third and fourth pilots, subjects were paid performance incentives of up to \$4 and \$6, respectively, with no payment for completion. Initial performance—i.e., in the first two rounds—generally increased with the performance-based incentives and by the fourth pilot, matched the average initial performance of the experts on the same sequences. As a result, we implemented the payment scheme from the fourth pilot in the subsequent study.

²⁸Appendix D shows the instructions on Mechanical Turk. Conducting experiments on crowdsourcing platforms, such as Mechanical Turk, has risen greatly in popularity in recent years (Mason and Suri, 2012) and a body of research has shown that these subjects replicate the behavior of traditional laboratory subjects in many experiments (e.g. Berinsky, Huber and Lenz, 2012; Paolacci, Chandler and Ipeirotis, 2010; Goodman, Cryder and Cheema, 2013).

²⁹The actual average payment was \$1.70.

guidance about how long the task would take to complete or how the bonus would be calculated. We replicated the 31 forms for which experts completed 4 matched sequences, and we randomly assigned non-experts to these forms. For each of the 31 surveys, we analyze the first 4 responses from non-experts who take at least 2 minutes to complete the task,³⁰ creating a sample of 496 balanced responses from 124 respondents. We compare this non-expert sample to the 124 responses from the 31 experts who evaluate the same sequences.

Figure 7: Rounds 1 & 2 vs. rounds 3 & 4.



Note: 496 responses—248 in rounds 1 and 2, and 248 in rounds 3 and 4—by 124 non-experts matched to 124 responses by 31 experts. Non-experts improve marginally with repetition and more slowly than experts, especially on the most likely series outcome question.

Figure 7 shows cumulative distributions of absolute error for non-experts, separately for the first two sequences (solid) and the last two sequences (dashed), and separately for the series probability (7a) and most likely series outcome (7b); for comparison, we superimpose the expert results from Figure 6 in gray. First, we highlight the similar performance of experts and non-experts in rounds 1 and 2, demonstrating the effectiveness of our matching procedure. For the series probability, error distributions in rounds 1 and 2 are overlapping

³⁰Despite the performance-based incentives, some Mechanical Turk respondents complete the problems too quickly to have considered them in a thoughtful way. We committed to this restriction in pre-registration.

for experts and non-experts, and the difference between the populations is statistically indistinguishable by either the KS ($p = 0.999$) or MW ($p = 0.975$) test. In the first two rounds, mean error for the series probability is 15% for both groups ($p = 0.860$). For the most likely series outcome, non-experts make the correct choice on 36% of first and second round responses, compared to 35% for experts ($p = 0.900$), and neither distributional test rejects the null hypothesis of equivalence between the populations (KS: $p = 0.552$; MW: $p = 0.458$).³¹

For the series probability, experts and non-experts improve at similarly slow rates. In Figure 7a, the two populations appear matched not only in rounds 1 and 2, but also in rounds 3 and 4. Relative to the first pair of sequences, mean error for non-experts declines by 1.7 percentage points when evaluating the second pair, compared to a decline of 4.1 percentage points for experts, and the difference in these declines is not significantly different from zero ($p = 0.271$).

For the most likely series outcome, learning rates diverge starkly. In Figure 7b, the distribution of expert error in rounds 3 and 4 stochastically dominates the corresponding distribution for non-experts, and the difference between the populations is statistically significant under both the KS ($p = 0.022$) and MW ($p = 0.013$) tests. Experts learn to identify the most likely series outcome while non-experts do not. Non-experts choose the correct most likely series outcome at the same rate in rounds 1 and 2 as in rounds 3 and 4. Experts, by contrast, improve by 21 percentage points, and these rates of improvement differ significantly ($p = 0.023$). Expertise appears to accelerate learning in an unfamiliar setting, but only for the most likely outcome task.

³¹Experts and non-experts differ on other observables. Median completion times are 6 minutes and 18 seconds for non-experts and 8 minutes and 52 seconds for this subsample of experts ($p = 0.008$). And on the common fifth sequence, 15% of non-experts and 27% of experts report 99% or 100% for the series probability, while 77% of non-experts and 90% of experts choose the correct most likely series outcome.

5 Conclusion

As-if justifications of rationality contend that individuals approximate optimal behavior through heuristics. We find evidence in support of this conjecture, showing that the heuristics experts employ mimic complex calculations. At the same time, we find that this intuitive expertise is not very portable. The experts we study fail at a logically isomorphic but contextually distinct problem. Their skill is not conceptual mastery but something more intuitive and brittle. When the problem changes slightly, their expertise crumbles.

This finding entails a normative implication for experimental design. In experimental economics, abstraction is convention: actions and payoffs are clear, but their corresponding analogs in real life are often not. Though economists readily complete the analogy, experimental participants may be mystified by these elements—even when they are familiar with analogous quantities outside of the laboratory. If people employ context-dependent heuristics to make good decisions in everyday life, then studies which abstract away contextual cues break those heuristics and bias their findings against standard theories of rationality.

In many professions, experts face similar yet unfamiliar problems. Doctors encounter patients with novel presentations. Securities analysts follow firms with innovative business models. And managers change teams, roles, or companies. [Kahneman and Klein \(2009\)](#) caution against the assumption that prior expertise will prove valuable even in conceptually similar domains, speculating that the applicability of intuitive expertise is tightly circumscribed around direct experience. Our results show how narrow that radius can be.

By manipulating a minimal set of contextual cues, we measure the fragility of expertise—the change in performance associated with an epsilon change in context. Of course, two points do not illuminate much of the terrain. We cannot say what would have happened had we manipulated a different set of contextual cues. Nor can we extrapolate our results to problems that do not require consistency in probabilistic judgment. Indeed, the adeptness

of humans at non-logical tasks like pattern matching suggests that other types of expertise would be robust to epsilon (or greater) changes in context. Illuminating this terrain strikes us as fruitful for understanding how far as-if rational actors may travel.

References

- Al-Ubaydli, Omar and John A List (2013) “On the generalizability of experimental results in economics: with a response to Camerer,” *National Bureau of Economic Research*.
- Bassok, Miriam and Keith J Holyoak (1989) “Interdomain transfer between isomorphic topics in algebra and physics,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 15, p. 153.
- Benz, Matthias and Stephan Meier (2008) “Do people behave in experiments as in the field? Evidence from donations,” *Experimental Economics*, Vol. 11, pp. 268–281.
- Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz (2012) “Evaluating online labor markets for experimental research: Amazon.com’s Mechanical Turk,” *Political Analysis*, Vol. 20, pp. 351–368.
- Bonin, Holger, Thomas Dohmen, Armin Falk, David Huffman, and Uwe Sunde (2007) “Cross-sectional earnings risk and occupational sorting: The role of risk attitudes,” *Labour Economics*, Vol. 14, pp. 926–937.
- Burks, Stephen V, Jeffrey P Carpenter, Lorenz Goette, and Aldo Rustichini (2009) “Cognitive skills affect economic preferences, strategic behavior, and job attachment,” *Proceedings of the National Academy of Sciences*, Vol. 106, pp. 7745–7750.
- Camerer, Colin (2011) “The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List,” *Available at SSRN 1977749*.
- Charness, Gary, Edi Karni, and Dan Levin (2010) “On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda,” *Games and Economic Behavior*, Vol. 68, pp. 551–556.
- Chiappori, P-A, Steven Levitt, and Timothy Groseclose (2002) “Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer,” *American Economic Review*, pp. 1138–1151.
- Cosmides, Leda and John Tooby (1996) “Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty,” *Cognition*, Vol. 58, pp. 1–73.
- Dyer, Douglas and John H Kagel (1996) “Bidding in common value auctions: How the commercial construction industry corrects for the winner’s curse,” *Management Science*, Vol. 42, pp. 1463–1475.
- Dyer, Douglas, John H Kagel, and Dan Levin (1989) “A comparison of naive and experienced bidders in common value offer auctions: A laboratory analysis,” *The Economic Journal*, Vol. 99, pp. 108–115.

- Falk, Armin and James J Heckman (2009) “Lab Experiments Are a Major Source of Knowledge in the Social Sciences,” *Science*, Vol. 326, pp. 535–538.
- Friedman, Milton (1953) *Essays in positive economics*: University of Chicago Press.
- Gigerenzer, Gerd and Klaus Hug (1992) “Domain-specific reasoning: Social contracts, cheating, and perspective change,” *Cognition*, Vol. 43, pp. 127–171.
- Goodman, Joseph K, Cynthia E Cryder, and Amar Cheema (2013) “Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples,” *Journal of Behavioral Decision Making*, Vol. 26, pp. 213–224.
- Grether, David M (1980) “Bayes rule as a descriptive model: The representativeness heuristic,” *The Quarterly Journal of Economics*, pp. 537–557.
- (1992) “Testing Bayes rule and the representativeness heuristic: Some experimental evidence,” *Journal of Economic Behavior & Organization*, Vol. 17, pp. 31–57.
- Haigh, Michael S and John A List (2005) “Do professional traders exhibit myopic loss aversion? An experimental analysis,” *The Journal of Finance*, Vol. 60, pp. 523–534.
- Harrison, Glenn W (1989) “Theory And Misbehavior Of First Price Auctions,” *The American Economic Review*, Vol. 79, p. 749.
- Harrison, Glenn W and John A List (2004) “Field Experiments,” *Journal of Economic Literature*, Vol. 42, pp. 1009–1055.
- (2008) “Naturally Occurring Markets and Exogenous Laboratory Experiments: A Case Study of the Winner’s Curse*,” *The Economic Journal*, Vol. 118, pp. 822–843.
- Hertwig, Ralph and Andreas Ortmann (2001) “Experimental practices in economics: A methodological challenge for psychologists?” *Behavioral and Brain Sciences*, Vol. 24, pp. 383–403.
- Kagel, John H and Dan Levin (1986) “The winner’s curse and public information in common value auctions,” *The American economic review*, pp. 894–920.
- Kahneman, Daniel and Gary Klein (2009) “Conditions for intuitive expertise: a failure to disagree.,” *American Psychologist*, Vol. 64, p. 515.
- Kessler, Judd B and Alvin E Roth (2014) “Don’t Take ‘No’ For An Answer: An Experiment With Actual Organ Donor Registrations,” *National Bureau of Economic Research*.
- Kovash, Kenneth and Steven D Levitt (2009) “Professionals Do Not Play Minimax: Evidence from Major League Baseball and the National Football League,” *NBER Working Paper*.

- Levitt, Steven D and John A List (2007a) “Viewpoint: On the generalizability of lab behaviour to the field,” *Canadian Journal of Economics/Revue canadienne d’économique*, Vol. 40, pp. 347–370.
- (2007b) “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?” *The Journal of Economic Perspectives*, Vol. 21, pp. 153–174.
- (2008) “Homo economicus Evolves,” *Science*, Vol. 319, pp. 909–910.
- Levitt, Steven D, John A List, and David H Reiley (2010) “What happens in the field stays in the field: Exploring whether professionals play minimax in laboratory experiments,” *Econometrica*, Vol. 78, pp. 1413–1434.
- Levitt, Steven D, John A List, and Sally E Sadoff (2011) “Checkmate: Exploring Backward Induction among Chess Players,” *The American Economic Review*, Vol. 101, p. 975.
- List, John A (2006) “The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions,” *Journal of Political Economy*, Vol. 114.
- List, John A and Michael S Haigh (2005) “A simple test of expected utility theory using professional traders,” *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102, pp. 945–948.
- Loewenstein, George (1999) “Experimental economics from the vantage-point of behavioural economics,” *The Economic Journal*, Vol. 109, pp. 25–34.
- Mason, Winter and Siddharth Suri (2012) “Conducting behavioral research on Amazon’s Mechanical Turk,” *Behavior research methods*, Vol. 44, pp. 1–23.
- Meier, Stephan and Charles Sprenger (2010) “Present-biased preferences and credit card borrowing,” *American Economic Journal: Applied Economics*, pp. 193–210.
- Neelin, Janet, Hugo Sonnenschein, and Matthew Spiegel (1988) “A further test of noncooperative bargaining theory: Comment,” *The American Economic Review*, Vol. 78, pp. 824–836.
- Palacios-Huerta, Ignacio (2003) “Professionals play minimax,” *The Review of Economic Studies*, Vol. 70, pp. 395–415.
- Palacios-Huerta, Ignacio and Oscar Volij (2008) “Experientia docet: Professionals play minimax in laboratory experiments,” *Econometrica*, Vol. 76, pp. 71–115.
- (2009) “Field centipedes,” *The American Economic Review*, Vol. 99, pp. 1619–1635.
- Paolacci, Gabriele, Jesse Chandler, and Panagiotis G Ipeirotis (2010) “Running experiments on amazon mechanical turk,” *Judgment and Decision making*, Vol. 5, pp. 411–419.

- Stoop, Jan, Charles N Noussair, and Daan Van Soest (2012) “From the lab to the field: Cooperation among fishermen,” *Journal of Political Economy*, Vol. 120, pp. 1027–1056.
- Sutter, Matthias, Martin G Kocher, Daniela Glätzle-Rützler, and Stefan T Trautmann (2013) “Impatience and Uncertainty: Experimental Decisions Predict Adolescents’ Field Behavior,” *The American Economic Review*, Vol. 103, p. 510.
- Tversky, Amos and Daniel Kahneman (1983) “Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment.,” *Psychological review*, Vol. 90, p. 293.
- Wooders, John (2010) “Does experience teach? Professionals and minimax play in the lab,” *Econometrica*, pp. 1143–1154.

Appendix for online publication

A Materials from the NBA playoff forecasting study

Figure 8: Example invitation for experts (email)

From: [REDACTED]
Sent: Thursday, April 16, 2015 12:04 PM
To: [REDACTED]
Cc: [REDACTED]
Subject: Western Conference Playoff Predictions: Vote now!

Hi folks:

You have been amazingly accurate in your playoff predictions in recent years, so let's do it again.

Ballots here:
[Warriors-Pelicans](#) | [Rockets-Mavs](#) | [Clippers-Spurs](#) | [Grizzlies-Blazers](#)


The turnaround time will be pretty tight, so please make your predictions ASAP. Shouldn't take too long.


Many thanks.

Figure 9: Example invitation for readers (blog post)

Predict: Nets-Raptors and Blazers-Rockets

4/18/2014

 Share with Facebook

 Share with Twitter

0

Shares



Royce Webb, Director, Content Analytics

We want to see what TrueHoop readers predict will happen in two series:
Nets-Raptors and Blazers-Rockets.



We're looking for your best, most accurate opinion about each series.

You can vote here: [Nets-Raptors](#) | [Blazers-Rockets](#)

We will publish the results at TrueHoop next week.

Thank you.

http://espn.go.com/blog/truehoop/post/_id/67583/predict-nets-raptors-and-blazers-rockets

Figure 10: Example survey, first page.

Eastern Conference Semifinals

Cleveland Cavaliers vs. Chicago Bulls

Format is 2-2-1-1-1.

Please predict all seven games.

Thanks for participating.

***Required**

Panel of voters

Your Name *

Cleveland Cavaliers vs. Chicago Bulls

What is the probability that Cleveland wins Game 1 vs. Chicago? *

Monday, May 4, at Cleveland

- 5%
- 15%
- 25%
- 35%
- 45%
- 50%
- 55%
- 65%
- 75%
- 85%
- 95%

Figure 11: Expert forecast published on ESPN.com

ESPN Forecast: Finals predictions

6/3/2015 - NBA,
GOLDEN STATE WARRIORS +1 more

Share with Facebook Share with Twitter

411

ESPN.com

Shares





It's the NBA's best team against the best player in the world. Which team -- the [Golden State Warriors](#) or the [Cleveland Cavaliers](#) -- will hoist the Larry O'Brien Trophy when the dust settles? We asked our ESPN Forecast panel who will prevail in the NBA Finals.

Our panel is predicting a six-game series win for [Stephen Curry & Co.](#) -- securing the franchise's first championship since 1975 -- over [LeBron James](#) and championship-starved Cleveland.

How it works: The ESPN Forecast panel predicted each team's likelihood of winning each of the seven potential games and each team's likelihood of winning the series. The top-line "Forecast" is based on a combination of the mean, median and mode of the predictions for each series.

(1) Warriors vs. (2) Cavaliers

Forecast: Warriors In 6		
		
Win in 4	8.8%	3.7%
Win in 5	19.5%	7.1%
Win in 6	16.6%	13.6%
Win in 7	20.3%	10.4%
Overall	65.2%	34.8%

B Unobserved dependence

One concern is that reported game-by-game probabilities may not be sequentially independent. We calculate summary statistics implied by a given sequence of game-by-game probabilities under the assumption that probabilities assigned to individual games do not depend on the outcomes of prior games. However, respondents may believe, for example, that the probability of winning game 6 depends on whether the series home team leads 3 games to 2 or trails 2 games to 3. If so, the probability she reports for game 6 will depend on the probability that she attaches to the series home team leading 3-2 relative to the probability of the series home team trailing 2-3. Given this sequential dependence, the likelihoods of the 8 possible series outcomes (e.g., series home team wins 4-2) may differ from their values under the independence assumption. For example, a response which assigns probabilities of 0.9 to each of the first 4 games implies, under the independence assumption, that the probability of the series home team winning 4 games to 0 is $0.9^4 = 0.66$. However, the respondent may instead believe that the home series team will win the first game with probability 0.9 and that the remaining games will be won with certainty by the winner of game 1. Under these beliefs, the probability of the series home team winning 4 games to 0 is equivalent to the game 1 probability of 0.9.³²

We use a flexible dependence framework to estimate bounds on the series probability for each reported sequence, finding that unobserved sequential dependence cannot explain our results: reported series probabilities by experts are far more likely to fall within these bounds than those reported by readers. Consider a sequential structure in which the outcome of game n depends on s_n , the state of the series prior to game n (e.g., the series home team

³²Note that under this dependence structure, game outcomes depend on the series score, not the manner in which that score was reached. Order effects are partially accounted for in our dependence structure. Whenever a team has won all of the prior games in the series, a dependence structure based on individual game outcomes is equivalent to one based on series score. However, an exhaustive accounting for the order of outcomes would require many more parameters than our model calls for (68 vs. 16).

leads 3-2 before game 6). We denote the probability with which the series home team wins game n in state s_n as $p_{n|s_n}$. Under this structure, the full distribution of outcomes is defined by 16 conditional probabilities $(p_1, p_{2|1-0}, p_{2|0-1}, \dots, p_{6|3-2}, p_{6|2-3}, p_7)$, rather than the 7 game-by-game probabilities under sequential independence (p_1, \dots, p_7) . Note that both p_1 and p_7 are sequentially independent under in our model—the former because it begins the sequence, and the latter because there is only one state in which a seventh game is played (i.e., when the series is tied 3 games apiece).

In our model, the respondent reports a sequence of unconditional probabilities in which the probability for game n , p_n , equals her expectation over the states in which game n is played. These probabilities are defined recursively with p_1 fixed, $p_2 = p_1 \cdot p_{2|1-0} + (1 - p_1) \cdot p_{2|0-1}$, and $p_n = p_{n-1} \cdot p_{n|s'} + (1 - p_{n-1}) \cdot p_{n|s''}$, where s' is the state in which the series home team wins game $n - 1$, and s'' is the state in which the series home team loses game $n - 1$. The conditional probabilities define the distribution of series outcomes. For example, the probability of the series home team winning 4 games to 0 equals $p_1 \cdot p_{2|1-0} \cdot p_{3|2-0} \cdot p_{4|3-0}$. From these series outcomes, exact summary statistics can be calculated.

Since we do not observe the parameters of the dependence structure—i.e., the conditional probabilities—we cannot identify the exact summary statistics implied by such a structure. However, the reported game-by-game probabilities constrain the values which the conditional probabilities can take, and these constraints define bounds on summary statistics of interest. We calculate bounds on the series probability, p_{series} , by finding, for a given sequence, its minimum and maximum values under these constraints:³³

$$p_{\text{series}} \in \left[\min_{p_{n|s_n}} p_{\text{series}}(p_{n|s_n}), \max_{p_{n|s_n}} p_{\text{series}}(p_{n|s_n}) \right]$$

$$\text{s.t. } p_{n|s_n} \in [0, 1] \ \& \ p_n = p_{n-1} \cdot p_{n|s'} + (1 - p_{n-1}) \cdot p_{n|s''}$$

³³We solve for the minimum and maximum values using a convex optimizer, initializing the parameters at their values under an assumption of independence.

Figure 2 displays the bounds on p_{series} for commonly reported sequences of game-by-game probabilities. For example, a sequence of 95% in all games can be rationalized by a series probability as low as 95%—if $p_{2|1-0} = 1, p_{3|2-0} = 1, p_{4|3-0} = 1$, and all other conditional probabilities equal 0.

C Selection of probabilities

Another concern is that the differences between expert and reader responses in Figure 1 may result from asymmetries in the task, rather than differences in consistency. At issue is that each response reports both game-by-game probabilities and statistics which summarize that sequence—i.e., respondents choose the distributions they evaluate. In particular, readers may choose sequences whose implied summary statistics are more difficult to compute than the implied values of sequences chosen by experts.

To address this concern, we construct the largest possible matched sample in which the distribution of sequences is identical for experts and readers.³⁴ While there are billions of combinations of expert and reader responses that produce such a sample, there are zero such samples in which the readers evaluate the series probability more consistently, on average, than experts. Specifically, there is no sample in which the net area between the cumulative distributions of expert and reader errors is negative. In other words, it is impossible to construct a large matched sample in which experts exhibit greater inconsistency than readers.³⁵

³⁴For the series probability, we restrict the full set of responses to the 19 sequences that experts and readers each report at least once. Of these, 9 have exactly one expert response and one reader response. The remaining 10 sequences differ in the number of expert and reader responses. For these sequences, we sample from the panel with more responses for that sequence. Each matching contains 18 responses from the 9 sequences that are matched perfectly in the data and 34 responses from the 10 sequences in which the number of expert and reader responses differs.

³⁵This statement is based on the most extreme match—in which the most inconsistent expert responses are matched with the least inconsistent reader responses—rather than observation of the entire set of feasible matches.

A related concern is that experts may more frequently than readers report series probabilities close to 50%, thereby restricting their maximum error relative to readers. This supposition is inconsistent with the data, as distributions of summary statistics are comparable for experts and amateurs. While readers tend to report extreme series probabilities more often than experts, those series probabilities correspond to the trivial and frequently reported sequences in which the series home team has either a 95% or 5% chance of winning every game—for which reported summary statistics are nearly always correct. In other words, readers report extreme summary statistics only when they know that they are correct. Ignoring these trivial sequences as well as uniform sequences of 50%, reported series probabilities by experts and readers are on average 19 and 20 percentage points, respectively, from 50% ($p = 0.41$).

D Materials from the jars and marbles study

Figure 12: Example survey invitation for experts

From: [REDACTED]
Sent: Friday, December 18, 2015 10:48 AM
To: [REDACTED]
Subject: New NBA survey: Vote now

You have been selected for a new survey designed to improve ESPN's NBA playoffs forecasting.

The survey is five brief pages -- 10 questions total.

Your participation will also contribute to academic research, and be greatly appreciated by me and our partners at Microsoft Research.

If you can complete it by Tuesday, Dec. 22, that's ideal. If you need longer, please let me know.

Hit me with any questions you have.

Thank you!

Figure 13: Description of task on Mechanical Turk

Instructions

We are conducting an academic survey about probabilistic judgments. The survey has 5 problems of an identical format -- and 10 questions in total.

A bonus between \$0 and \$6 will be paid based on correctness. We anticipate that the average bonus will be \$2.

Please read the questions closely and answer to the best of your abilities. Please do not attempt this survey more than once.

Make sure to leave this window open as you complete the survey. When you are finished, you will return to this page to paste the code into the box.

Template note for Requesters - To verify that Workers actually complete your survey, require each Worker to enter a unique survey completion code to your HIT. Consult with your survey service provider on how to generate this code at the end of your survey.

Survey link: https://stanforduniversity.qualtrics.com/SE/?SID=SV_b7mM0TA3yFYNtb

Provide the survey code here:

E Pilot study

We designed the jars and marbles study after a pilot study found poor performance by the expert panel on a jars-and-marbles problem of the same format. In the pilot study, the marble proportions were common to all participants and corresponded to a frequently reported sequence of game-by-game probabilities for an NBA playoff series, with 65 black marbles in jars 1, 2, 5, and 7, and 45 black marbles in jars 3, 4, and 6.³⁶ We then compared responses by survey participants to NBA predictions made by a non-overlapping set of experts, finding far higher accuracy in the NBA domain.

29 experts completed this pilot survey. In the NBA domain, 17 responses report this sequence and evaluate the most likely series outcome; 13 of these also evaluate the series probability. Of these 13, none reports a series probability more than 12 percentage points from the correct value. By contrast, 9 of the 29 experts (31%) who take the pilot survey do so. The difference is similarly extreme at a 5 percentage point cutoff: 9 of 13 (69%) NBA responses are within this error bound, compared to just 8 of 29 (28%) pilot survey responses. For the most likely series outcome, 11 of 17 (65%) NBA responses provide the exact implied

³⁶Respondents also evaluated two other sequences with marble proportions that did not mimic game-by-game probabilities for a typical NBA playoff series.

value, compared to just 14 of 29 (48%) pilot survey responses.

This performance gap may be explained by unobserved differences between these groups, rather than differences in contextual cues. By matching marble proportions to game-by-game probabilities reported by the same respondent, the following study allows for within-respondent comparisons—thereby alleviating selection concerns, isolating the effect of the unfamiliarity on the portability of expertise, and providing stronger evidence in support of the conclusion that the experts fail to apply their expertise in the unfamiliar domain.

F Pre-registration documents

December 17, 2015

Pre-registration for "*Expertise in the Field Fades in the Lab*"
by Etan A. Green, Justin M. Rao, and David Rothschild

Summary of work so far, as in the September 3, 2015 version of the paper

Over 100 experts from ESPN completed a task wherein they predicted, for a number of NBA playoff series, 1) the probability that the favored team would win each game of the series (7 values), and 2) the probability that the favored team would win the series (1 value). We find that the reported series probabilities in (2) are highly consistent with the implied series probabilities from the game-by-game predictions in (1)---both in absolute terms and relative to a panel of fans performing the same task.

To test whether this skill is portable to a lab context, we invited a handful of these experts to participate in a pilot study involving a formally identical task. Subjects were told that 7 jars contained 100 marbles each and that the marbles were either black or red. In addition, subjects were told the proportions of black and red marbles in each jar. Subjects were then told that a single marble would be drawn randomly from each jar, with 7 marbles drawn in total. Subjects were asked to then state the probability that at least 4 black marbles would be drawn in total. We gave all subjects a common sequence of proportions that is identical to a commonly elicited series of game-by-game probabilities. Despite this formal equivalence, expert subjects performed considerably worse in the jars and marbles setting than subjects who reported those same probabilities in the NBA setting.

Experimental design

We plan to test whether the results from our pilot study hold when the expert subjects evaluate sequences of jars whose proportions of black and red marbles are identical to sequences of game-by-game probabilities that they reported for an NBA playoff series (i.e., as opposed to a common, generic sequence).

To do this, we identified 116 experts who forecast at least one NBA playoff series that met the following 2 criteria: 1) the sequence had a unique most likely outcome, and 2) the game 7 prediction was not 50% (such sequences may give the impression that two outcomes are equally likely). Among these experts, we chose the 4 sequences that met these criteria and were maximally different from each other (using a maxmin algorithm that chose the set of 4 sequences whose most similar pair, measured as the sum of the absolute differences in probabilities between their elements, is greatest). For experts who report fewer than 4 sequences, we added commonly reported, non-redundant sequences.

We generated personalized surveys 5 pages in length, with two questions per page, of the following format:

Imagine **7 jars with 100 marbles each**. The marbles are either black or red.

Jar 1 has **75 black** marbles and 25 red marbles.

Jar 2 has **75 black** marbles and 25 red marbles.

Jar 3 has **55 black** marbles and 45 red marbles.

Jar 4 has **55 black** marbles and 45 red marbles.

Jar 5 has **75 black** marbles and 25 red marbles.

Jar 6 has **55 black** marbles and 45 red marbles.

Jar 7 has **75 black** marbles and 25 red marbles.

One marble will be drawn randomly from each jar in order, starting with Jar 1 and ending with Jar 7.

7 marbles will be drawn in total. **What is the probability that 4 or more will be black?**

Please enter a percentage between 0 and 100.

The 7 drawn marbles will either be majority black or majority red. **From which jar will the 4th marble of the same color most likely be drawn?**

- The 4th **black** marble will be drawn from **Jar 4**. The 4th **red** marble will be drawn from **Jar 4**.
- The 4th **black** marble will be drawn from **Jar 5**. The 4th **red** marble will be drawn from **Jar 5**.
- The 4th **black** marble will be drawn from **Jar 6**. The 4th **red** marble will be drawn from **Jar 6**.
- The 4th **black** marble will be drawn from **Jar 7**. The 4th **red** marble will be drawn from **Jar 7**.

>>

The only difference between pages is the number of black and red marbles listed in each jar. The first 4 pages show the 4 respondent-specific sequences in random order. The 5th page is

comprehension test common to all responds: a sequence of 7 jars in which each has 95 black marbles (for which the correct answers are trivially easy).

Survey distribution and timeline

An editor at ESPN will email personalized links to each of the 116 expert respondents at 10am EST on Friday, December 18. The editor will issue a reminder email to those who have not completed the survey early the following week, and again the week of January 3rd. We will close the survey at the end of the workday on Friday, January 15th, and only thereafter will we analyze the responses.

Plans for analysis

We will reconstruct the same figures as in the September 9th version of the paper. We will measure correctness for the most likely outcome by the *probability* distance between the chosen and true most likely outcome, in addition to the distance in games/jars of those outcomes. We will also explore sources of heterogeneity in correctness, in particular attributes of the sequence, responses to the comprehension check, with and without answers that involve decimal precision (suggestive of using a computer), question ordering, and the expert's level of experience and statistical background (as assessed by the editor)..

February 3, 2016

Pre-registration for “*Expertise in the Field Fades in the Lab*”, part 2
by Etan A. Green, Justin M. Rao, and David Rothschild

Summary of results since September 3, 2015 version of the paper

We presented the same sequences of game-by-game probabilities reported in the NBA context to the same experts, but now rendered in the abstract jars and marbles language described in the 9/3/2015 version of the paper and the first pre-registration document.

We found that expert performance in the abstract domain is considerably lower than the consistency shown in the NBA domain. However, performance improves with repetition: error rates for the third and fourth sequences were lower than those for the first and second sequences. Given that the sequences were randomly ordered, these findings suggest some measure of learning. One interpretation is that the experts are learning to apply their expertise in the abstract context. A second interpretation is that the experts are learning how to solve the problems independent without relying on their expertise.

Experimental design

We plan to arbitrate between two explanations by presenting the same sequences to a population of novices, or those who lack experience in making NBA predictions. If the two groups learn at different rates, a likely explanation is that expertise either aids or hinders adaptation to the unfamiliar context.

Specifically, we will run the same jars and marbles experiment with subjects on Amazon’s Mechanical Turk (MTurk), randomly assigning each subject a particular expert’s sequence. We will incentive subjects by paying them strictly based on accuracy. They will be promised a bonus of between \$0 and \$6 with an expected bonus of \$2. No other information will be given on how the bonus is calculated or how much time we expect the task to take. We will restrict our sample to subjects who take at least two minutes to complete the survey, so as to weed out those who do not take the task seriously. Finally, we will solicit subjects until we have 4 complete responses, each taking at least two minutes to complete, for each sequence.

Survey distribution and timeline

The survey will be sent out today (2/3/16). We expect completion within an hour.

Plans for analysis

We plan to compare performance on the jars and marbles task between experts and novices. Specifically, we will compare rates of learning between the two groups.