A Sharp Test of the Portability of Expertise[*]

Etan A. Green[†], Justin M. Rao[‡], and David Rothschild[§]

Microsoft Research

February 21, 2016

Abstract

We measure the contextual dependence of expertise by observing experts perform a task conceptually identical to—but contextually distinct from—the task in which they are experienced. We find that these experts fail to apply their expertise in the decontextualized environment, indicating that unfamiliarity restricts the portability of expertise. However, we also find that the experts improve with repetition in the new setting—and more quickly than comparably able novices—suggesting that experts learn to apply their expertise on the initially foreign task.

[*] We thank seminar participants at Microsoft Research, Jenn Wortman Vaughan in particular, as well as Jim Andreoni, Hengchen Dai, David Reiley, Joe Simmons, and Charlie Sprenger, for helpful comments and suggestions. We are grateful to Henry Abbott, Kevin Arnovitz, and Royce Webb for creating panels of experts and novices, and for inviting us to work with them on their polling. We also thank Microsoft Research for generous financial support.

[†] etangreen@gmail.com

[‡] justin.rao@microsoft.com

[§] davidmr@microsoft.com

Electronic copy available at: http://ssrn.com/abstract=2656268

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 perfect logician immediately discerns the logical rules of a new environment, the boundedly rational actor develops adaptive heuristics through experience. As a result, mental shortcuts honed in one context may leave actors ill-equipped in another. Kagel and Levin (1986) illustrate this concern in an experimental study of repeated common value auctions. They find that laboratory subjects learn to avoid the winner's curse by bidding a fraction of their private signals, but when minor contextual details change, players revert to bidding their signals and experience the winner's curse again. Individuals often find themselves in contexts outside of their direct experience. The extent to which economic theories predict behavior in these cases depends on the portability of their expertise.

A handful of studies measure this portability by observing real-world experts perform

[1] For instance, Harrison and List (2004) hypothesize that "naturally occurring [auction] markets are efficient because certain traders use heuristics to avoid the inferential error that underlies the winner's curse."

Electronic copy available at: http://ssrn.com/abstract=2656268

unfamiliar tasks that conceptually replicate those in which they are experienced. These studies, which we review in Section 2, have produced conflicting results. As we detail, this disagreement may stem from the obscurity of the unfamiliar tasks: each of these studies examines the behavior of experts on conventional laboratory games that differ markedly from their expertise.

In this paper, we conduct a sharp test of the portability of expertise by observing experts perform nearly the same task in which they are experienced. This analogous task differs from our subjects' domain of expertise on a minimal set of contextual cues; on formal dimensions, it is isomorphic. Our main finding is that these experts fail to apply their expertise in the unfamiliar environment, implying that unfamiliarity restricts the portability of expertise. However, we also find that the experts improve with repetition in the new setting—and more quickly than comparably able novices—suggesting that relevant expertise accelerates learning on unfamiliar tasks.

The experts we study are members of a forecasting panel who regularly make probabilistic predictions about individual events and sets of events. Formally, the task is similar to predicting both the probability of default on individual mortgages in a mortgage-backed security and the probability that the security will be downgraded, or to 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 individual events (e.g., mortgage defaults) are consistent with her predictions for an encompassing set of events (e.g., a downgrade). We compare the consistency of predictions by the same experts between a task that is familiar to them and a conceptually isomorphic task in which the context is rendered abstractly (as drawing balls from urns).

Specifically, we observe predictions from a panel of more than one hundred basketball experts run by the ESPN sports network, which are published on ESPN.com. Panelists predict the outcomes of playoff series in the National Basketball Association (NBA), which follow a best-of-7 format, ending after one team wins 4 games. Before each series, these experts are asked to report: 1) 7 probabilities corresponding to the likelihoods that the favored team in the series wins each of the 7 games (should the 5th, 6th, or 7th game be played), 2) the probability that the favored team wins the series (hereafter, series probability), and 3) the most likely series outcome chosen from 8 possible series outcomes (i.e., favored team in 4, 5, 6, or 7 games, or the underdog in 4, 5, 6, or 7 games). Assuming that respondents interpret the game-by-game probabilities as sequentially independent—an assumption that is justified by the data—the game-by-game probabilities imply both the series probability and the most likely series outcome. We evaluate the consistency of a respondent's predictions by comparing her reported series predictions with those implied by her game-by-game probabilities. ESPN solicits the same predictions from readers of a basketball blog on ESPN.com, and we use these predictions to construct a non-expert baseline.

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 game-by-game and series predictions made by these experts in their domain of expertise. The average deviation between an expert's predicted series probability and the value implied by the game-by-game probabilities is just 6.7 percentage points, compared to 13 percentage points for readers. Similarly, experts report the implied most likely series outcome on 49% of responses, whereas readers do so only 39% of the time.

The same ESPN editor then asked the same expert panel to perform a conceptually identical exercise. Specifically, the task described 7 ordered jars, each with black and red marbles in specified quantities, and each with 100 marbles total. Subjects were told that a single marble would be drawn sequentially from each jar. They were then asked to report 1) the probability that 4 or more black marbles will be drawn in total—i.e., the probability that black will "win" the series of draws; and 2) the jar they expect the 4th marble of the same color to be drawn from—i.e., the most likely best-of-7 outcome. Each expert observed 5 sequences of jars, 4 of which were matched to the NBA context—i.e., each sequence contained black and red marbles in identical proportions to game-by-game probabilities reported by that same expert for an NBA playoff series. The fifth sequence, shown to all respondents, contained 95 black marbles (and 5 red marbles) in each jar. The jars-and-marbles problem is a close copy of the NBA playoff forecasting task: the questions were conceptually identical and respondents observed the same sequences of probabilities on the unfamiliar task that they had previously solved on the familiar one. Moreover, the link between the contexts was made explicit: the editor described the marbles exercise as "designed to improve ESPN's NBA playoffs forecasting."

The experts appear motivated to solve the jars-and-marbles problems, taking their time and giving precise answers. Nonetheless, they display significantly less consistency on the unfamiliar marbles task, both in economic and statistical terms. Among matched sequences, reported series probabilities differ from their implied values by an average of 14 percentage points in the unfamiliar context—twice the error rate in the NBA context. Similar results pertain for the most likely series outcome: 26% more responses predict the implied most likely series outcome on the familiar task than on the unfamiliar one. Furthermore, errors are weakly correlated across the familiar and unfamiliar contexts, suggesting that the experts behaved as if facing new problems, rather than ones they had already solved.

We interpret these results to imply that the unfamiliarity of the jars-and-marbles exercise inhibits the transfer of expertise. The jars-and-marbles problem is unfamiliar to NBA experts in two ways: the problem is presented in the abstract language of jars and marbles instead of games and probabilities, and the proportions of marbles in jars are imposed, whereas the game-by-game probabilities are elicited from the experts themselves.[2] Research in cognitive

[2] Reporting the game-by-game probabilities may help familiarize the respondent with the series.

science shows that individuals have difficulty transferring expertise from a specific to an abstract domain (Loewenstein, 1999)—e.g., solving an algebra problem after learning to solve a conceptually identical problem in physics (Bassok and Holyoak, 1989).[3] Analogously, NBA experts have difficulty applying skills learned from games and teams to an abstract, though formally equivalent, question involving jars and marbles. Our results suggest that the internal consistency displayed by the experts in the NBA is learned implicitly. While many of our expert subjects have statistical training, few explicitly calculate summary statistics for the complex distributions they are asked to assess. Instead, NBA experts may possess separate intuitions for 1) outcomes of games, and 2) the outcomes of the series, and it may be that with experience, these intuitions become consistent.

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 accuracy increases with repetition. For the series probability, the average error drops from 16 percentage points for the first and second sequences to 11 percentage points for the third and fourth, a decline that is statistically significant but insufficient to reach the 7 percentage point rate in the NBA context. For the most likely series outcome, repetition erodes the effect of unfamiliarity completely.

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%. Instead, the

[3] The Wason selection task offers a complementary insight: when problems are rendered in familiar, rather than abstract terms, subjects prove more capable. In 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 either learn to apply their domain expertise as they become more familiar with the problem, or they refine a heuristic separate from their expertise. We arbitrate between these interpretations by assigning the same jars-and-marbles problems to a population of comparatively able novices on Mechanical Turk, calibrating performance incentives to match first- and second-round accuracy between the experts and novices. We find that with repetition, the novices improve only marginally—and more slowly than the experts—suggesting that expertise plays a role in the experts' improvement on the unfamiliar task.

Collectively, our 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, the main result that unfamiliarity restricts the portability of expertise raises questions about the extent to which behavior by experts in abstract laboratory settings (e.g. Haigh and List, 2005; List and Haigh, 2005; Harrison and List, 2008) predicts behavior by those actors in more familiar contexts. Our results also speak to questions about the influence of unfamiliarity on non-experts. In contrast to many laboratory experiments in social psychology, in which elements of familiar environments are recognizable, the convention in economics is to render laboratory experiments in abstract terms, in which theoretical concepts like incentives and beliefs are readily identifiable, but contextual cues are not (Hertwig and Ortmann, 2001).[4] Our results indicate that individuals behave differently in abstract laboratory environments than in the conceptually similar environments with which they are familiar.

Our findings also show that establishing familiarity through repetition can bring outside experience to bear in the lab. We observe that rates of improvement on an unfamiliar task are correlated with underlying expertise. As a result, laboratory subjects who learn quickly may have stronger external validity, making them a useful sub-sample for analysis.

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

2 Related Literature

Two sets of studies on experts, one with professional soccer players and a second with chess grandmasters, examine the portability of expertise to traditional laboratory games. Chiappori, Levitt and Groseclose (2002) and Palacios-Huerta (2003) show that in professional soccer, penalty kicks are consistent with minimax predictions: kickers and goalies appear to 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. One concern is that soccer players may not expertly play mixed strategies in the field—tests of optimal randomization in penalty kicks are underpowered (Kovash and Levitt, 2009), and a failure to reject the null of optimal play does not imply optimal play. Levitt et al. (2010) find that professional poker players, whose expertise in randomization is arguably less ambiguous than that of soccer players,[5] do not apply that expertise in the lab.

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 grand-

[5] One justification for this argument is that experience is more frequent for poker players than for soccer players: poker players play many hands in a sitting, whereas soccer matches end in penalty kicks infrequently, and when they do, relatively few kicks decide the game.

masters 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. Addressing this concern, Levitt et al. (2011) show that grandmasters exhibit greater skill when playing the "race to 100" game, which like chess is winner-take-all. However, seemingly incidental rule changes (e.g., whether the game can be maximally incremented by 9 or 10) nevertheless induce large changes in behavior.

One interpretation of these conflicting results is that studies which compare expert behavior across vastly different environments have difficulty isolating the effect of unfamiliarity on expertise. For instance, subtle framing effects may differ between purportedly comparable designs. In the centipede game, these framing effects may motivate one group of subjects to play "the right way" and another to make the most money. By contrast, our study lessens the distance between familiar and unfamiliar domains, allowing us to more credibly measure the portability of expertise.

A separate literature measures the degree to which preferences, rather than expertise, correlate across lab and field. The findings here are mixed: in some cases, preferences in the lab predict those in the field (e.g. Benz and Meier, 2008), 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 more than one hundred 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 their 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 survey 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. 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 exception of the game 1 probability, none of these quantities are reliably estimated by Vegas betting lines or prediction markets prior to game 1.[6] Appendix A shows examples of the survey invitation and the online survey form. Our sample comprises 165 experts, who each complete between 1 and 32 surveys over the observation window (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 472 reader responses over 2 series.[7]

[6] Betting 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.

[7] This 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. Unfortunately, inconsistent name

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.[8] 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.

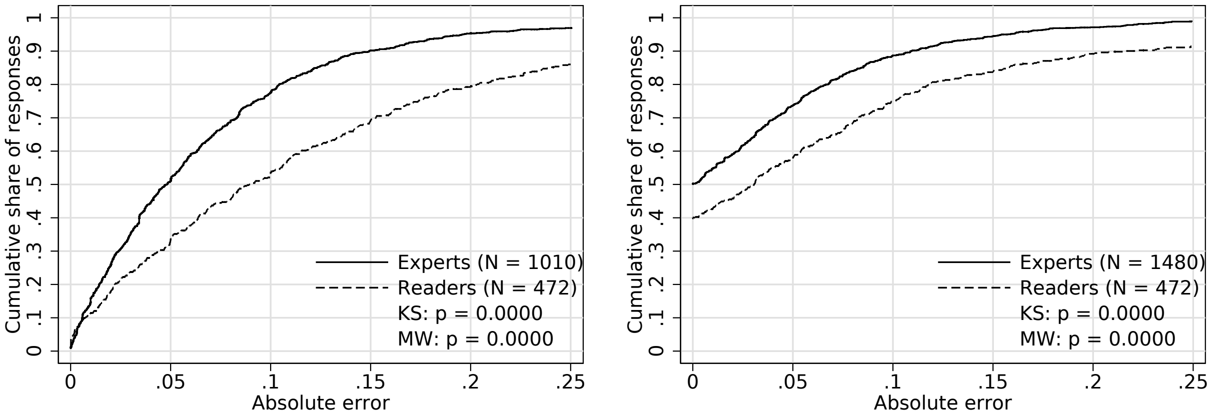
We measure the internal consistency of a prediction by its absolute error. For the series probability, the absolute error is the difference between the reported and implied values; for the most likely series outcome, the absolute error is the difference between the probability of the reported most likely series outcome and the probability of the implied most likely 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.[9]

reporting prevents us from identifying multiple responses from the same reader.

[8] For 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.

[9] 16% 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 intermediate probability—between 15% and 81%.

Figure 1: NBA playoffs: experts vs. readers.



(a) Series probability

(b) most likely series outcome

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.

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. In Figure 1b, 49% of expert responses report the most likely series outcome implied by their game-by-game predictions, compared to 39% of readers. For both the series probability and most likely series 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.

We measure the statistical significance of the difference in consistency between experts and readers 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 are reported in the legends of Figures 1a and 1b. For each summary statistic, both tests reject the null hypothesis, implying that experts demonstrate statistically greater internal consistency than readers.[10]

Experts also outperform a set of heuristics when reporting the series probability. The mean absolute error among expert responses is 6.7 percentage points. If each expert had instead 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 8.3, 8.6, 9.6, and 9.4 percentage points, respectively. The differences in mean error between each of these heuristics and the actual expert responses are all significant at the $p < 10^{-4}$ level. By contrast, reader responses underperform most of these heuristics. The mean error rate among reader responses is 12.9 percentage points, compared to 10.9 percentage points ($p = 0.001$), 9.7 percentage points ($p < 10^{-4}$), and 10.0 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 14.3 percentage points ($p = 0.032$).

One concern with the comparison between experts and readers is that the two groups 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 B, we show

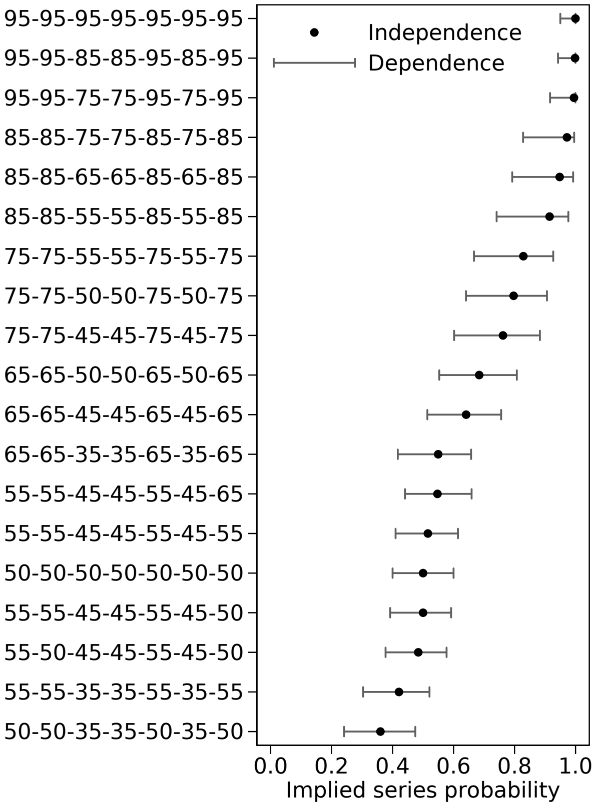
[10] This implication only follows if one sample stochastically dominates the other. If instead, the empirical cumulative distribution functions cross, then both the KS and MW tests may reject the null hypothesis even if it is not clear that one population is more internally consistent than the other. However, in all cases in this paper for which two-sample KS and MW tests reject the null hypothesis, inspection of the empirical cumulative distribution functions reveals first-order stochastically dominance over the vast majority of the range.

that differences in observed consistency cannot be explained by differential selection of the game-by-game probabilities. Specifically, we show that it is impossible to construct a large matched sample—in which the distribution of game-by-game probabilities is equivalent for experts and readers—for which readers demonstrate greater internal consistency than experts.

A second concern 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. 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 C 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 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 have to systematically misevaluate those beliefs in the direction of the implied value under independence. We contend

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.

that a more likely interpretation is that the experts report game-by-game probabilities as if game outcomes are more or less independent events.[11]

Collectively, these results suggest that the experts we study possess expertise in a task that is familiar to them. The next set of studies asks whether that expertise can be applied to an unfamiliar yet conceptually identical task, or equivalently, whether the nature of that

[11] This analysis also shows that differences in dependence structures cannot explain observed differences in consistency between experts and readers, 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 67% for readers.

expertise is a general skill in evaluating summary statistics of complex distributions or is instead context specific.

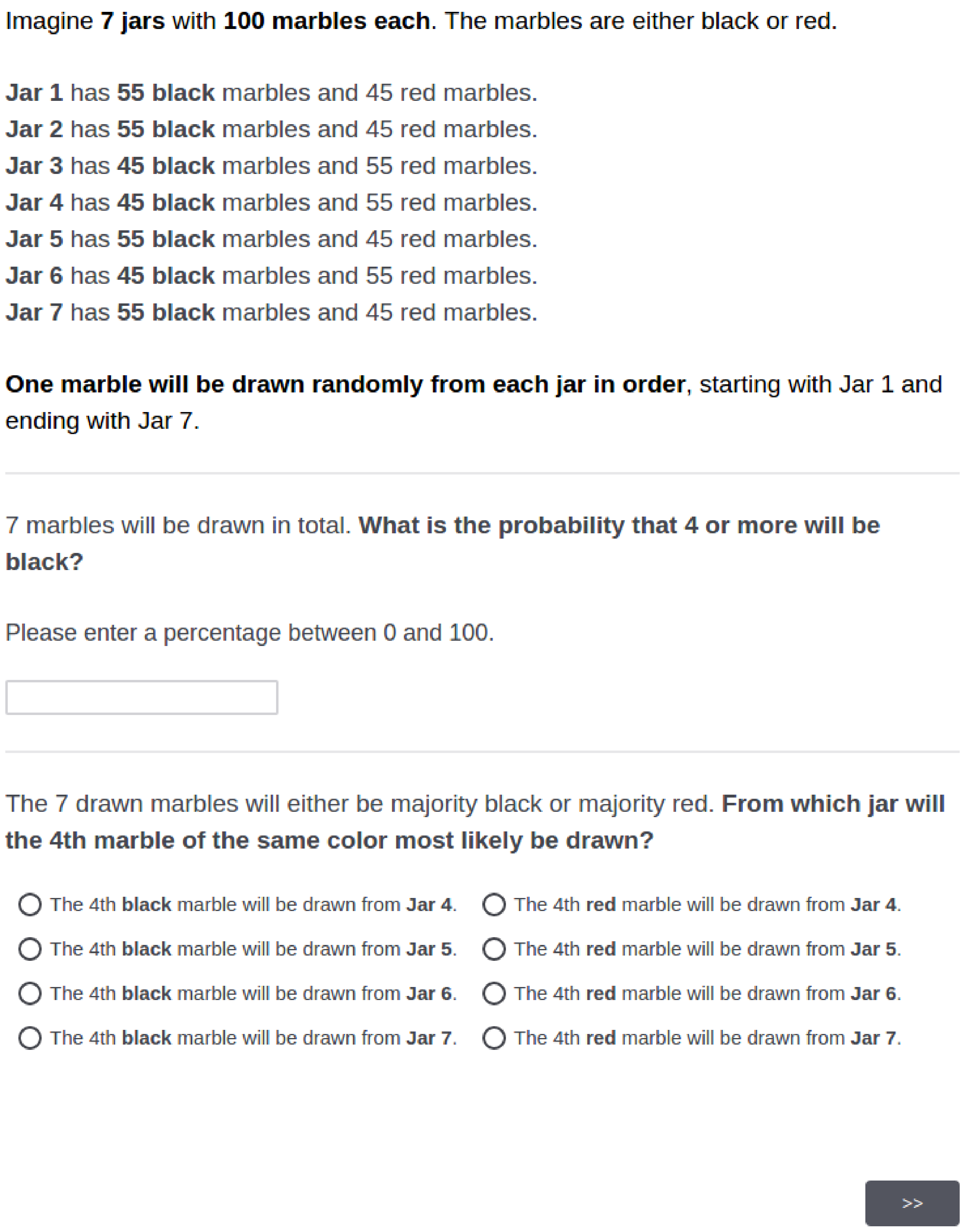
4 Jars and marbles

Six months after the conclusion of the 2015 NBA playoffs, the same ESPN editor asked the expert panel to participate in a survey "designed to improve ESPN's NBA playoffs forecasting...[and] to contribute to academic research."[12] We designed the survey to conceptually mimic the playoff prediction task in an unfamiliar environment. Specifically, the survey 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 from the survey, 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—red in the 4th, 5th, 6th, or 7th jar, or black 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

[12] Pre-registration documents are available at `www.etangreen.com`. Appendix D shows an example of the survey invitation.

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**.

and marbles abstract away from teams and games, and participants observe sequences of proportions instead of reporting game-by-game probabilities.

Respondents completed a personalized survey comprising five sets of jars-and-marbles problems, with each set on a separate page. The first four 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.[13] Matching marble proportions to game-by-game probabilities allows us to compare performance by the same expert on a formally identical problem. The fifth page showed the same sequence to all respondents: 7 jars containing 95 black marbles (and 5 red marbles) each. This sequence measures conceptual understanding at the end of the survey. The first four pages were randomly ordered, and respondents could not navigate to previously completed pages.

We designed this survey 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.[14] 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.[15] However, this performance gap may be

[13] When 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.

[14] The sequence comprised 65 black marbles in jars 1, 2, 5, and 7, and 45 black marbles in jars 3, 4, and 6; respondents also evaluated two other sequences with marble proportions that did not mimic game-by-game probabilities for a typical NBA playoff series.

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

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, this 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.[16]

Of the 119 experts solicited, 44 reported estimates for at least one 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 survey, and 98 participated in some capacity. Moreover, 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.[17] For our initial results, in which we compare performance on matched sequences across domains, we analyze the 143 matched sequences from all 44 study participants. However, 13 of these participants did not complete four matched sequences. 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 four matched sequences. All p-values for mean comparisons reported in this section are from two-way tests with standard errors clustered by respondent.

Low motivation, a concern in many laboratory experiments, does not appear to plague our study. The median completion time is 9 minutes and 7 seconds, or almost 2 minutes per

[16] The pilot study raises a second concern—that performing multiple studies on the same population may corrupt responses in later studies. In particular, practice with problems of the same format should enhance subsequent performance, rather than diminish it. As a result, this dynamic suggests that our findings of poor performance on the second study are conservative. Without practice on a pilot study, the experts presumably would have performed even worse.

[17] For 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 ($p = 0.687$). Note that for survey participants, these figures reflect performance on all NBA responses, not just matched sequences.

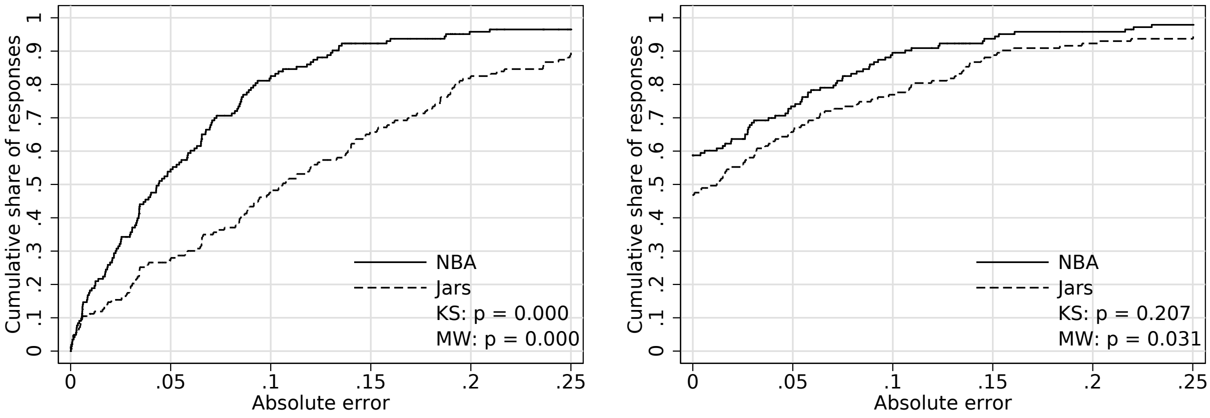
page, and the minimum completion time is 3 minutes and 23 seconds.[18] Error rates for those who finish the survey faster than the median completion time and slower than the median are statistically indistinguishable, suggesting that the fastest finishers do not greatly sacrifice accuracy.[19] Also indicative of attentiveness, respondents round their series probabilities to the nearest 5% *less* often when evaluating jars and marbles (rounding on 55% of responses) than when predicting NBA playoff outcomes (68%), a statistically significant difference ($p = 0.028$) Finally, one subject appeared to solve the problems either mathematically or by simulation, taking more than 20 minutes and correctly answering each question to the nearest percentage point 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.

Despite their apparent attentiveness, the experts are generally unable to apply their expertise in the unfamiliar study setting. 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% ($p < 0.001$). 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

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

[19] For 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 ($p = 0.597$).

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



(a) Series probability

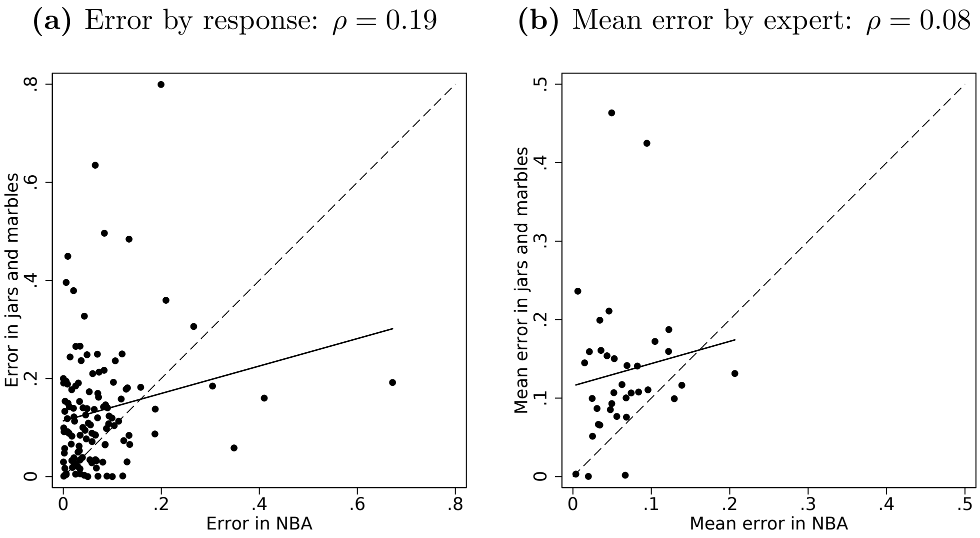
(b) most likely series outcome

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.

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. Moreover, they fail to solve problems that are formally identical to those they solved before.

Unfamiliarity not only erodes performance—it decouples performance between the familiar and unfamiliar domains. Figure 5 shows scatter plots of absolute error in series

Figure 5: Series probability error rates across domains.


(a) Error by response: $\rho = 0.19$

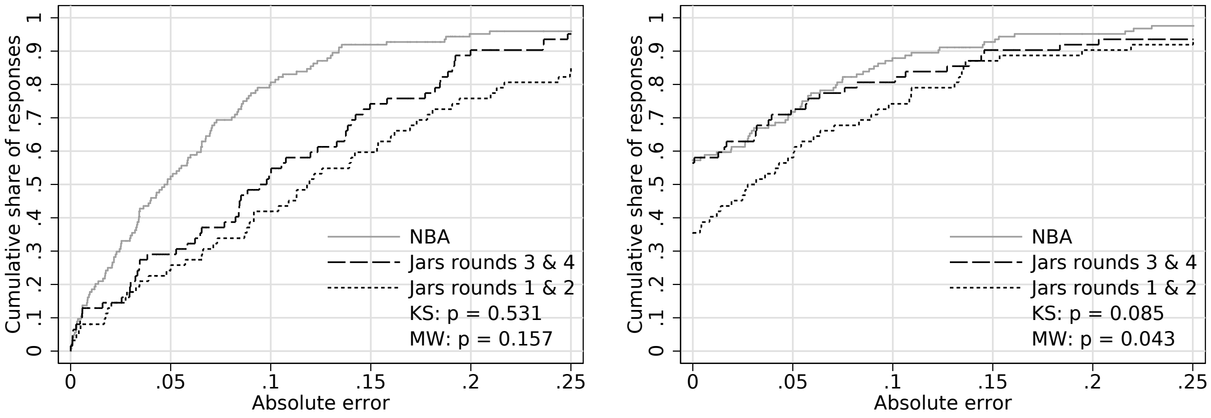
(b) Mean error by expert: $\rho = 0.08$

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.

probability judgements by context; Figure 5a compares errors for each matched response, and Figure 5b compares mean error rates by respondent. Performance in one context poorly predicts performance in the other: errors in series probability judgements are correlated at 0.19 by response and 0.08 by respondent. The decoupling is similarly severe for evaluations of most likely series outcomes. Correctness in these judgements 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.

We also find that exposure makes the abstract domain more familiar: 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

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



(a) Series probability

(b) most likely series outcome

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.

figures, the error distributions for the third and fourth rounds stochastically dominate the error distributions for the first and seconds 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, has a series probability of 99.98%—i.e., drawing 4 or more black marbles is virtually guaranteed.[20] 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%. (Experts report the corresponding sequence of game-by-game probabilities 4 times for NBA playoff series, and each time they select a series probability of 99% or 100%.) Two possible interpretations remain. Either the experts learn to apply their expertise to the formally identical problem, or they 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.[21] We match initial performance on the marbles exercise between experts and non-experts by calibrating the incentives offered to non-experts.[22] 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

[20] For 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."

[21] Pre-registration documents are available at `www.etangreen.com`.

[22] We 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.

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.[23] We advertised a survey about probabilistic judgement 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.[24] Subjects were not given any further guidance about how long the survey would take to complete or how the bonus would be calculated. We replicated the 31 surveys for which experts completed 4 matched sequences, and we randomly assigned non-experts to these surveys. For each of the 31 surveys, we analyze the first 4 responses from non-experts who take at least 2 minutes to complete the survey,[25] 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.

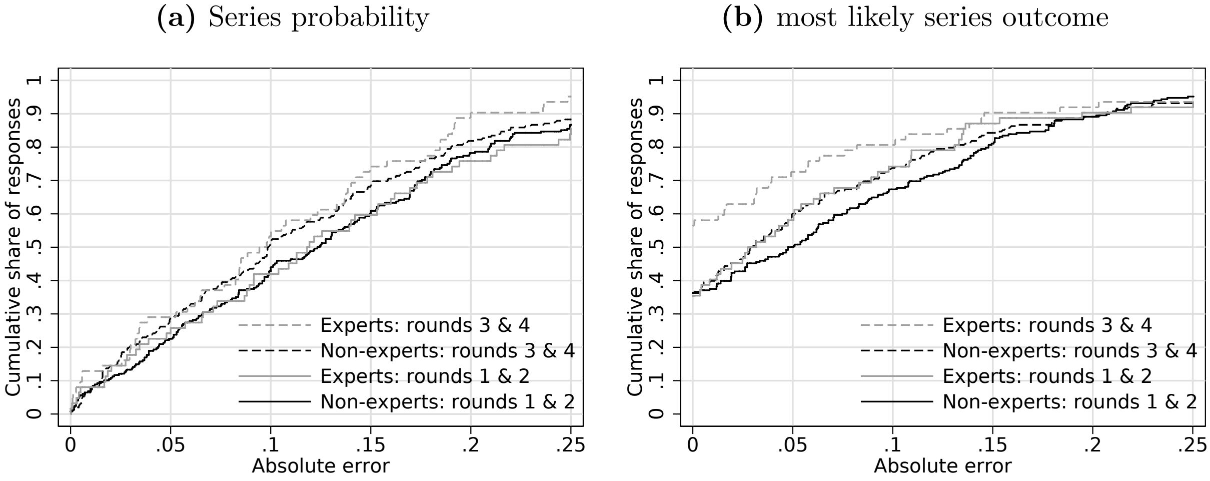
Figure 7 shows cumulative distributions of absolute error for non-experts, separately for the first two sequences (solid) and the last two sequences (dashed), and separately for the series probability (7a) and most likely series outcome (7b); for comparison, we superimpose the expert results from Figure 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

[23] Appendix 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).

[24] The actual average payment was \$1.70.

[25] Despite 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 criterion in pre-registration.

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



(a) Series probability

(b) most likely series outcome

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.

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: $p = 0.458$). However, experts 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 ($p = 0.008$). 65% of series probabilities reported by non-experts are rounded to the nearest 5 percentage points, compared to 53% for expert responses ($p = 0.013$). 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. These comparisons suggest that while experts and non-experts are closely matched on initial performance, they differ along dimensions other than expertise.

Second, we compare rates of improvement for experts and non-experts. 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 ($p = 0.271$). Divergence between the populations is more stark for the most likely series outcome. 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 ($p = 0.023$). We interpret this result to imply that expertise in a formally identical problem accelerates learning in an unfamiliar setting.

5 Conclusion

This paper shows that unfamiliarity restricts the portability of expertise. We study experts who are adept at making difficult probabilistic judgements on a familiar task. And we find that these experts fail to apply their expertise on a formally identical but contextually distinct problem. Their expertise is not conceptual mastery but something more intuitive and context dependent. When the problem is made unfamiliar, their expertise fades.

We make the problem unfamiliar by manipulating the contextual cues that experts observe. This is akin to a worker who takes a similar job at a different firm or in a different industry—for instance, a buyer in a department store who moves from housewares to jewelry, or an agent who negotiates first on behalf of musicians and then on behalf of athletes. Our results suggest that such actors develop expertise intuitively, and then fail to apply that expertise in unfamiliar, but conceptually analogous, environments.

Problems can be made unfamiliar in other ways—for instance, when parameters change. A number of laboratory experiments show that learning does not necessarily imply conceptual mastery. In the repeated common-value auctions studied by Kagel and Levin (1986) and described earlier, subjects learn to avoid the winner's curse with practice, only to experience it again when the number of bidders changes. In a similar vein, Neelin, Sonnenschein and Spiegel (1988) observe sequential bargaining games with discounting,[26] in which subjects play a 2-round game, followed by a 3-round game, followed by a 5-round game. After a practice game, subjects make the subgame-perfect first offer in the initial 2-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. In our experiment, experts improve with repetition of the jars-and-marbles problem. Yet when shown a final version with unfamiliar values, they reveal a conceptual misunderstanding of the binomial distribution. Evidently, slight changes in context or cues can deprive an expert of the benefits of her expertise.

[26] In these games, two players exchange offers over how to split a shrinking pie. In the first round, one player offers a split. If the second player rejects the split, she makes a offers a split of a smaller sum. This sequence alternates until either one player accepts an offer, in which case the pie is divided as agreed to, or the pie disappears completely.

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.

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.

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.

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.

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.

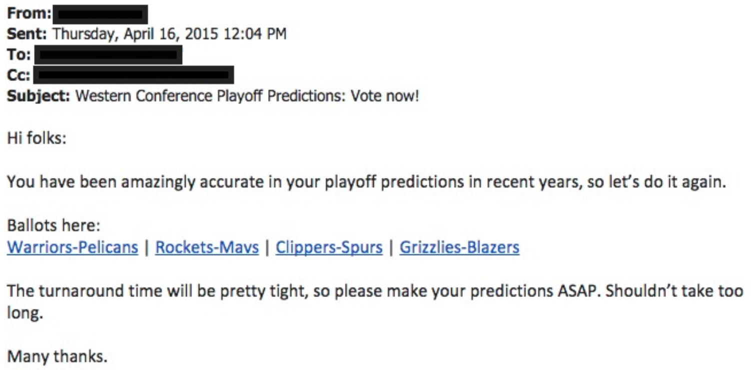
Tversky, Amos and Daniel Kahneman (1983) "Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment.," *Psychological review*, Vol. 90, p. 293.

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:

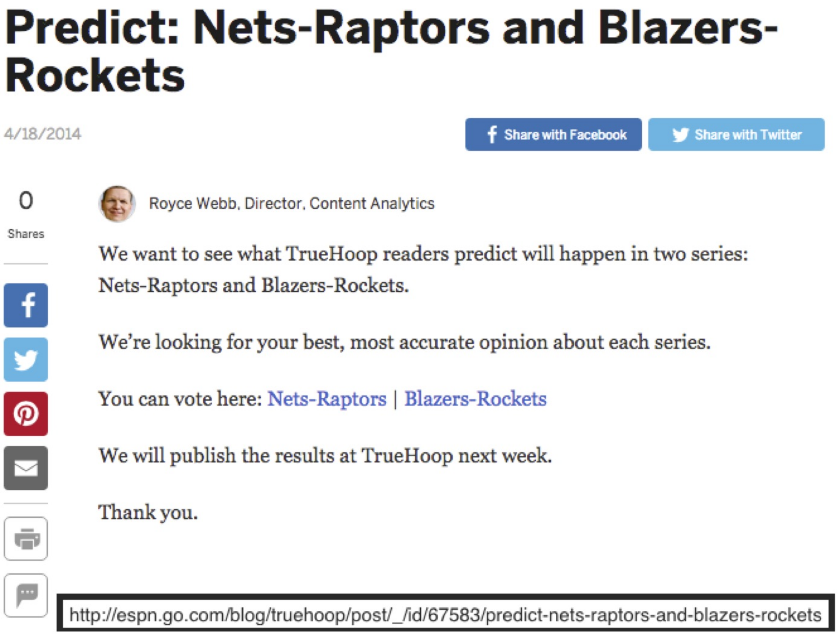
You have been amazingly accurate in your playoff predictions in recent years, so let's do it again.

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

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.

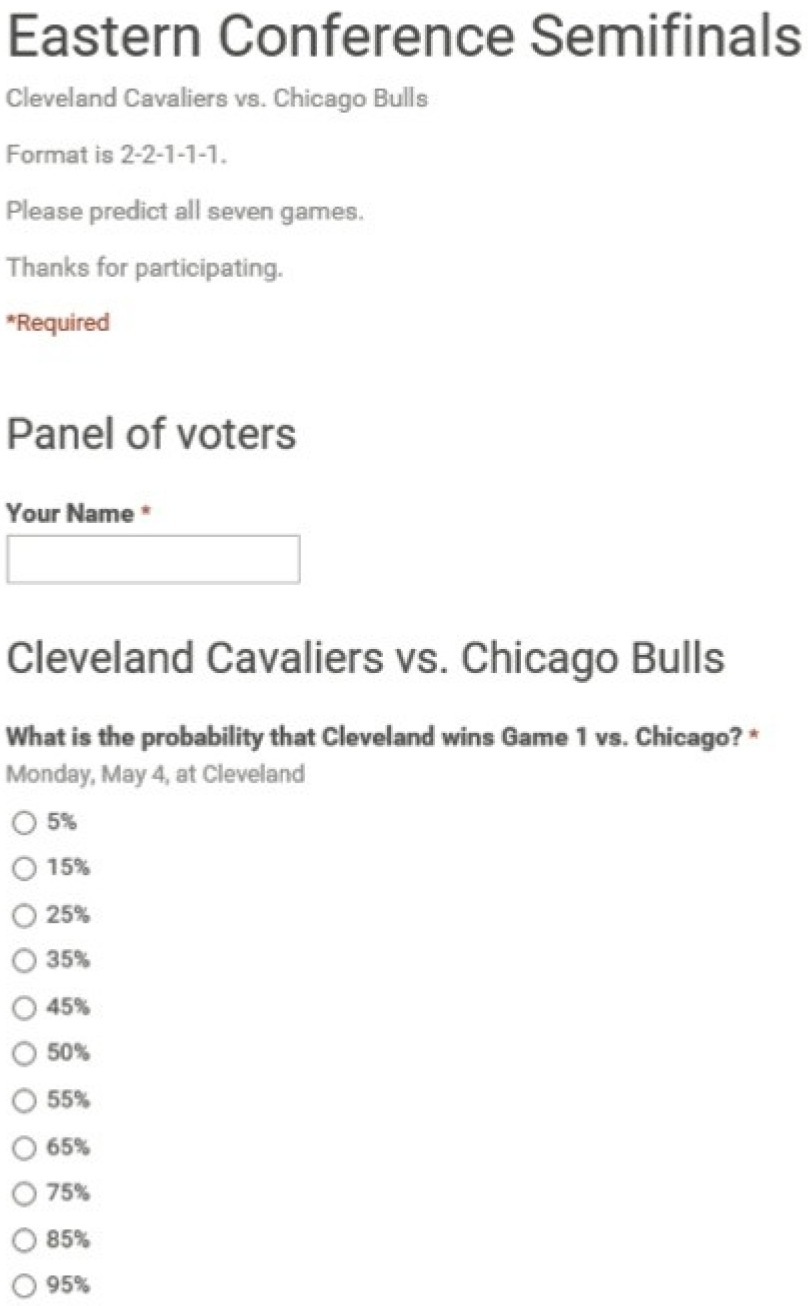
You can vote here: Nets-Raptors | Blazers-Rockets

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%

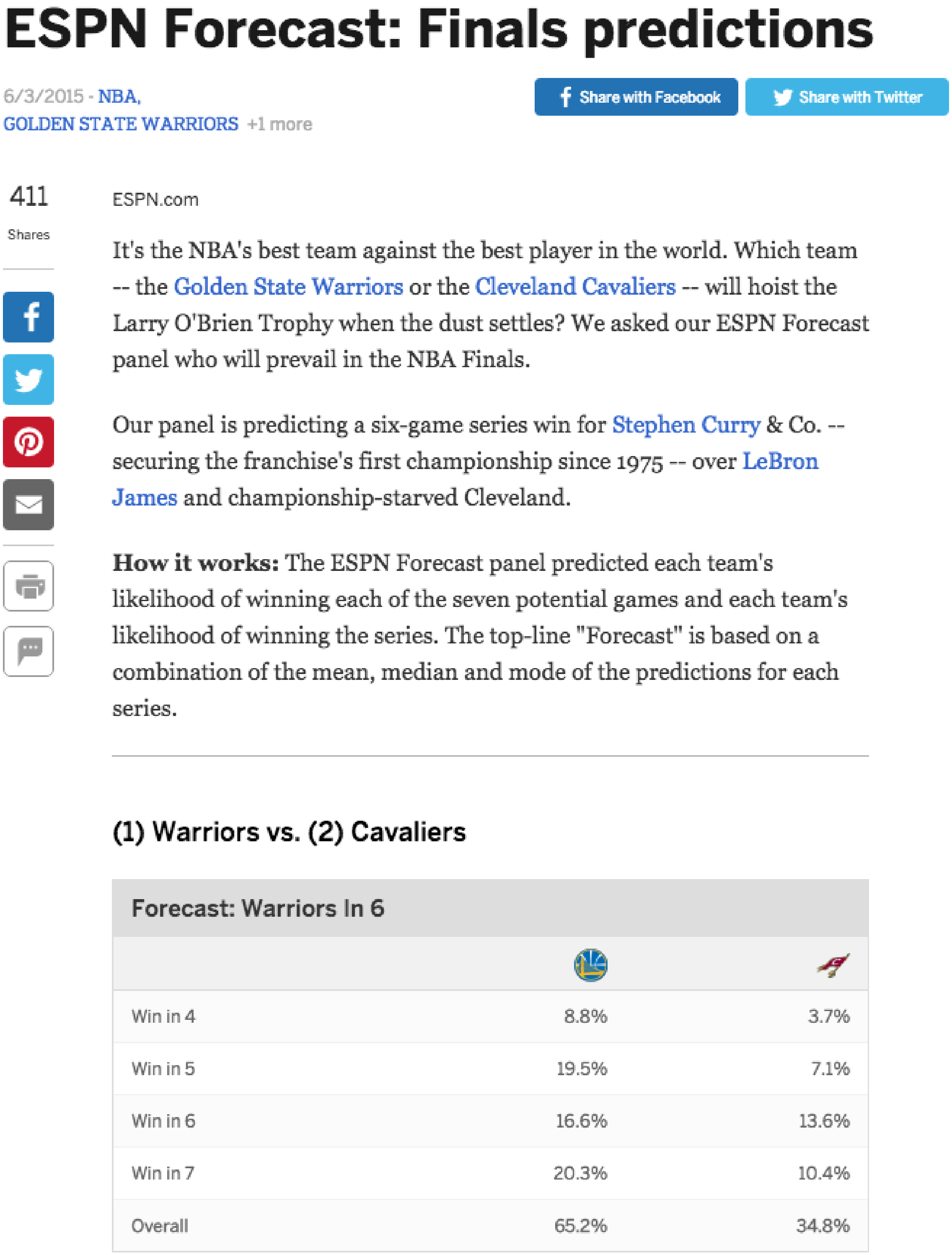
○ 65%

○ 75%

○ 85%

○ 95%

Figure 11: Expert forecast published on ESPN.com



ESPN Forecast: Finals predictions

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

ESPN.com

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 Selection of probabilities

One 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 series probabilities and most likely series outcomes 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.[27] While there are more than half a billion combinations of expert and reader responses that produce such a sample, there are zero such samples—for either the series probability or the most likely series outcome—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.[28]

A related concern is that experts may more frequently than readers report summary statistics near the midlines of their respective ranges—i.e., 50% for series probabilities and 4-3 or 3-4 for most likely series outcomes—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

[27] For the series probability (most likely series outcome), we restrict the full set of responses to the 20 (24) sequences that experts and readers each report at least once. Of these, 11 (14) have an identical number of expert and reader responses. We sample among the remaining 9 (10) sequences, in each sample selecting from the panel with more responses for that sequence. Each matched sample contains 24 (30) responses for sequences that are matched perfectly in the data and 38 (30) responses for sequences that sampled from the matched sequences. That is, each matched sample contains 62 (60) responses—31 (30) from experts and 31 (30) from readers—43 (45) of which are identical across samples, and 19 (15) of which vary.

[28] This 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.

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%.[29] Distributions of reported most likely series outcomes are also similar for experts and readers. Experts report the median outcomes of 4-3 and 3-4 in 33% of responses, compared to 28% for readers ($p = 0.02$). However, experts report extreme outcomes of 4-0, 4-1, 1-4, and 0-4 more frequently than readers—30% to 18% ($p < 10^{-4}$)—contrary to the supposition that experts more greatly limit their potential for inconsistency.

C Unobserved dependence

A second 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

[29] The p-value of the difference between these values is 0.32.

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.[30]

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 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 $(p_1, p_{2|1\text{-}0}, p_{2|0\text{-}1}, \dots, p_{6|3\text{-}2}, p_{6|2\text{-}3}, p_7)$, 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, $p_2 = p_1 \cdot p_{2|1\text{-}0} + (1 - p_1) \cdot p_{2|0\text{-}1}$, 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 $p_1 \cdot p_{2|1\text{-}0} \cdot p_{3|2\text{-}0} \cdot p_{4|3\text{-}0}$. From

[30] Note 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).

these series outcomes, exact summary statistics can be calculated.

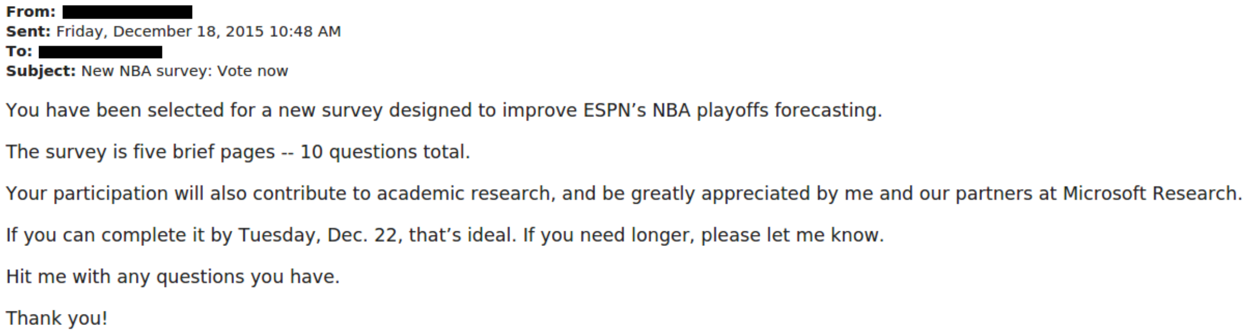
Since we do not observe the parameters of the dependence structure—i.e., the conditional probabilities—we cannot identify the exact summary statistics implied by such a structure. However, the reported game-by-game probabilities constrain the values which the conditional probabilities can take, and these 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:[31]

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

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 $p_{2|1\text{-}0} = 1, p_{3|2\text{-}0} = 1, p_{4|3\text{-}0} = 1$, and all other conditional probabilities equal 0.

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.

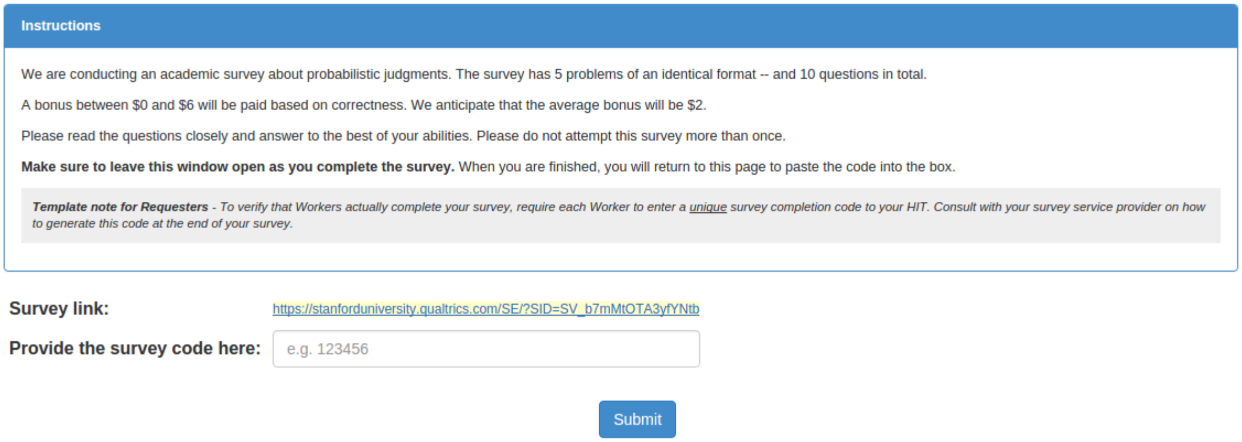
If you can complete it by Tuesday, Dec. 22, that's ideal. If you need longer, please let me know.

Hit me with any questions you have.

Thank you!

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

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_b7mMtOTA3yfYNtb

Provide the survey code here: e.g. 123456