Current Directions in Psychological Science
2014, Vol. 23(4) 290–295

DOI: 10.1177/0963721414534257
cdps.sagepub.com

Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate

Philip E. Tetlock, Barbara A. Mellers, Nick Rohrbaugh, and Eva Chen
University of Pennsylvania

Abstract

Forecasting tournaments are level-playing-field competitions that reveal which individuals, teams, or algorithms generate more accurate probability estimates on which topics. This article describes a massive geopolitical tournament that tested clashing views on the feasibility of improving judgmental accuracy and on the best methods of doing so. The tournament's winner, the Good Judgment Project, outperformed the simple average of the crowd by (a) designing new forms of cognitive-debiasing training, (b) incentivizing rigorous thinking in teams and prediction markets, (c) skimming top talent into elite collaborative teams of "super forecasters," and (d) fine-tuning aggregation algorithms for distilling greater wisdom from crowds. Tournaments have the potential to open closed minds and increase assertion-to-evidence ratios in polarized scientific and policy debates.

Keywords

forecasting, tournaments, probability, accuracy

When we scratch beneath the rhetorical veneer of high-stakes policy debates, we find clashing predictions about the perils and promises of competing policy proposals—how should central bankers balance the risks of pushing interest rates either too low or too high, or how should arms-control negotiators avoid being either too conciliatory or too confrontational? And when we scratch beneath the predictions, we often find elaborate theories of human nature that are rarely rattled by apparent prediction failures (Jervis, 1976, 2010; MacCoun, 1998; Tetlock, 2005).

The key word, however, is "apparent." It is difficult to falsify vague-verbiage predictions such as, "All else equal, if we go down path *x*, this outcome might happen." Outside the lab, ceteris paribus is never satisfied. And "might happen" could mean a probability as low as .1 or as high as .9, depending on context (Wallsten, Budescu, & Zwick, 1993). Moreover, forecasters in trouble always have recourse to classic belief-system defenses, such as the close-call-counterfactual and off-on-timing defenses: the unexpected event either almost happened or will happen eventually (Tetlock, 2005; Tetlock & Mellers, 2011).

Forecasting tournaments are potential epistemic game changers because they let us assess the track records of proponents of clashing views with far more precision than is normally possible. If proponents of one policy assign probabilities palpably closer to reality than do proponents of another, the losers are under pressure to either acknowledge mistakes or trivialize them. If the losers continue losing in later rounds, the pressure to reassess builds. And the pressure builds faster in a tournament that requires transparent knowledge claims than it does in the real world, which offers lots of rhetorical cover for concealing forecasting failures.

This article explores how both scientists and policymakers can use tournaments to test lab-based theories in real-world settings and render intractable debates tractable. We have divided this article into two sections. The first covers key psychological lessons learned thus far from first-generation tournaments, and the second

Corresponding Author:
Philip E. Tetlock, 3720 Walnut St., Solomon Labs, University of Pennsylvania, Philadelphia, PA 19104
E-mail: tetlock@wharton.upenn.edu

explores the full potential of the next generation of tournaments to test competing views of human nature underlying policy debates.

I. Lessons From First-Generation Tournaments

Tournaments are sometimes seen as gimmicky contests that reward lucky winners—and consign losers to undeserved ignominy. Properly designed, however, tournaments are powerful tools for answering otherwise impossible-to-answer questions—about issues ranging from the evolution of cooperation (Axelrod, 2006) to optimal strategies of judgment and choice under uncertainty (Armstrong, 2001; Erev et al., 2010; Makridakis et al., 1982).

In 2005, Tetlock reported the first large-scale geopolitical forecasting exercise in *Expert Political Judgment: How Good Is It? How Can We Know?* Tetlock used psychological methods to assess the accuracy of experts' probability judgments about a wide array of events, from the stability of multi-ethnic states to the growth prospects of emerging markets to the risks of interstate conflict. He elicited about 29,000 predictions from 284 subject-matter experts and found support for the real-world robustness of several psychological phenomena, including overconfidence, hindsight bias, self-serving biases in counterfactual reasoning, and the difficulty of outperforming even simple statistical models. For instance, experts who scored high on need for closure were especially likely to exaggerate the degree to which they performed above chance level on longer-range forecasts inside their domain of expertise—and to balk at acknowledging mistakes.

Tetlock's (2005) study focused on judgmental shortcomings and is often cited for showing not only that experts know less than they think they do, but also that their predictions often cannot outperform random guessing (Gardner, 2011). But what researchers find is often a function of how hard they look, and the study was not designed to incentivize optimal performance. There was no formal competition with public winners and losers. Indeed, to secure cooperation from wary professionals, Tetlock had to promise anonymity. The winners and losers were abstractions—for instance, the lower-need-for-closure "foxes" outperformed the higher-need-for-closure "hedgehogs."

By contrast, the Intelligence Advanced Research Projects Activity (IARPA) tournament of 2011 through 2013, sponsored by the U.S. intelligence community, was a true tournament and, unlike most research on judgment and choice (Baron, 2000), it focused on optimal, not typical, performance. The official goal was to identify which of five competing research programs could generate the most accurate probability judgments. Each program conducted its own mini-tournament and was free to develop its own methods for best sampling, eliciting, and aggregating forecasts. The official metric was cumulative Brier scores (the square deviation between the forecast and the outcome, scored as 0 for nonoccurrence and 1 for occurrence; Brier, 1950), across time and across more than 200 questions selected by the intelligence community (questions on topics ranging from Brent crude oil prices to Sino-Japanese clashes in the East China Sea to leadership turnover in Russia and Zimbabwe). The winner was whoever invented the fastest methods of pushing probabilities toward .0 for things that did not happen and toward 1.0 for things that did happen, without triggering steep penalties for false positives (too close to 1.0 for nonevents) and false negatives (too close to .0 for events).

The Good Judgment Project (GJP)[1] won the IARPA tournament: Its best wisdom-of-the-crowd algorithms were on the right side of 50/50 on 86.2% of all daily forecasts, outperforming the simple average of the control group (forecasters randomly assigned to a working-alone, no-training condition) by 60% and other teams by 40%. The tournament was not, however, just a horse race. GJP randomly assigned its forecasters to cells in factorial designs that tested hypotheses about the psychological drivers of accuracy. We discovered four such drivers: (a) recruitment and retention of better forecasters (accounting for roughly 10% of the advantage of GJP forecasters over those in other research programs); (b) cognitive-debiasing training (accounting for about a 10% advantage of the training condition over the no-training condition); (c) more engaging work environments, in the form of collaborative teamwork and prediction markets (accounting for a roughly 10% boost relative to forecasters working alone); and (d) better statistical methods of distilling the wisdom of the crowd—and winnowing out the madness (the log-odds-extremizing algorithm of Satopää, Baron, et al., 2014, Satopää, Jensen, Mellers, Tetlock, & Ungar, in press, and Baron, Ungar, Mellers, and Tetlock, 2014, which contributed an additional 35% boost above unweighted averaging of forecasts).

GJP also added a controversial twist to its winning strategy. It created "super-forecaster" teams by skimming off the top 2% of forecasters each year of the tournament and assigning them to elite teams. We say "controversial" because GJP informally surveyed experts and found flatly contradictory opinions on the wisdom of this strategy, from the bearish "Expect nothing. Your lucky 'supers' will soon regress toward the mean" (e.g., in the spirit of Hartzmark, 1991) and "The 'super' label will swell their heads" (e.g., Levitt & March, 1988) to the bullish "Expect good things. The best predictors of future performance are past performance and IQ—and your supers have both factors going for them" (e.g., in the spirit of Hunter

& Hunter, 1984) and "Supers will also get a self-fulfilling-prophecy boost—and derive the benefits that tracking confers on high-ability students" (i.e., stimulation from peers; e.g., Betts & Shkolnik, 2000).

The experts were divided, but the data were unequivocal: Super forecasters performed superbly. Averaged forecasts of GJP's super forecasters (five teams of 12 forecasters each) in Year 2 handily beat the Brier-score goals that the IARPA set for Year 4, and all other research programs. They showed no regression toward the mean from one year to the next, and they improved on all the standard psychometric indices of judgmental accuracy, including calibration, discrimination, and area under the curve (Mellers, Ungar, et al., 2014).

Although the superior performance of "supers" has been established and re-established, the "How super are supers?" debate continues. We plan experiments in future tournaments to resolve questions such as, how well would regular forecasters have done if they had received the classic self-fulfilling-prophecy designation of "late bloomers" (Rosenthal, 1966)? How well would supers have done if they had not been labeled "super"? Can some supers transcend Kahneman's (2011) System 1 biases, or are they simply adept at System 2 self-correction?

Finally, we should not assume that the winners of IARPA-style tournaments will also be the biggest net contributors to scientific knowledge. The other programs made important discoveries, including (a) Budescu and Chen's (in press) development of a method of distinguishing forecasters who free ride from those who contribute to team performance; (b) Steyvers, Wallsten, Merkle, and Turner's (2014) proposal that the area under the curve is a better metric than the Brier score for evaluating forecasts, because it makes only ordinal assumptions about probability scales—and allows estimation of false-positive rates that decision makers can use to estimate the expected utility of response options; and (c) the demonstration that aggregation of individual judgments works best when it follows recalibration designed to correct for biases such as under- and overconfidence (Turner, Steyvers, Merkle, Budescu, & Wallsten, 2013) and violations of additivity (Karvetski, Olson, Mandel, & Twardy, 2013).

II. Can Tournaments Transform Reflexive Opponents Into Flexible Perspective Takers?

Tournaments let us test the clashing hunches that psychologists have about judgmental accuracy. But tournaments also raise a deeper epistemic question: When should we test theories, and when theorists? The distinction is well defined in highly mathematized fields in which it is easy to see when theorists introduce fudge factors to absorb awkward facts. But the distinction is fuzzier in psychology, where theorists have lots of wiggle room in deciding what counts as a fair test because their theories take the form of logically entwined bundles of natural-language propositions, and researchers have great flexibility in operationalizing constructs (Cronbach, 1986; Tetlock & Manstead, 1985) and analyzing data (Simmons, Nelson, & Simonsohn, 2011).

Tournaments are the light at the end of this dark tunnel. They can be structured not only to incentivize eclectic innovation (the goal of IARPA-style tournaments) but to facilitate Kahneman-style adversarial collaborations in which clashing camps can make "reputational bets" on empirical outcomes in well-defined laboratory or natural settings. For instance, defenders and critics of the concept of general cognitive ability could, in principle, have agreed—before seeing any IARPA data—to adjust confidence in their positions if their clashing expectations about the predictive power of the Ravens Advanced Progressive Matrices were or were not borne out. (Thus far, defenders have the edge; Mellers, Stone, et al., 2014.) Similar competitions could have been organized around a variety of other psychological debates, such as the following:

1. What roles do cross-situational cognitive styles, such as actively open-minded thinking and need for cognition, play? (Thus far, they have been shown to play a significant one; Mellers, Stone, et al., 2014.)
2. Is it better for forecasters to work separately—and garner the benefits of independence (e.g., canceling random errors)—or to work interactively and learn from each other? (Thus far, the answer is "interactively"; Mellers, Ungar, et al., 2014.)
3. If interactive working conditions are better, then is it better for forecasters to work in collaborative teams or in prediction markets? (Thus far, both have been shown to help—roughly equally; Atanasov et al., 2014.)
4. How feasible is it to teach people to avoid well-known cognitive biases—and to become better probability assessors? (Thus far, the payoffs from training have been shown to be substantial; Mellers, Ungar, et al., 2014; Stone & Luu, 2013.)
5. Which poses the greater threat to accuracy: excessive conservatism or excessive volatility (under- vs. overreactions to news)? (Thus far, both have been shown to be sources of error, with conservatism posing a bigger problem; Atanasov et al., 2014.)

Tournaments can also be used to broker debates on higher-stakes policy issues. For instance, psychologists disagree on both the laboratory and real-world predictive power of unconscious racial biases (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). The policy stakes are high because, in the courtroom, expert witnesses have argued that automatic biases are so pervasive and potent that it is virtually impossible for organizations to comply with equal-employment-opportunity laws without holding managers accountable for achieving numerical goals (Bielby, 2000, 2008). The standard equal-employment-opportunity forms of accountability are, in this view, insufficient. To resolve the dispute, Tetlock and Mitchell (2009) proposed designing lab and field tests of the durability of automatic biases under varying accountability pressures—and eliciting ex ante reputational bets from opposing camps about effect sizes.

This approach to scientific-dispute resolution is strikingly similar to the adversarial-collaboration model of Kahneman and Klein (2009) and Mellers, Hertwig, and Kahneman (2001), and it could be adapted to a wide range of disputes. For instance, the Open Science Collaboration (2012) is exploring the replicability of findings in top psychological journals, with the goal of identifying which features of studies predict which results hold up. The timing of this initiative is not coincidental: It parallels severe critiques of data-analytic practices in social psychology, from *p* hacking to outright fraud (Simmons et al., 2011).

Tournaments put the Open Science Collaboration in a fresh light. Scientists talk a lot about good methodological judgment, but they rarely try to quantify it. Tournaments allow us to assess how well experts can predict replicability relative to regression models that weight objective features of studies (e.g., social cognition vs. pure cognition; sample sizes and between- vs. within-subjects designs; perceived political spin on results). Past work has suggested that the humans will be overtaken by statistical models, even models that simply mimic the cognitive strategies of human forecasters (Dawes, 1998). Regardless of outcome, however, tournaments will speed the process of figuring out how reliable our collective databases are—and which theoretical camps were right about which strengths or weaknesses.

We suspect that when participants in polarized debates begin playing in competitive tournaments, they will learn to make more circumspect claims. The prospect of imminent falsification looms large in tournaments—creating a form of public accountability likely to suppress self-justification and stimulate self-critical thinking (Lerner & Tetlock, 1999). And when the evidence begins to break decisively in favor of one side or the other, tournaments should accelerate depolarization of probability judgments. Tournaments impose reputational penalties on bad Bayesian updaters, which may be why Moore et al. (2014) found so little overconfidence among GJP forecasters (a level comparable to that of weather forecasters, who are often upheld as paragons of good judgment).

Closing Thoughts

Tournaments can nudge players in polarized debates toward the right epistemic direction. Imagine a world in which the comment that "someone has good psychological intuitions" is more than a vague hunch but is instead based on objective performance in reputational bets against prominent psychologists in major controversies.

This scenario is realizable. Prediction-market and tournament-design specialists know how to set up such contests (Hanson, 2007; Tetlock, 2005), and we urge the Association for Psychological Science to sponsor such competitions. Such sponsorship would be timely. It would signal recognition that our peer-review systems have shortcomings and that we are trying to offset them by incentivizing clashing schools of thought to relentlessly second-guess each other's knowledge claims.

Recommended Reading

Gardner, D. (2011). (See References). Offers compelling contemporary and historical examples illustrating how it is possible for charlatans to flourish when we fail to systematically monitor judgmental accuracy.

Kahneman, D. (2011). (See References). Provides an elegant overview of the heuristics-and-biases research program, written by arguably the most influential psychologist of the last century.

Tetlock, P. E. (2005). (See References). Reports the first large-scale effort to use psychological theories and methods to study judgmental accuracy in world politics.

Tetlock, P. E. (2009, September/October). Reading tarot on K Street. *The National Interest*, 57–67. Makes the case for keeping objective score on political pundits and consultants promising to improve the quality of judgment.

Tetlock, P. E., & Mellers, B. A. (2011). (See References). Explores how difficult it is for people and organizations to commit to improving judgmental accuracy when they are caught in the middle of accountability blame games.

Acknowledgments

The U.S. Government is authorized to reproduce reprints for government purposes notwithstanding any copyright annotation. The views expressed herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Intelligence Advanced Research Projects Activity, the Department of the Interior National Business Center, or the U.S. Government.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by Research Contract D11PC20061 to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity via the Department of Interior National Business Center.

Note

1. To learn more about the Good Judgment Project and participate if so inclined, visit goodjudgmentproject.com.

References

Armstrong, J. S. (Ed.). (2001). *Principles of forecasting: A handbook for researchers and practitioners*. New York, NY: Springer.

Atanasov, P., Rescober, P., Stone, E., Swift, S., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2014). *Crowd forecasting with prediction polls versus prediction markets*. Manuscript in preparation.

Axelrod, R. M. (2006). *The evolution of cooperation*. New York, NY: Basic Books.

Baron, J. (2000). *Thinking and deciding*. New York, NY: Cambridge University Press.

Baron, J., Ungar, L. H., Mellers, B. A., & Tetlock, P. E. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*. Advance online publication. doi: 10.1287/deca.2014.0293

Betts, J. R., & Shkolnik, J. L. (2000). Key difficulties in identifying the effects of ability grouping on student achievement. *Economics of Education Review*, *19*, 21–26.

Bielby, W. T. (2000). Minimizing workplace gender and racial bias. *Contemporary Sociology*, *29*, 120–129.

Bielby, W. T. (2008). Promoting racial diversity at work: Challenges and solutions. In A. P. Brief (Ed.), *Diversity at work* (pp. 53–86). Cambridge, England: Cambridge University Press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Budescu, D., & Chen, E. (in press). Identifying expertise to extract the wisdom of crowds. *Management Science*. Advanced online publication. doi: http://dx.doi.org/10.1287/mnsc.2014.1909

Cronbach, L. J. (1986). Social inquiry by and for earthlings. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science: Pluralisms and subjectivities* (pp. 83–107). Chicago, IL: University of Chicago Press.

Dawes, R. M. (1998). Behavioral decision making and judgment. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vols. 1–2, 4th ed., pp. 497–548), New York, NY: McGraw-Hill.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., & Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*, 15–47.

Gardner, D. (2011). *Future babble: Why expert predictions are next to worthless, and you can do better*. New York, NY: Dutton.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17–41.

Hanson, R. (2007). Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets*, *1*, 3–15.

Hartzmark, M. L. (1991). Luck versus forecast ability: Determinants of trader performance in futures markets. *The Journal of Business*, *64*, 49–74.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72–98.

Jervis, R. (1976). *Perception and misperception in international politics* (Vol. 49). Princeton, NJ: Princeton University Press.

Jervis, R. (2010). *Why intelligence fails: Lessons from the Iranian Revolution and the Iraq War*. Ithaca, NY: Cornell University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. London, England: Macmillan.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*, 515–526.

Karvetski, C. W., Olson, K. C., Mandel, D. R., & Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis*, *10*, 305–326.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*, 255–275.

Levitt, B., & March, J. G. (1988). Organizational learning. *Annual Review of Sociology*, *14*, 319–340.

MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology*, *49*, 259–287.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*, 111–153.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269–275.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., . . . Tetlock, P. (2014). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. Manuscript submitted for publication.

Mellers, B. A., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . Tetlock, P. (2014). Psychological strategies for winning geopolitical forecasting tournaments. *Psychological Science*. Advance online publication. doi:10.1177/0956797614524255

Moore, D. A., Swift, S. A., Minster, A., Mellers, B. A., Ungar, L., Tetlock, P. E., . . . Tenney, E. R. (2014). *Confidence and accuracy over time*. Manuscript in preparation.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*, 171–192.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. East Norwalk, CT: Appleton-Century-Crofts.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*, 344–356.

Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (in press). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *Annals of Applied Statistics*.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Steyvers, M., Wallsten, T. S., Merkle, E. C., & Turner, B. M. (2014). Evaluating probabilistic forecasts with Bayesian signal detection models. *Risk Analysis*, *34*, 435–452.

Stone, E., & Luu, J. (2012, November). *Improving forecaster accuracy using automatized calibration training experts*. Poster session presented at the 33rd Annual Conference of the Society for Judgment and Decision Making, Minneapolis, MN.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Tetlock, P. E., & Manstead, A. S. (1985). Impression management versus intrapsychic explanations in social psychology: A useful dichotomy? *Psychological Review*, *92*, 59–77.

Tetlock, P. E., & Mellers, B. A. (2011). Intelligent management of intelligence agencies: Beyond accountability ping-pong. *American Psychologist*, *66*, 542–554.

Tetlock, P. E., & Mitchell, G. (2009). Implicit bias and accountability systems: What must organizations do to prevent discrimination? *Research in Organizational Behavior*, *29*, 3–38.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2013). Forecast aggregation via recalibration. *Machine Learning*. Advance online publication. doi:10.1007/s10994-013-5401-4

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, *39*, 176–190