Judgment and Decision Making, Vol. 12, No. 6, November 2017, pp. 610–626

Accountability and adaptive performance under uncertainty: A long-term view

Welton Chang* Pavel Atanasov† Shefali Patil‡ Barbara A. Mellers§ Philip E. Tetlock¶

Abstract

Accountability pressures are a ubiquitous feature of social systems: virtually everyone must answer to someone for something. Behavioral research has, however, warned that accountability, specifically a focus on being responsible for outcomes, tends to produce suboptimal judgments. We qualify this view by demonstrating the long-term adaptive benefits of outcome accountability in uncertain, dynamic environments. More than a thousand randomly assigned forecasters participated in a ten-month forecasting tournament in conditions of control, process, outcome or hybrid accountability. Accountable forecasters outperformed non-accountable ones. Holding forecasters accountable to outcomes ("getting it right") boosted forecasting accuracy beyond holding them accountable for process ("thinking the right way"). The performance gap grew over time. Process accountability promoted more effective knowledge sharing, improving accuracy among observers. Hybrid (process plus outcome) accountability boosted accuracy relative to process, and improved knowledge sharing relative to outcome accountability. Overall, outcome and process accountability appear to make complementary contributions to performance when forecasters confront moderately noisy, dynamic environments where signal extraction requires both knowledge pooling and individual judgments.

Keywords: forecasting, process accountability, outcome accountability

1 Introduction

Accountability ground rules, that specify who must answer to whom for what, are essential features of human social life (Tetlock, 1985). One key distinction running through discussions of these ground rules is that between "process" versus "outcome" accountability (Lerner & Tetlock, 1999). In the ideal-type, pure-process accountability regime, people expect to justify their efforts and strategies to achieve results. The focus is on inputs, not outcomes. Under pure outcome accountability, the focus flips: people expect to answer for end-state results, with no interest in explanations of how they did it.

With certain notable exceptions (see de Langhe, van Osselaer & Wierenga, 2011; Patil, Tetlock & Mellers, 2016), judgment and decision making scholars are skeptical of the value of holding people accountable for outcomes (for a review, see Patil, Vieider & Tetlock, 2013) while simultaneously espousing the virtues of being accountable for process. The worry is that outcome accountability, or being responsible for results, can induce performance-debilitating levels of evaluative apprehension — thereby increasing commitment to sunk costs (Simonson & Staw, 1992), dampening complex thinking (Siegel-Jacobs & Yates, 1996), attenuating attentiveness (Brtek & Motowidlo, 2002), reducing epistemic motivation (De Dreu, Beersma, Stroebe & Euwema, 2006), and hurting overall decision quality (Ashton, 1992; Chaiken, 1980; Hagafors & Brehmer, 1983). For these reasons, many scholars have instead advocated holding people accountable for process, or the ways in which they go about making decisions, a form of accountability thought likelier to stimulate higher levels of deliberate processing, increased reflection and more learning (Ford & Weldon, 1981; Simonson & Nye, 1992).

Although past work has deepened our understanding of the effects of process and outcome accountability, close inspection reveals at least three serious limits on the generalizability of these findings.

First, these studies have largely been conducted in stable task environments with relatively little role for randomness in choice-outcome relationships. For instance, subjects made judgments about the attitudes of others (Brtek & Motowidlo, 2002; Siegel-Jacobs & Yates, 1996) and negotiated agree-

The authors thank the Intelligence Advanced Research Projects Agency for their support. This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

*3720 Walnut Street, University of Pennsylvania, Philadelphia PA, 19104. Email: welton@sas.upenn.edu.

†Pytho.

‡University of Texas, Austin.

§University of Pennsylvania

ments (De Dreu et al., 2006) — tasks in which following best practices of data gathering and analysis reasonably reliably deliver better outcomes. We know much less about how accountability shapes performance in dynamic environments with substantial roles for chance in action-outcome contingencies, tasks that require people to flexibly adapt to changes in predictive relationships (Grant & Ashford, 2008; Levitt & March, 1988; Patil & Tetlock, 2014). We should not assume that contextual factors that affect proficiency in tightly structured laboratory tasks in one direction will affect adaptivity in messier real-world settings in the same manner (Griffin, Neal & Parker, 2007).

The second limitation is that virtually none of the experimental work described above measures the influence of accountability systems over time. Previous work has been confined largely to single-session laboratory experiments in which subjects had little or no opportunity for learning and adjustment.

The third limitation is that previous research has used laboratory tasks in which the experimenters know the correct answers and the best practices or processes for reaching them could be specified precisely ex-ante. This limits external validity to the real-world in which no one knows the right answers or when even a best practice will lead to desired outcomes.

The current study, a large-scale longitudinal experiment, overcomes these limitations. We explore the short and longer-term effects of process, outcome, and hybrid accountability on *adaptive performance,* which we define as the ability to adjust to uncertain, complex and dynamic tasks (Griffin, Parker & Mason, 2010; Huang, Ryan, Zabel & Palmer, 2014).

We tested these effects in a field experiment involving thousands of subjects in a geopolitical forecasting tournament lasting ten months. The forecasting tournament provided an optimal setting for examining these effects for three reasons. First, geopolitical forecasting is inherently uncertain and dynamic. No one knows exactly how far the forecasting accuracy frontier can be pushed out ex ante (Tetlock & Mellers, 2011a). Thus, the tournament setup let us capture adaptive performance — i.e., improvements in forecasting accuracy that reflect a tendency to adjust to changing conditions. Second, the tournament enabled us to measure longer-term effects of accountability. Subjects had the opportunity to answer as many as 137 forecasting questions over ten months, often with updates to original forecasts, providing us with multiple judgment points over an extended period. Third, the tournament let us examine the effects of process, outcome, and hybrid accountability on accuracy improvements no only within individuals, but also any gains from knowledge sharing. We developed a system that (a) enabled fellow forecasters to rate how much they relied on shared information, and (b) measured how much shared information boosted knowledge consumers' performance.

Our study contributes to the debate over whether accountability improves judgments, a topic still debated as certain kinds of accountability within certain contexts can have no effect or decrease performance on tasks. In addition to painting a more balanced portrait of the pros and cons of process, hybrid and outcome accountability, our study highlights the importance of situating accountability in different task environments, which allows researchers to develop a more comprehensive understanding of the varied impacts accountability. Overall, by taking into account task environment and performance over time, we provide insights that more closely mirror how accountability operates in real-world settings.

2 Theory and Hypothesis Development

2.1 The Effects of Process, Outcome, and Hybrid Accountability on Adaptive Performance

A wide array of empirical precedents suggest that accountability for accuracy induces people to be more vigilant information processors and better belief updaters (Lerner & Tetlock, 1999; Chang, Vieider & Tetlock, in preparation). However, previous reviews of the accountability literature have also noted the instances when accountability either had no effect or detrimental effects on overall task performance (Lerner & Tetlock, 1999; Frink & Klimoski, 1998). Thus, accountability is not an unalloyed good: the context within which accountability exists matters to whether it improves the judgments of those who must live under the current regime's rules. If, for example, subjects were accountable to an audience that incentivized humorous explanations rather than accurate forecasts, subjects would predictably be incentivized to offer comments that do not advance the epistemic goals of accurate geopolitical forecasting. To the extent that the tournament-style experiment sets the conditions for putting these epistemic goals at the forefront, we posit that accountable forecasters will outperform those who are not accountable.

The context of accountability matters to how subjects react to being responsible for their choices and judgments. In dynamic environments, people need to shift between conforming to standard practices during periods of stability and deviating during periods of change (Bigley & Roberts, 2001; Patil & Tetlock, 2014). This fluctuation constitutes adaptive performance (Griffin et al., 2007; Griffin et al., 2010). It is important to note that adaptive performance does not mean simply changing strategies, but rather changing strategies *that results in better outcomes*. This requires a deepening understanding of the environment — an understanding developed by trial-and-error experimentation with strategies for maximizing outcomes (Lee, Edmondson, Thomke & Wor-

line, 2004; Mellers et al., 2014; Tetlock & Mellers, 2011a). This is usually achievable only after lengthy exposure to the evolving patterns of signals and noise in the environment (Thomke, 1998).

For these reasons, we expect that, in the short-term, process accountability could enhance adaptive performance relative to outcome and hybrid accountability. Previous research suggests that when people feel process accountable to well-defined guidelines, they tend to default into the low-effort coping strategy of conforming to standard practices (Patil & Tetlock, 2014; Patil et al., 2016). Conformity is a prudent political strategy because even if results are not achieved, one can claim that one did all one could within the bounds of what is currently deemed best practice (Edelman, 1992; Meyer & Rowan, 1977; Patil et al., 2013).

However, in dynamic environments, using standard practices can also boost performance for a number of reasons (Sutcliffe & McNamara, 2001). For one, it is well-known that people are susceptible to a host of cognitive biases (Kahneman, 2011), which can cause systematic mistakes (Northcraft & Neale, 1987). Standard practices that incorporate debiasing guidance can protect people from slipping into tempting but avoidable errors (Sutcliffe & McNamara, 2001). Furthermore, to the degree that standard practices reflect the current stock of organizational know-how, shared knowledge of cause-effect relationships that have worked in the past (Szulanski, 1996), standard practices can direct decision makers' attention to relevant information in an uncertain environment (Dean & Sharfman, 1996; March & Simon, 1958), sparing decision makers the frustrations of repeating the mistakes of their predecessors.

By contrast, subjects working under pure outcome accountability regimes are not protected by practice guidelines. They are simply expected to deliver results. As Siegel-Jacobs and Yates (1996: 2) note: "while outcome accountability may provide an additional incentive to produce a positively evaluated response, there is no guidance inherent in [. . .] how to achieve that goal (somewhat like simply shouting "get a hit" to the batter in a baseball game)." Thus, faced with evaluative pressures to deliver results in an unfamiliar environment, people under outcome accountability may flail under the uncertainty, sensing that they have been tasked with predicting the unpredictable (Patil et al., 2016). This may occur even if standard-practice guidelines are provided but not incorporated in the accountability systems in ways that permit "I-was-following-best-practices" excuses or explanations. A similar-but-weaker pattern should be observed under hybrid accountability because a portion of people's evaluations are still contingent on delivering outcomes that it may or may not be possible to deliver. Decision makers under hybrid accountability are under cross-pressure: pulled between "staying safe" with standard practices and reaching for novel solutions to achieve outcomes (Patil et al., 2013).

In the long run, however, outcome accountability could trump both process and hybrid accountability because standard practices tend to inhibit exploration and learning in uncertain, dynamic environments. Under process accountability, we expect that evaluation of processes will focus forecasters on how they think about thinking, in order to ensure consistency with standard practices (Feldman & March, 1981). Although this form of introspection can sometimes help to correct biased intuitions (Kahneman, 2011), it can also become oppressive. People working under process accountability are likely to feel self-conscious about how well their thinking sizes up against accepted practices. And because self-consciousness can induce excessive monitoring of behaviors (Baumeister, 1984; Carver & Scheier, 1978), it can also have the unintended effect of reducing reliability and success (Langer, 1978; Langer & Imber, 1979). As Martens and Landers (1972: 359) note: "Direct evaluation of the performance process [. . .] results in greater [. . .] impairment than the evaluation of the performance outcome only." In sum, we hypothesize that in the long-term, process accountability will hamper adaptive performance relative to outcome and hybrid accountability.

By contrast, outcome accountability could boost adaptive performance in the long-term because decision makers can move flexibly between introspection and execution to figure out which strategies are working in changing environments. Under outcome accountability, decision makers are less wedded to standard practices, and can engage in trial-and-error in the long-term (Patil et al., 2016). For example, in a flight simulation experiment, Skitka, Mosier, and Burdick (1999) found that compared to process accountability, outcome accountability increased pilots' skill in improvising during periods of change. This liberating effect of outcome accountability is one reason why scholars have advocated outcome-defined stretch goals that challenge people to push the bounds of what has been done before (Sitkin, See, Miller, Lawless & Carton, 2011). Because stretch goals move people into uncharted territories, they induce more openness in thinking about alternative strategies to achieve outcomes (March, 1991; March & Olsen, 1976; Sitkin, 1992). Furthermore, because outcome goals move people to the novel and unfamiliar, they can gain a sense of enthusiasm, curiosity, and urgency — all of which stimulate exploration and learning (Argyris & Schön, 1978; Barnett & Pratt, 2000; Greve, 1998). For these reasons, we expect that in the long-term, outcome accountability will enhance adaptive performance.

On balance, we expect hybrid accountability will have effects that fall between those of process and outcome accountability. Hybrid accountability should induce some degree of flexibility in shifting between standard practices and experimenting with novel strategies — in effect, harnessing the positive aspects of each accountability system. Some researchers have found that hybrid forms of accountability and reward systems can induce complex, flexible thinking (Green, Visser & Tetlock, 2000). Under hybrid account-

ability, decision makers try to achieve competing demands — learning from accumulated organizational knowledge and experimenting with the new — both of which promote performance (Feldman & Pentland, 2003). Hybrid accountability incentivizes this flexibility. That said, a large body of work gives us grounds to worry about people's capacity to balance competing performance criteria (Fischhoff & Chauvin, 2011). For example, research on collective versus individual rewards in teams shows that contradictory pressures to achieve opposing goals leads people to focus on one goal and ignore the other (Quigley, Tesluk, Locke & Bartol, 2007). Hybrid systems can also lead to analysis-paralysis, unduly delaying decisions (Ethiraj & Levinthal, 2009).

2.2 Effective Knowledge Transfer

Adaptive performance of individuals, or simple aggregations of crowd knowledge, is not the only reason that organizations create accountability systems. Another dependent variable central to organizational learning is effective knowledge transfer. Process accountability systems might be especially well-suited to promote such transfer which is arguably as important as knowledge creation (Kogut & Zander, 1992). Transferring knowledge among organizational members allows an organization to make fuller use of each individual's privately-developed expertise, boosting the effectiveness of an organization by elevating the knowledge base of all organizational members (Grant, 1996; Wernerfelt, 1984). But simply transferring content is not necessarily beneficial. For knowledge transfer to be effective, the intended recipients need to use the information and derive performance benefits from it (Levin & Cross, 2004).

We noted earlier that process and hybrid accountable decision makers are likelier to engage in self-conscious introspection because their conduct is being explicitly monitored (Carver & Scheier, 1978; Langer, 1978). Although monitoring can impair adaptive performance by imposing too much pressure to conform to standard practices, it can improve knowledge sharing and transfer. Because process and hybrid accountable decision makers have to defend their thought processes to evaluators, they are likelier to try to convert tacit knowledge into explicit forms that make their underlying reasoning process more understandable, transparent and persuasive. Tacit knowledge is by definition harder to codify than explicit knowledge (Hansen, 1999; Nonaka, 1994; Zander & Kogut, 1995). Although this conversion process can draw decision makers' attention from experimenting with new strategies, reducing adaptive performance, it can facilitate the communication of previously private knowledge. Others in the organization are likelier to understand and use information organized into shared schemata. Conversely, because outcome accountable decision makers are not being evaluated for their thinking, they are less likely to make this conversion effort, leading to less effective knowledge transfer than for their process and hybrid accountable counterparts.

Thus we ask whether process and hybrid accountability will boost effective knowledge transfer relative to outcome accountability.

3 Methods

3.1 Overview of Studies

We tested the effects of interest in the fourth year of a multi-year geopolitical forecasting competition sponsored by the U.S. Intelligence Advanced Research Projects Activity (IARPA). Five research teams (including ours) competed in a tournament with the goal of developing innovative and accurate methods for predicting a wide range of geopolitical events. The first three years of the tournament focused on producing accurate aggregate forecasts (Mellers et al., 2014, Atanasov et al., in press). The fourth year spanned from August 22, 2014 to June 10, 2015.

In both studies, subjects were presented with a diverse array of 137 total forecasting questions, which included: Will Russia officially annex any additional Ukrainian territory before 1 January 2015? Will the World Health Organization report any confirmed cases of Ebola in a European Union member state before 1 June 2015? Will North Korea test a long-range missile before 1 June 2015? They could place probability predictions and written justifications, and update their views at any time before each question resolved.

We aimed in the forecasting tournament research for a form of accountability toward the middle of the continuum, not telling people exactly what to think but also not leaving them completely to their own devices.

3.2 Defining Process, Outcome, and Hybrid Accountability

As noted, one prominent distinction among performance-monitoring systems is that between process and outcome accountability (Patil et al., 2016; Pitesa & Thau, 2013; Zhang & Mittal, 2007). Little experimental work has been done testing hybridized accountability systems and thus we set out to contribute to understand of efficacy of incentivizing both.

Process accountability evaluates employees on how they go about achieving results, not on the results themselves (de Langhe et al., 2011; Siegel-Jacobs & Yates, 1996; Tetlock, Vieider, Patil & Grant, 2013). But process accountability is not a unitary construct. It is best viewed as a tight-loose continuum of approaches to performance management (see Gersick & Hackman, 1990; Hackman & Wageman, 1995; March & Simon, 1958). At the tight end of the continuum, we find ultra-bureaucratic, assembly-line forms of accountability that specify each step of how to do the job. At the

loose end, we find the Rorschach-like, open-to-interpretation forms of accountability common in laboratory experiments on cognitive debiasing, in which subjects know as little as possible about the preferences of the audience to whom they must answer. In laboratory settings, subjects are typically asked to provide justifications for their choices and judgments. Subjects are only informed that researchers will examine their rationales and that they will be evaluated based on their quality.

Of course, organizations are not limited to a dichotomous choice between outcome or process accountability (Hochwarter et al., 2007; Tetlock & Mellers, 2011b). They can also evaluate performance based on "hybrids" or blends process and outcome criteria. Our hybrid approach employed process and outcome accountability evaluation criteria together.

Expanding on the above point, while it is possible to design a hybrid system, process and outcome accountability are qualitatively different constructs, each multidimensional in its own right — so any comparison will be problematic. For instance, if process accountability under-performs relative to outcome, one can always argue that the system deployed an unfairly weak version of process accountability: too vague guidelines, wrong guidelines, burdensome guidelines, too slow feedback, a misfit between guidelines and task environment. Or one could flip the argument and say that outcome accountability had its own distinctive pattern of unfair advantages: its greater simplicity and transparency, rapidity of feedback, or the plus of greater flexibility due to the heterogeneity of the experimental task.

Our process accountability manipulation attempted to hold people responsible for using what we later define as the "CHAMPS KNOW" guidelines for making high-quality forecasts, which are analogous to best-practice guidelines given to trainee physicians in diagnosing completely new patients or how to craft an intelligence assessment for analysts. These guidelines help forecasters pose useful questions about specific problems (e.g., look for comparison classes that offer clues to how often events of this sort, such as incumbents falling from power, have happened in the past) but do not offer anything close to the restrictive guidance that would make it impossible for subjects to break out from slavishly following them. We call this middle-ground solution "autonomous-professional" process accountability.

Outcome accountable forecasters in our research received the same training guidelines but were evaluated solely on the bottom-line accuracy of their judgments, not on the process they used to reach these judgments (Kausel, Culbertson, Leiva, Slaughter & Jackson, 2015; Patil et al., 2013; Simonson & Staw, 1992). Outcome accountability sends the message: here is our process advice but feel free to do whatever it takes to get the best possible answers because ultimately you will be held responsible for results.

We only had one opportunity at designing outcome, hybrid and process systems for this experiment and so we cannot claim to have covered all reasonable instantiations of each type of accountability.

That said, the process vs. outcome question is of fundamental social-organizational-political interest — so there is a need to encourage work on it. We made a good faith effort to construct a form of process accountability, organized around a training system that did repeatedly work in this task environment (Chang et al., 2016). We urged people to use their judgment in deciding which guidelines to stress for particular problems — and we stressed that the quality of one's explanation for one's forecast would be the sole basis for judging performance, not the accuracy of the forecast. In this sense, the process accountability manipulation resembled the rather open-ended process manipulations used in the lab literature on accountability, which have been found to be moderately effective in reducing certain biases (Lerner & Tetlock, 1999; Chang et al., 2016). As we explain later, process accountable forecasters had two process-specific opportunities to improve their scores in ways that were not incentivized for pure outcome forecasters: first, they could be more diligent at the actions supporting forecasting and second they could more fully utilize what they learned during initial forecasting training.

3.3 Subjects

A total of 1,850 subjects participated in the experiment. (Table 1 shows their demographic characteristics.) Subjects were recruited from professional societies, research centers, alumni associations, science blogs, as well as via word of mouth. Subjects completed a battery of psychometric and political knowledge tests prior to forecasting.

Subjects were largely U.S. citizens (69%), male (78%), and highly educated (94% had completed four years of higher education and 58% had post-graduate training). Of the 1,850 subjects, 39% had participated in the previous (third) year of the forecasting tournament. Random assignment produced proportional representation of subject characteristics across accountability conditions.

A no-accountability (control) comparison provides a benchmark for gauging the impact of accountability regimes on forecasting performance.[1] Control comparison forecasters participated in the same tournament and answered the same forecasting questions but were placed into a separate platform for logistical purposes. The forecasters for the control were recruited into the experiment through popular media coverage of previous years of the forecasting tournament — and were subjected to fewer entry requirements (e.g., we did not ask them to complete an individual differences survey upon intake). Upon induction, forecasters were randomly assigned to: a) an outcome-accountability condi-

[1] Internally referred to as the "Massively Open-Online Forecasting Competition" or MOOC

TABLE 1: Demographic characteristics of subjects by experimental condition.

Condition	N	Male (%)	Age (SD)	Education, % BA/BS	US Citizen	Retention (%)
Independent						
Outcome	215	76%	35.8 (11.6)	93%	68%	86%
Hybrid	212	72%	36.0 (11.8)	93%	71%	82%
Process	207	79%	36.0 (11.5)	94%	69%	78%
Team						
Outcome	404	78%	35.1 (10.7)	96%	69%	85%
Hybrid	410	78%	34.9 (11.6)	93%	69%	80%
Process	402	80%	36.1 (12.1)	94%	70%	81%

tion featuring a modified version of the forecasting training we provided to forecasters in the main experiment (approximately 15 minutes in length), and b) a no accountability condition, in which forecasters received no feedback on either the process or outcomes of their forecasting.

3.4 Design

Subjects were assigned (randomly, except for the control condition, as noted earlier) to one of four accountability conditions (control, outcome, process, hybrid) and two collaboration conditions (independent forecasters, cooperative teams). The treatment conditions utilized a full-factorial design. All control forecasters worked independently. Independent subjects had no access to the predictions or written justifications of others. They received scores and rankings based solely on their own performance. Teams were a part of the overall research design of the tournament (as noted earlier, the goal was to use any means possible to boost performance). Team conditions were composed of 30 teams of 13 individuals. Teams were encouraged and enabled to share predictions, rationales and individual messages. Mellers et al. (2014) describe the teaming manipulation and its impact on accuracy. Teaming is not a focus of the current study.[2]

Between seven and twenty subjects (3.5% to 6.5%) from each condition withdrew during the preseason practice period, which lasted 18 days. We replenished conditions from a waiting list, randomly assigning forecasters such that the conditions were evenly matched on the first day of scored forecasting.

[2]Effects of teaming were approximately orthogonal to those of accountability: we found no significant team-accountability interaction effects. Relevant analyses are thus collapsed across teams and independent forecasters.

3.5 Pre-Accountability Manipulation in Forecasting Training

At the start of the fourth season, before we randomly assigned subjects to their respective accountability conditions, we delivered a 90-minute online training module to all subjects. This training, which was encapsulated by the acronym CHAMPS KNOW, had previously been reported as effective in improving forecasting accuracy (Mellers et al., 2015; Mellers et al., 2014).

CHAMPS focused on psychological principles for improving prediction and probabilistic reasoning. KNOW focused on context — core concepts drawn from political science that were relevant to the subject matter of the forecasting problems the subjects encountered. Comparison Classes (C) encouraged forecasters to take the "outside view" by seeking out relevant reference classes and calculating base rates. Hunt for Information (H) taught forecasters how to find information for forecasting. Adjusting (A) taught the importance of reviewing forecasts to account for new events and the passage of time. The section on mathematical and statistical models (M) instructed forecasters to seek out formal models that captured past time-series or cross-sectional patterns. Post-mortems (P) offered advice on how to review past forecasting mistakes (as well as successes) and discover ways to improve. Select appropriate effort (S) covered the concept of cognitive triage, or allocating effort where it is likeliest to pay off. Another part of the training module explained the Brier scoring rule for assessing accuracy.

Training also covered political knowledge conveyed via the acronym KNOW. Know the Power Players (K) educated forecasters on best practices for analyzing political actors' goals, capabilities and constraints. Norms & Protocols (N) called attention to the laws, regulations, and protocols of important institutions (e.g., the United Nations, national constitutions, international credit rating agencies, electoral commissions) that shape geopolitical events. Other perspectives (O) trained forecasters not to overlook bottom-up sources

of influence (e.g., populist movements, cultural conflicts, grassroots ideologies). Finally, Wildcards (W) reminded forecasters of the limits of prediction in light of irreducible uncertainty in the world. Mellers et al. (2014) demonstrated that a predecessor of this training curriculum significantly improved forecasting accuracy. Control condition forecasters did not receive training.

3.6 Accountability Manipulation

The final portion of training explained how each forecaster would be held accountable (with the exception of the control condition). Accountability manipulations mimicked previous research (e.g., de Langhe et al., 2011; Patil et al., 2016; Siegel-Jacobs & Yates, 1996). Outcome-accountable forecasters were told that their goal was maximizing accuracy, independent of the process used to generate the forecasts. Process-accountable forecasters were informed that they would be evaluated based on the quality of their forecasting process, exemplified by empirically supported guidelines such as CHAMPS KNOW and behavioral measures of engagement. As an argument for the legitimacy of process accountability, forecasters were told that their process scores would not suffer due to bad luck on a forecasting question. Process-accountable subjects received information about how their "process" scores would be calculated.

Hybrid-accountable forecasters were told that their scores would reflect equal weights on both outcomes (accuracy) and process (application of evidence-based guidelines in ways other forecasters find useful).

3.7 Feedback

Forecasters in all conditions received regular performance feedback. Those in the process-accountability condition received monthly process scores (explained in detail below) as well as a detailed letter explaining how to interpret the scores; subjects in the outcome accountability conditions received accuracy scores whenever a question resolved for the questions they participated in forecasting (3 times per month on average); and, subjects in the hybrid accountability condition received both scores.[3] Additionally, process accountable forecasters received examples of well-written rationales, since a major part of the forecasting process was to develop high quality justifications for their forecasts. All subjects received feedback on the outcomes of questions they forecasted, allowing them to compare their predictions to the ground truth.

Past research on "good process" guided our development of process scores, which consisted of two components, which we termed "objective" and "subjective", with a total minimum value of zero and maximum value of 100 points. First, decision makers' processes should follow precedent, reflecting what worked in the past (Sutcliffe & McNamara, 2001). Following precedent enables more predictable and stable judgments. Objective process scores captured past process-following behavior. We examined data from the previous three years of the tournament to find a set of behavioral measures that were associated with higher levels of accuracy. The forecasting accuracy component was operationalized as the mean of the standardized Brier scores for each forecaster. The behavioral variables most predictive of accuracy were identified via least absolute shrinkage and selection operator (LASSO) regression.

The objective process score was thus set as a function of the: (1) mean number of forecasts per question; (2) mean number of hyperlinks included in each written forecaster justification; (3) number of clicks on links to relevant news stories within the forecasting platform; and (4) the number of analytic products ("key comments") written.[4] Each component of the objective process score was equally weighted after being log transformed and standardized (Dawes, 1979). Objective process scores were rescaled so that the worst scoring subject earned a score of 0 and the best earned 50. Forecasters learned about the general methodology for generating objective scores, but not the exact variables used in the scoring or the mathematical formulas translating behavioral measures into process scores.

Subjective process evaluation constituted the other half of the process score. This component was based on the notion that "good process" is the result of inter-subjective validation and agreement (Hackman & Wageman, 1995). Forecasting is too complex to capture all possible useful guidelines — so the best way to evaluate proper process adherence is to ask other trained forecasters to rate the qualitative commentary. We calculated this component by asking other independent forecasters to rate the utility of reasoning and justifications (referred to internally as "key comments") that subjects submitted alongside their forecasts. These forecast justifications (with the associated probability estimate and identifying information removed) were assigned to many trained raters who then read and judged the quality of the rationales provided.[5] The average forecast justification was rated independently by 10.8 (sd = 6.1) raters. Subjects were provided, on a monthly basis, with process feedback, which consisted of exemplar rationales that were highly rated and demonstrated proper use of CHAMPS KNOW principles as part of their overall package of feedback in the first three months

[3]We included a copy of the exact monthly feedback letter to process accountable forecasters in the supplement.

[4]In the experiment, forecasters were asked to identify their own comments as "key" if they thought they were particularly useful and used concepts from the forecasting training. We selected four supplementary metrics (bringing the total to nine) for process forecasters on teams: mean forum post length, number of reply comments generated, number of team-mail messages received, and number of up-votes given to teammates comments.

[5]We excluded key comments that did not include any original material but featured only text from a pull-down menu. One forecaster generated approximately 60% of all invalid key comments.

of tournament, and subjective scores thereafter. Sample of forecaster feedback is shown in Appendix A1.

The raters were a separate group of forecasters in another experimental condition in the tournament who were not in competition with those in the accountability experiment.[6] All raters received CHAMPS KNOW training. They assessed each justification on a scale from one to five, based on the usefulness of each forecast justification in aiding their own probability estimates, as well as the extent to which justification writers insightfully applied CHAMPS KNOW training concepts.[7] Thus, by design, highly rated justifications would not only follow standard-practice guidelines but also feature insights that other forecasters find helpful to their own forecasting practice. The emphasis on usefulness made the evaluation somewhat open-ended, discouraging thoughtless retrieval of CHAMPS KNOW principles.

To correct for potential rater biases (i.e., harshness or leniency) we debiased the ratings.[8] The subjective process scores were then rescaled to a 0–50 scale and combined with the objective process scores into the overall process score such that the highest possible process score for any forecaster was 100. This total process score was provided to process and hybrid accountable subjects every month along with exemplars of rationales (the same exemplars for everyone), which demonstrated good process execution.

3.8 Adaptive Performance

We used forecasting accuracy scores as the measure of adaptive performance. Accuracy scores reflected subjects' ability to cope with an uncertain and dynamic forecasting task. In order to correctly answer forecasting questions, subjects had to effectively deal with complex situations and irreducible uncertainty by adapting strategies to achieve the best scores possible. We used the Brier scoring rule (Brier, 1950) to measure forecasting accuracy. The Brier scoring rule is commonly used and "strictly proper": it incentivizes forecasters to report their true beliefs avoiding both underconfidence and overconfidence. Brier scores are the sums of squared deviations between probability forecasts and ground truth (in which ground truth is coded as "1" if the event occurs and "0" otherwise) and range from 0 (best) to 2 (worst). For example, suppose a forecaster reported that option A of a two-option question was 75% likely (thus, option B is 25% likely), and outcome A occurred. The forecaster's Brier score would be $(1-0.75)^2 + (0-0.25)^2 = 0.125$.

Scores were calculated based on a forecaster's estimate for each day and then averaged over all days at the question level. Adaptive performance at the individual level is thus defined as each individual's Brier score computed over time and across questions.[9] A mixed effects model accounted for variance in accuracy across questions, so that forecasters who attempted more difficult questions can be compared with those that tended to select easier ones. Outcome and hybrid accountable forecasters received Brier scores on each question and were informed of their mean score across all of the questions they attempted, as well as their overall accuracy ranking. Process accountable forecasters learned the outcomes of the questions they answered, and received detailed breakdown of their Brier scores after the end of the forecasting season.

In addition to individual accuracy, we also measured aggregate accuracy by experimental condition. To do this, we first calculated simple averages of forecasts across subjects in an experimental condition, then compared these aggregate forecasts across conditions. Conceptually, aggregate forecasts represent the cumulative predictive knowledge of a condition across all its subjects. Importantly, analysis at the aggregated level allows for direct comparison across conditions without the need for statistical adjustments because forecasts are simultaneously available for the entire duration of each question.

3.9 Effective Knowledge Transfer

Effective knowledge transfer from person to person is a two-step process: first that the information is actually used and second that it has a positive impact (Levin & Cross, 2004). We developed a net index of effective knowledge transfer by combining three measures: the quantity of forecast justifications written, the number of times that these forecast justifications led to forecast updates by raters (i.e., after reading the justifications, the raters updated their own forecasts), and the degree to which these subsequent forecast updates improved accuracy. The proportion of forecast justifications that led to updates gave us a measure of persuasiveness. Accuracy enhancement was measured as the extent to which raters became more accurate by adjusting their predictions. For example, if a rater updated her probabilistic estimate from 30% to 20% for an event that did not occur, the forecast justification that was credited for the update would receive points for the boost in accuracy. Appendix A2 in the supplement

[6]The raters were told to consider these analytic products as "Tips", which could help them improve their own predictions.

[7]A comment rating of "1" corresponded to a "not at all useful" label, "3" to "moderately useful" and "5" to "extremely useful".

[8]We calculated how each rating differed on average from other ratings of the same forecast justification and subtracted this mean difference from all their ratings. We then averaged the debiased ratings for each product, then across products generated by all individual forecasters and rescaled these scores such that the worst observed Subjective Process Score is 0 and the best is 50.

[9]For individual subject analyses, we rescaled the Brier score at the question level, to account for variance in question difficulty. We did so to avoid penalizing individuals for attempting difficult questions. Rescaling involves subtracting the mean score for all subjects from the raw score and dividing by the standard deviation. This yields a distribution with mean zero and a standard deviation of one, although not necessarily a normal distribution.

provides an example of a written forecast justification, as well as the belief revision and scoring workflows.

4 Results

4.1 Does Accountability Impair or Improve Forecasting?

The comparison of relevance is on accuracy of the no-accountability vs. outcome accountability conditions. The outcome of interest in this part of the study was the accuracy of simple averages of forecasts in each condition. Accuracy was measured using the Brier scoring rule and the absolute distance rule (like the Brier score but using the absolute value of the distance between the judgment and the 1/0 outcome rather than its square). Forecasts included were generated between August 22, 2014 and June 9, 2015, and covered 135 questions. For the 135 questions, Brier scores were higher (indicating worse forecasting performance) in the no accountability (M = .41, SD = .31) vs. outcome accountability (M = .28, SD = .22, Cohen's d = .49, $t(134) = 5.56$, $p < .001$). See Table 1. These results demonstrate the effectiveness of accountability in generating higher levels of forecasting performance rather than dampening it. However, the subjects were not well matched for this comparison; they were not randomly assigned to the control condition. *One difference was that forecasters under no accountability received no forecasting training, while those in outcome accountability condition did receive training. To assess the approximate impact of training on accuracy of aggregated predictions, we turn to a comparison of trained vs. non-trained outcome accountable forecasters in the accountability comparison arm (only the trained forecasters were included in subsequent analyses). The comparison showed that Brier scores for aggregated predictions of trained forecasters (M = 0.28, SD = 0.25) did not differ significantly from those made by untrained forecasters (M = 0.27, SD = 0.21, t(134) = 0.87, p > .10, n.s.).*

In the accountability comparison arm of the experiment, aggregated forecasts made by process accountable forecasters yielded significantly worse Brier scores (M = .33, SD = .21), than those made by outcome accountable (M = .28, SD = .25, Cohen's d = .23, $t(134) = 3.60$, $p < .001$), and hybrid accountable forecasters (M = .30, SD = 0.25, Cohen's d = .15, $t(134) = 2.91$, $p < 0.01$). Bonferroni adjustment for multiple comparisons was applied to these comparisons.

In summary, we found that the difference between outcome and process accountability (Cohen's d = .23) was approximately half as large as the difference between outcome accountability and no accountability in the MOOF experiment (Cohen's d = .49). This analysis implies that outcome accountability was about twice as effective at improving accuracy as process accountability, when compared to a no-accountability baseline. The training vs. no training comparison pointed to the result that the differences in accuracy across experimental conditions were driven by generally being held accountable, and not by differences in forecasting training.

TABLE 2: Comparison of Brier scores for aggregated predictions by experimental condition. All conditions featured forecasting training, except the last row.

	Brier Score	
Accountability	Mean	SD
Control Comparison		
Outcome	0.282	0.223
No accountability	0.411	0.308
Accountability Comparisons		
Process	0.330	0.209
Hybrid	0.297	0.230
Outcome	0.277	0.253
Outcome, No Training	0.269	0.207

4.2 Adaptive Performance in the Short and Long Term

4.2.1 Aggregate Accuracy over Time

To ask whether short-term, process accountability would enhance adaptive performance relative to outcome and hybrid accountability, we examined patterns of aggregate crowd performance across accountability conditions. To assess aggregate performance, we first averaged the probability estimates of the forecasts for each forecaster on each question within each accountability condition, using as an unweighted linear opinion pool (ULinOP) estimate. We then calculated the daily Brier scores for the accountability condition ULinOP estimates for each question. For example, if a forecasting question was open from September 1 to 30, we scored the accuracy of the averaged estimates for each condition on each of the 30 days by applying the known outcome. While the set of forecasting questions changed over time, on any given day the available set was identical across conditions. This produced a comparison of the wisdom of crowds, where each experimental condition was treated its own separate crowd.

Our adaptive performance model distinguishes between two types of time measures: across and within questions. If adaptive performance spurred by accountability is a more generalizable phenomena, then we would expect accuracy to improve *across* questions from the beginning to the end of the experiment. If adaptive performance is more the result of a narrower phenomena, then we would expect it to be the result of learning within questions, that is, accuracy for a

specific question improving from the beginning to the end of the life cycle of a question. Timing across questions is possible since some questions were posed early and others posed later during the months-long experiment. If performance differences are larger for late than for early questions, this may point to differential rates of learning across accountability conditions. Thus we took the estimate of the mid-point between question start and end date. The within-question measure was scaled so that 0% denoted the start date of a question, 100% the end date, and 50% denoted the mid-point. This within-question measure enabled us to estimate how much forecasters improve their probability estimates on each specific questions over time. To estimate the causal role of accountability on adaptive performance for the two types of learning in question, we used a mixed-effects model (Bates, 2010) with random intercepts for each question and random slopes for within-question timing.

With the Brier score measure of accuracy, the interaction between accountability and within question timing yields similar estimates within and across questions. The interaction effects of accountability and across-question timing on accuracy were less robust than the within-question timing effects. Namely, Brier scores did not diverge between outcome and hybrid for early vs. late questions, while the divergence was marginally significant between outcome and process groups (b = 0.012, t=2.53, $p < .05$). The outcome group outperformed the process group by a larger margin for late than for early questions.

With the absolute distance measure of accuracy, the effects of accountability on performance varied over time within question and across questions. The interaction effects for the accountability vs. timing within questions were significant, denoting that that the benchmark condition, outcome accountability, extended its lead over both the hybrid and process groups between the start and end date of each question (b = 0.085, t = 22.15, and b=0.0121, t = 31.54, respectively; p <.001 for both). Across questions, outcome accountability group extended its lead over hybrid (b = 0.054, t = 9.50, p < .001) and process (b = 0.104, t = 18.29, p < .001) groups over the course of the season, i.e. the advantage was larger for late than for early questions.

In summary, the outcome accountability condition generated probability estimates that were, in aggregate, more accurate than those elicited from the process accountability group, and the differences in accuracy grew over time, both within and across questions. Figure 1 illustrates the combined effects of within and across question timing by showing displaying comparative accuracy over the course of the season.

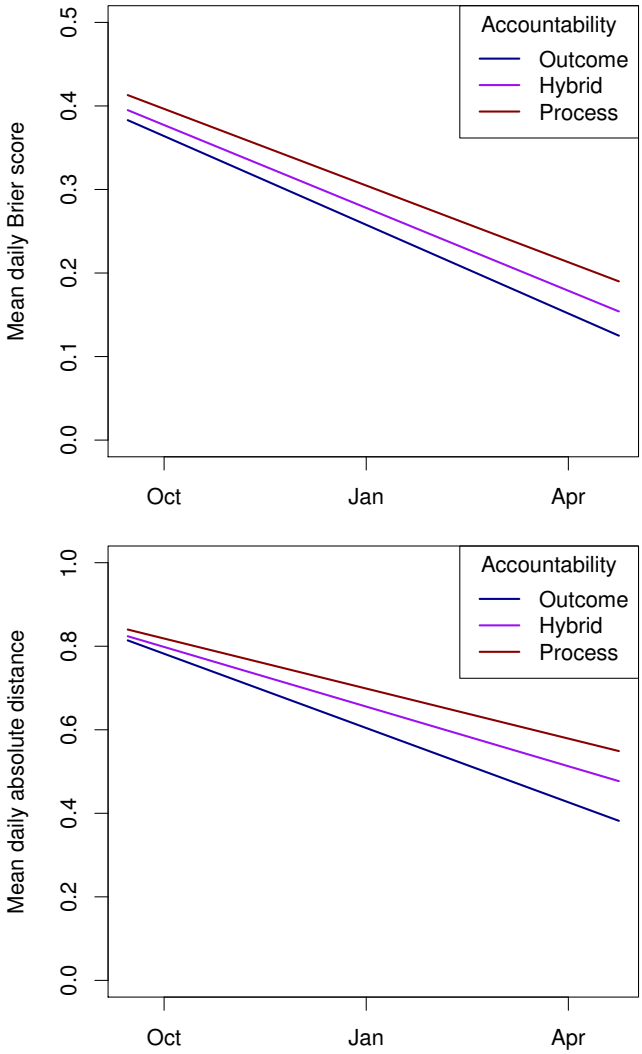

FIGURE 1: Accuracy of simple averages of forecasts over the course of a forecasting season by accountability condition.

4.2.2 Individual Accuracy Over Time

The analyses so far focus on aggregate accuracy of different groups, rather than individual performance of group members. How did individual subject performance differ across experimental conditions? To answer this, we used accuracy measures for individual subjects at the question level. Due to data sparseness, we did not break down accuracy scores by date within a question. Subjects received accuracy feedback in terms of Brier score, so this was used as the main outcome of interest. Absolute distance measure of accuracy was added in a sensitivity analysis. Accuracy scores were clustered per question, by specifying random question intercepts in the mixed model.

Focusing only on the main effects of accountability on individual accuracy, we found that process accountable subjects were marginally less accurate than outcome accountable peers In terms of Brier score (b = 0.018, t = 2.14, p < .05), but similar in terms of absolute score (b = 0.018, t = 1.01, p > .10) measures. The model did not detect signifi-

TABLE 3: Forecasting accuracy over time for simple averages. Higher values denote lower accuracy. Mixed-effects model coefficients, standard errors in parentheses. Process and hybrid accountability are compared to outcome accountability as the references.

	a. Brier Score	b. Absolute Distance
Intercept	0.518 (0.044)**	1.018 (0.058)**
Team	−0.080 (0.001)**	−0.116 (0.001)**
Accountability		
Hybrid	0.007 (0.003)*	−0.014 (0.004)**
Process	0.021 (0.003)**	−0.015 (0.004)**
Timing within questions	−0.219 (0.016)**	−0.382 (0.021)**
Timing across questions	−0.214 (0.007)**	−0.288 (0.093)**
Accountability × Time within Q		
Hybrid	0.032 (0.003)**	0.085 (0.004)**
Process	0.042 (0.003)**	0.121 (0.004)**
Accountability × Time across Q		
Hybrid	−0.002 (0.005)	0.054 (0.006)**
Process	0.012 (0.005)*	0.104 (0.006)**
Time within Q random slopes	Yes	Yes
Question random intercepts	Yes	Yes

* p<0.05, ** p<0.01.

cant difference between outcome and hybrid conditions. See Table 4a.

We then asked whether accuracy varied across questions that appeared early vs. late in the season. Subjects in the outcome and process accountability conditions produced estimates of similar accuracy on questions that appeared earlier in the season, but the outcome condition gained an advantage for questions that appeared later on. The accountability-time interaction effects were statistically significant for both Brier score (b = 0.041, t = 3.31, p < .01) and absolute distance (b = 0.055, t = 3.26, p < .01) accuracy measures. There interaction effects for differences between outcome and hybrid groups were close to zero, and were not statistically significant, denoting that performance between these conditions did not diverge over time. See Table 4b.

In summary, we examined the forecasting accuracy among individuals and for aggregate forecasts in each accountability condition, using both Brier score and absolute distance measures of accuracy. In but one of the above specifications, outcome accountability resulted in more accurate estimates than process accountability. On average, the accuracy differences were larger for questions appearing later in the tournament, and for forecasts at the late days within each question. The hybrid accountability group performed on par with outcome accountability throughout.

4.3 Best Practice Application

The strong performance of outcome and hybrid accountability is surprising, because application of standard practices had strong association with prediction accuracy in prior years, and process accountable forecasters were incentivized to apply these guidelines more diligently. To examine the point of breakdown for this process-outcome chain, we estimated the predictive relationship between CHAMPS KNOW guideline use, defined as the average number of CHAMPS KNOW self-reports per subject (subjects could tag their comments on the platform), and accuracy. We found that, across all conditions, forecasters who reported using standard practices produced more accurate predictions. This pattern held equally across the three accountability conditions. Yet, despite process group's greater incentive to report applying CHAMPS KNOW, they underperformed their hybrid and outcome accountable peers in terms of accuracy. See Appendix A3 in the supplement.

4.4 Effective Knowledge Transfer

We posited that process accountability would contribute more to effective knowledge transfer than would outcome or hybrid accountability. We break down effective knowledge transfer into two components: 1) persuasiveness: the power of analytical products, in this case, forecast justifica-

Table 4a. Forecasting accuracy over time among individual subjects. Higher values denote lower accuracy. Mixed-effects model coefficients, standard errors in parentheses. Process and hybrid accountability are compared to outcome accountability as the references.

	a. Brier Score	b. Absolute Distance
Intercept	0.373 (0.020)**	0.611 (0.028)**
Team	−0.028 (0.008)**	−0.093 (0.017)**
Accountability		
Hybrid	−0.002 (0.009)	−0.004 (0.017)
Process	0.018 (0.009)*	0.017 (0.017)
Question random intercepts	Yes	Yes

* p<0.05, ** p<0.01.

Table 4b. Forecasting accuracy over time among individual subjects. Higher values denote lower accuracy. Mixed-effects model coefficients, standard errors in parentheses. Process and hybrid accountability are compared to outcome accountability as the references.

	a. Brier Score	b. Absolute Distance
Intercept	0.494 (0.046)**	0.828 (0.057)**
Team	−0.028 (0.08)**	−0.093 (0.017)**
Accountability		
Hybrid	−0.003 (0.010)	−0.004 (0.019)
Process	0.001 (0.010)	−0.008 (0.019)
Timing across questions	−0.222 (0.008)**	−0.396 (0.092)**
Accountability × Time across Q		
Hybrid	0.002 (0.012)	−0.0002 (0.017)
Process	0.041 (0.012)**	0.055 (0.017)**
Question random intercepts	Yes	Yes

* p<0.05, ** p<0.01.

tions to convince raters to update their predictions, and 2) accuracy improvement, the subsequent boost in accuracy attributable to the forecast justifications. Additional analyses on justification ratings are described in Appendix A4 in the supplement.

Process-accountable subjects produced the most persuasive comments (10.6% of impressions resulted in a prediction update), followed closely by those in the hybrid condition (9.9%). Products by outcome-accountable forecasters were credited with belief updates only 8.2% of the time. Thus, process and hybrid accountable forecasters were approximately 20%-25% more likely to convince raters to update their predictions. To assess the statistical significance of differences in persuasiveness, we used a generalized linear mixed model, with random effects for question and justification writer, and fixed effects for teams and accountability condition. Justifications written under process (z =3.87, p < 0.001) and hybrid (z =3.49, p < 0.001) conditions were significantly more likely to result in a belief update than those composed under accountability.

Across all conditions, forecast updates resulted in accuracy improvements 85% of the time. In other words, when forecasters updated their predictions, they tended to move in the right direction. There were no differences among conditions in the average accuracy improvement associated with updates based on forecast justifications. In other words, when forecast justifications spurred raters to update their predictions, there was a similar decrease in prediction error associated with justifications written by process, hybrid and outcome accountable forecasters (p >.20 for all pairwise comparisons). Figure 2 summarizes the results.

Thus forecast justifications generated under process and hybrid accountability did lead to more effective knowledge transfer, in the sense that knowledge consumers were more likely to be persuaded to update their beliefs and *any* resulting forecast updating (across all conditions) boosted accuracy. Conditional on an update, however, the accuracy improvements were approximately equal across accountability conditions.

5 General Discussion

5.1 Overview and Limitations

This study, a long-term naturalistic experiment, revealed that accountable forecasters performed better than their non-accountable counterparts in terms of forecasting accuracy, and that outcome accountability produced better adaptive performance than process accountability, an effect that grew over time. This effect was present both within the specific questions answered as well as across the body of topics covered by forecasters, suggesting that a form of generalized learning occurred. Overall, the benefits for judgmental improvement from having any accountability at all were consistent with past findings demonstrating how accountability impact judgments.

These results pose a challenge for critics of outcome accountability who have warned, for example, that outcome accountability encourages over-fitting or over-explaining recent events, making us more vulnerable to be fooled by noise. The risk of overfitting to noisy outcomes is real. But outcomes also include valuable signals that may not be captured by standard practices in complex, dynamic settings. Outcomes

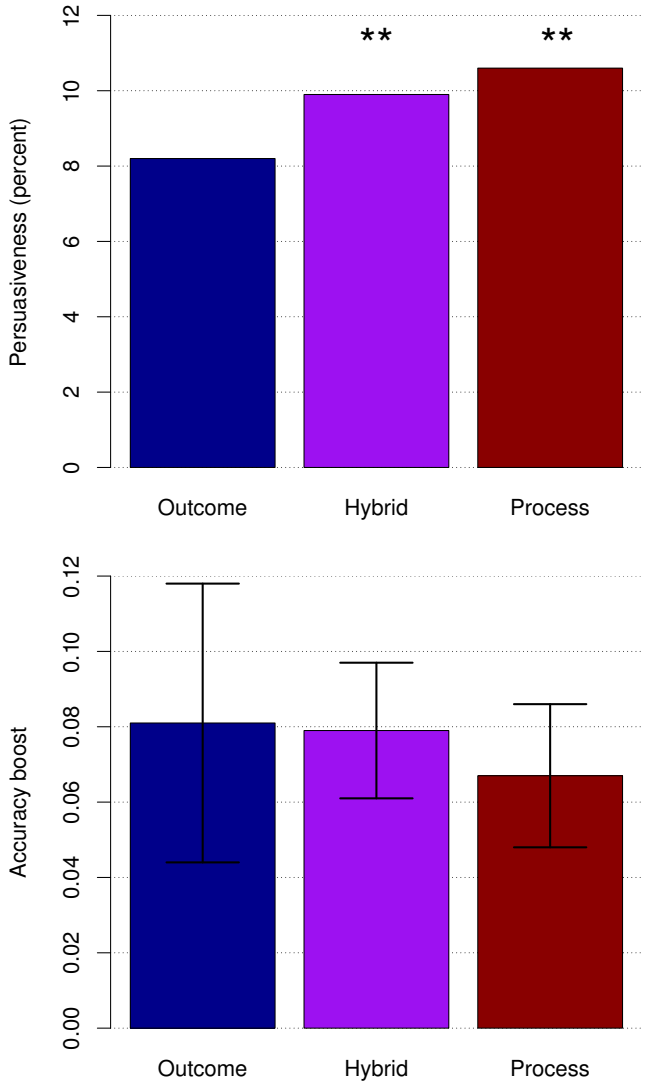

FIGURE 2: Persuasiveness and accuracy boost for analytical products generated by outcome, hybrid and process conditions. Persuasiveness is defined as the proportion of analytic product impressions resulting in belief updates. Accuracy boost is calculated as the mean Brier score improvement, before vs. after belief updates credited to key comments.

may teach valuable lessons that we have not had the opportunity to learn before. In some cases, learning these lessons will bring adaptive advantages that out-weigh the benefits of diligent implementation of best-practice guidelines under process accountability.

The experimental setting we describe may have been conducive to effective learning from outcomes: forecasters faced more than 130 diverse but always rigorously resolvable forecasting problems over 10 months and received clear feedback. Questions opened and closed throughout the season providing repeated opportunities for learning. Such conditions are in the spirit of the environments where deliberate practice, as defined by Ericsson (1993, p. 367), is likeliest to pay off: "The subjects should receive immediate feedback and knowledge of the result of their performance. The subject should repeatedly perform the same or similar tasks." Research on the benefits of deliberate practice has focused on controllable environments, with little irreducible uncertainty. We show that outcome accountability can improve learning in settings beset with uncertainty as well. It is an open question whether our results would hold up as well in environments with more variety in task formats or fewer opportunities to learn from feedback.

Skeptics may still justifiably ask: How can we know that the CHAMPS KNOW guidelines were the "best" best practices possible? Could a different, better set of guidelines have propelled process-accountable forecasters to higher accuracy, potentially outperforming the hybrid and outcome forecasters? We can never be certain. But we do know, based on three years of study and experimental testing, that the CHAMPS KNOW training did significantly improve accuracy above that of control conditions (Mellers et al., 2014, 2015; Chang et al., 2016). We also learned that some behaviors were strongly associated with better performance (e.g. frequent forecast updating), while others were not (e.g. attempting more questions). The process accountability metrics were far from arbitrary; they had a strong empirical basis and proven track record stretching back through the first three years of the forecasting tournament.

Crucially, the choices our team encountered in defining and scoring good process closely mirror the challenges that real-world organizations face in framing process guidelines. Few organizations have the luxury to subject their guidelines and potential alternatives to rigorous, experimental testing. With that in mind, the current research is arguably most relevant to environments with moderate amounts of irreducible uncertainty, such as forecasting tasks with time horizons of 12 to 24 months (Tetlock & Gardner, 2015). Proponents of outcome accountability may argue that boosting human performance in such domains is especially valuable because they are likeliest to remain in the domain of human, rather than algorithmic, forecasting.

Process accountability may be most effective in settings where coordination and efficiency are paramount, or those in which standard practices are more tightly linked to outcomes. But as Kahneman & Klein (2009, p. 63) point out, these are also the settings in which human judgment tends to underperform algorithms: "when validity is very high, in highly predictable environments [...] ceiling effects are encountered and occasional lapses of attention can cause humans to fail." The more concisely we can define the process guidelines, the easier it will be to transfer the job to algorithms.

The benefits of process accountability were most obvious in enhancing knowledge transfer. Subjects under hybrid and process accountability followed process guidelines more

diligently. These efforts paid off: not only in higher rated utility, but also in persuasiveness. This result underscores the importance of developing common schemata and guidelines for exchanging information in work settings. However, while process accountable subjects were better at persuasion, the guidance they produced was not any more effective in boosting accuracy than that of hybrid and outcome accountable subjects. There may often be mismatches between how much people think they are learning from each other's explanations and the value of those explanations in boosting performance.

5.2 Future Directions

Future research should aggressively explore boundary conditions on the effectiveness of different types of accountability on different types of personnel working in different types of environments. We now know that people working under outcome accountability outperform those under process accountability in a key way — they adapt to varying task demands and cues over time. But other key questions remain: when should we expect outcome accountability to reliably under- or over-perform process accountability? Could process accountable subjects improve their performance with more detailed feedback that explicitly laid out ways they could improve on each aspect of CHAMPS KNOW? And when should we conclude that hybrid accountability systems provide organizations with enough of the distinctive benefits of both process and outcome to justify the greater strain on people that complex incentive schemes are known to impose?

The current study asked subjects to tackle the task of forecasting political-economic outcomes — an activity with no neatly prepackaged answers. As Patil et al., (2016) notes, although proponents of process accountability point to the psychological protections that the best-practices defense affords against uncontrollable outcomes, a strict focus on process accountability may delay the recognition of randomness in the environment and the search for innovative alternative strategies. Pure process regimes may give agents too much psychological safety and too weak incentives to look beyond the rules. However, process accountability, coupled with precise and tailored feedback to subjects on each aspect of CHAMPS KNOW could have improved performance beyond the current results. Even more granular outcome feedback could have been provided on a regular basis as well, such as calibration, resolution, and scope sensitivity. While out of scope for our study, future work should devote resources to providing individualized feedback at scale and over time.

Importantly, this recognition could help outcome-accountable agents determine the extent to which environmental randomness impacts their performance, because the validity of effort-outcome coupling is front and center. Whereas previous research has been decidedly negative on the psychological effects of a singular focus on outcomes (i.e., agents feel it is unfair to be evaluated when events are outside of their control — and associated disruptive stress might diminish cognitive performance), one benefit is that agents should be quicker to recognize environmental randomness and thus better positioned to grasp the probabilistic cue-structure of their world while process-accountable agents are stuck within the guidelines passed on by organizational traditions.

Additionally, the current results suggest that hybrid systems, rather than overwhelming agents, a prediction from the literature on goal orientation (Ford, Smith, Weissbein, Gully & Salas, 1998; Kozlowski et al., 2001), can potentially mitigate the downsides of both outcome and process systems. That said, many questions remain unanswered about how best to design these systems. For example, can the weights assigned to each type of measure (outcome or process) influence how agents learn and perform? Within our experiment, the hybrid system started and remained at the agnostic setting of 50/50. This ratio should however be calibrated to the skills and preferences of personnel, the task at hand, the environmental setting, and the extent to which perfect execution of process reliably produces desired outcomes. Training agents on how to operate within hybrid systems can help them navigate competing accountability pressures and follow a path towards improved long-term performance similar to what we observed in the outcome-only systems in our study. Hybrid accountable individuals may also be better positioned than either outcome or process-only individuals to develop and codify new best practices, because they face both outcome-based pressures to perform and a need to explain how the chosen procedures are linked to outcomes.

Finally, the current study focused on prediction, a cognitively challenging task of anticipating and quantifying uncertain futures. Although such a task mirrors many of the challenges faced by knowledge-economy workers, it would also be useful to examine whether the relative performance of process, outcome and hybrid accountability systems holds for cognitively challenging tasks, such as problem solving, creative endeavors, and dispute resolution. For now, we can say that for tasks requiring analysis and synthesis of information under uncertainty, our study provides grounds for cautious optimism that process and hybrid accountability may boost, rather than hinder, long-term performance.

References

Argyris, C., & Schön, D. (1978). *Organizational learning*. Reading, MA: Addison-Wesley.

Atanasov, P. D., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P. E., Ungar, L., & Mellers, B. (in press). Distilling the wisdom of crowds: Prediction markets versus prediction polls. *Management Science, 63*(3), 691–706.

Ashton, R. H. (1992). Effects of justification and a mechanical aid on judgment performance. *Organizational Behavior and Human Decision Processes, 52*(2), 292–306.

Barnett, C. K., & Pratt, M. G. (2000). From threat-rigidity to flexibility: Toward a learning model of autogenic crisis in organizations. *Journal of Organizational Change Management, 13*(1), 74–88.

Bates, D. M. (2010). lme4: Mixed-effects modeling with R. Available at http://lme4.r-forge.r-project.org/book.

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology, 46*(3), 610–620.

Bigley, G. A., & Roberts, K. H. (2001). The incident command system: High-reliability organizing for complex and volatile task environments. *Academy of Management Journal, 44*(6), 1281–1299.

Brier, G. W. (1950). The statistical theory of turbulence and the problem of diffusion in the atmosphere. *Journal of Meteorology, 7*(4), 283–290.

Brtek, M. D., & Motowidlo, S. J. (2002). Effects of procedure and outcome accountability on interview validity. *Journal of Applied Psychology, 87*(1), 185–191.

Carver, C. S., & Scheier, M. F. (1978). Self-focusing effects of dispositional self-consciousness, mirror presence, and audience presence. *Journal of Personality and Social Psychology, 36*(3), 324–332.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752–766.

Chang, W., Chen, E., Mellers, B. A., & Tetlock, P. E. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision, 11*(5), 509–526.

Chang, W., Vieider, F., & Tetlock, P. E. (in preparation). Accountability: A hierarchical Bayesian meta-analysis of effect sizes (and situated-identity analysis of research settings).

De Dreu, C. K. W., Beersma, B., Stroebe, K., & Euwema, M. C. (2006). Motivated information processing, strategic choice, and the quality of negotiated agreement. *Journal of Personality and Social Psychology, 90*(6), 927–943.

de Langhe, B., van Osselaer, S. M. J., & Wierenga, B. (2011). The effects of process and outcome accountability on judgment process and performance. *Organizational Behavior and Human Decision Processes, 115*, 238–252.

Dean, J. W., & Sharfman, M. P. (1996). Does decision process matter? A study of strategic decision-making effectiveness. *Academy of Management Journal, 39*(2), 368–396.

Edelman, L. B. (1992). Legal ambiguity and symbolic structures: Organizational mediation of civil rights law. *American Journal of Sociology, 97*(6), 1531–1576.

Ethiraj, S. K., & Levinthal, D. (2009). Hoping for A to Z while rewarding only A: Complex organizations and multiple goals. *Organization Science, 20*(1), 4–21.

Feldman, M. S., & March, J. G. (1981). Information in organizations as signal and symbol. *Administrative Science Quarterly, 26*, 171–186.

Feldman, M. S., & Pentland, B. T. (2003). Reconceptualizing organizational routines as a source of flexibility and change. *Administrative Science Quarterly, 48*(1), 94–118.

Fischhoff, B., & Chauvin, C. (2011). *Intelligence analysis: Behavioral and social scientific foundations*. Washington, D.C.: National Academic Press.

Ford, J. K., & Weldon, E. (1981). Forewarning and Accountability. *Personality and Social Psychology Bulletin, 7*(2), 264.

Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. *Journal of applied psychology, 83*(2), 218.

Gersick, C. J. G., & Hackman, J. R. (1990). Habitual routines in task-performing groups. *Organizational Behavior and Human Decision Processes, 47*(1), 65–97.

Grant, A. M., & Ashford, S. J. (2008). The dynamics of proactivity at work. In B. M. Staw & R. Sutton (Eds.), *Research in organizational behavior* (Vol. 28, pp. 3–34). Greenwich, CT: JAI.

Grant, R. M. (1996). Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organization Science, 7*(4), 375–387.

Green, M. C., Visser, P. S., & Tetlock, P. E. (2000). Coping with accountability cross-pressures: Low-effort evasive tactics and high-effort quests for complex compromises. *Personality and Social Psychology Bulletin, 26*(11), 1380–1391.

Greve, H. R. (1998). Performance, aspirations, and risky organizational change. *Administrative Science Quarterly, 43*(1), 58–86.

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal, 50*(2), 327–347.

Griffin, M. A., Parker, S. K., & Mason, C. M. (2010). Leader vision and the development of adaptive and proactive performance: A longitudinal study. *Journal of Applied Psychology, 95*(1), 174–182.

Hackman, J. R., & Wageman, R. (1995). Total quality management: Empirical, conceptual, and practical issues. *Administrative Science Quarterly, 40*, 309–342.

Hagafors, R., & Brehmer, B. (1983). Does having to justify one's judgments change the nature of the judgment process? *Organizational Behavior and Human Process, 31*,

223–232.

Hansen, M. T. (1999). The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly, 44*(1), 82–111.

Hochwarter, W. A., Ferris, G. R., Gavin, M. B., Perrewé, P. L., Hall, A. T., & Frink, D. D. (2007). Political skill as neutralizer of felt accountability — job tension effects on job performance ratings: A longitudinal investigation. *Organizational Behavior and Human Decision Processes, 102*(2), 226–239.

Huang, J. L., Ryan, A. M., Zabel, K. L., & Palmer, A. (2014). Personality and adaptive performance at work: A meta-analytic investigation. *Journal of Applied Psychology, 99*(1), 162–179.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar Straus & Giroux.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist*, *64*(6), 515.

Kausel, E. E., Culbertson, S. S., Leiva, P. I., Slaughter, J. E., & Jackson, A. T. (2015). Too arrogant for their own good? Why and when narcissists dismiss advice. *Organizational Behavior and Human Decision Processes, 131*, 33–50.

Kogut, B., & Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science, 3*(3), 383–397.

Kozlowski, S. W., Gully, S. M., Brown, K. G., Salas, E., Smith, E. M., & Nason, E. R. (2001). Effects of training goals and goal orientation traits on multidimensional training outcomes and performance adaptability. *Organizational Behavior and Human Decision Processes, 85*(1), 1–31.

Langer, E. J. (1978). Rethinking the role of thought in social interaction. In J. Harvey, W. Ickes, & R. Kidd (Eds.), *New Directions in Attribution Research* (Vol. 2, pp. 35–58). Hillsdale, NJ: Erlbaum.

Langer, E. J., & Imber, L. G. (1979). When practice makes imperfect: Debilitating effects of overlearning. *Journal of Personality and Social Psychology, 37*(11), 2014–2024.

Lee, F., Edmondson, A. C., Thomke, S., & Worline, M. (2004) The Mixed Effects of Inconsistency on Experimentation in Organizations. *Organization Science*, *15*, 310–326.

Levin, D. Z., & Cross, R. (2004). The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. *Management Science, 50*(11), 1477–1490.

Levinthal, D. A., & March, J. G. 1993. The myopia of learning. *Strategic Management Journal, 14*, 95–112.

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

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*, 71–86.

March, J. G., & Olsen, J. P. (1976). *Ambiguity and choice in organizations*. Bergen: Universitetsforlaget.

March, J. G., & Simon, H. A. (1958). *Organizations*. Oxford, England: Wiley.

Martens, R., & Landers, D. M. (1972). Evaluation potential as a determinant of coaction effects. *Journal of Experimental Social Psychology, 8*(4), 347–359.

Mellers, B. A., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., . . . Tetlock, P. E. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of experimental psychology: applied, 21*(1), 1–14.

Mellers, B. A., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . Swift, S. A. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*(5), 1106–1115.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology, 83*, 340–363.

Nickerson, R. S. (1987). *Understanding understanding*. New York, NY: Bolt Beranek and Newman.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14–37.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39*(1), 84–97.

Patil, S. V., & Tetlock, P. E. (2014). Punctuated incongruity: A new approach to managing trade-offs between conformity and deviation. In B. M. Staw & A. Brief (Eds.), *Research in organizational behavior*, pp. 155–171. Greenwich: JAI Press.

Patil, S. V., Tetlock, P. E., & Mellers, B. A. (2016). Accountability systems and group norms: Balancing the risks of mindless conformity and reckless deviation. *Journal of Behavioral Decision Making*.

Patil, S. V., Vieider, F., & Tetlock, P. E. (2014). Process versus outcome accountability. *In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), Oxford Handbook of Public Accountability. New York: Oxford University Press,* pp. 69–89.

Pitesa, M., & Thau, S. (2013). Masters of the universe: How power and accountability influence self-serving decisions under moral hazard. *Journal of Applied Psychology*, http://dx.doi.org/10.1037/a0031697.

Quigley, N., Tesluk, P., Locke, E. A., & Bartol, K. (2007). A multilevel investigation of the motivational mechanisms underlying knowledge sharing and performance. *Organization Science, 18*, 71–88.

Siegel-Jacobs, K., & Yates, J. F. (1996). Effects of procedural and outcome accountability on judgment quality. *Organizational Behavior and Human Decision Processes, 65*(1), 1–17.

Simonson, I., & Nye, P. (1992). The effect of accountability

on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes, 51*(3), 416–446.

Simonson, I., & Staw, B. M. (1992). Deescalation strategies: A comparison of techniques for reducing commitment to losing courses of action. *Journal of Applied Psychology, 77*(4), 419–426.

Sitkin, S. B. (1992). Learning through failure: The strategy of small losses. In L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 14, pp. 231–266). Greenwich, CT: JAI Press.

Sitkin, S. B., See, K. E., Miller, C. C., Lawless, M. W., & Carton, A. M. (2011). The paradox of stretch goals: Organizations in pursuit of the seemingly impossible. *Academy of Management Review, 36*(3), 544–566.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991–1006.

Sutcliffe, K. M., & McNamara, G. (2001). Controlling decision-making practice in organizations. *Organization Science, 12*(4), 484–501.

Szulanski, G. (1996). Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management Journal, 17*(S2), 27–43.

Tetlock, P.E. (1985). Accountability: The neglected social context of judgment and choice. In B. Staw & L. Cummings (Eds.), *Research in organizational behavior* (Vol. 7, pp. 297–332). Greenwich, CT: JAI Press.

Tetlock, P. E. (1998). Close-call counterfactuals and belief-system defenses: I was not almost wrong but I was almost right. *Journal of Personality and Social Psychology, 75*(3), 639–652.

Tetlock, P. E., & Mellers, B. A. (2011a). Intelligent management of intelligence agencies: Beyond accountability ping-pong. *American Psychologist, 66*(6), 542–554.

Tetlock, P. E., & Mellers, B. A. (2011b). Structuring accountability systems in organizations: Key trade-offs and critical unknowns. In B. Fischhoff & C. Chauvin (Eds.), *Intelligence Analysis: Behavioral and Social Scientific Foundations* (pp. 249–270). Washington, DC: National Academies Press.

Tetlock, P. E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of Personality and Social Psychology, 57*(4): 632–640.

Tetlock, P. E., Vieider, F., Patil, S. V., & Grant, A. M. (2013). Accountability and ideology: When left looks right and right looks left. *Organization Behavior and Human Decision Processes, 122*(1), 22–35.

Thomke, S. H. (1998). Managing experimentation in the design of new products. *Management Science, 44*(6), 743–762.

Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal, 5*(2), 171–180.

Zhang, Y., & Mittal, V. (2007). The attractiveness of enriched and impoverished options culture, self-construal, and regulatory focus. *Personality and Social Psychology Bulletin, 33*(4), 588–598.