Appendix: Supplementary Examples and Analyses

A1. Sample Feedback Letter to Process Accountable Forecasters

Copy of Feedback Letter to Process Forecasters

Dear [Forecaster],

As discussed in your training materials, you are assigned to an experimental condition in which you will be judged on the accuracy of your forecasts (outcome). Outlined below are the methods we will use to evaluate your process and the feedback you can expect from us.

Your process performance will be judged using both objective and subjective process measures of good judgment. The objective score will be based on a combination of behavioral measures that have a proven track record of predicting forecasting accuracy. We will provide you monthly feedback on where you stand relative to the other forecasters in terms of your objective process score.

The subjective process measures will be based on the evaluations of your key comments by raters with specialized training in CHAMPS KNOW concepts and process evaluation. Key comments are evaluated on quality of CHAMPS KNOW application, as well as their general usefulness. The quality of your key comments will be more important than their quantity, so you should carefully designate as "key" only the comments that best showcase your skill. We will evaluate most key comments for most forecasters.

These subjective measures are completely new this year; therefore, we will need time to assess the validity of the subjective measures before sharing the results with you. Depending on the reliability and validity of initial ratings, we may only be able to send you feedback on subjective process metrics at the end of season. Our process accountability research team is working hard to provide subjective feedback sooner than this.

While we test our subjective process metrics, we will provide you with a sample of insightful, high-quality comments from IFPs that have closed during the past month. These comments will serve as examples of good process—a set of "greatest hits"—that will give you the chance to assess how your contributions compare to these examples.

At the end of the season, we will announce Super Process Analysts similar to those who achieved Superforecaster status in Seasons 1 through 3. Super Process Analysts will be the individuals and teams who achieved the highest process scores in the tournament.

In summary, your process score will be based on both objective and subjective metrics. At the end of September, you will start receiving feedback on objective measures related to good judgment and examples of good process. We will also provide subjective feedback about your performance will as soon as we verify the quality and validity of these metrics.

We hope you enjoy the final year of the Good Judgment Project.

A2. Illustration of Knowledge Sharing Flow

Forecasting question: Will a no-fly zone over any part of Syria be officially announced before 1 March 2015?

Start date: October 22, 2016; End date: February 28, 2016; Outcome: Event did not occur.
Justification: On November 5, 2015, Forecaster A places a 15% probability estimate of event occurrence, and contributes the following justification, self-marking it as a key comment.

> "As many analysts have pointed out, a no-fly zone (NFZ) would require either 1) cooperation from the Assad regime or 2) destruction of Syria's air defenses. The following military, economic, and political factors weigh heavily against either event occurring.
> Military factors: 1. A NFZ would require hundreds of US aircraft and a significant amount of personnel; 2. There's no indication a NFZ would alter the campaign against ISIS (the stated US goal); […]
> Economic factors: 1. Estimated US military costs for a Syrian NFZ are ~$1 billion/month.
> Political factors: 1. The US would have to stage equipment/personnel nearby for rescuing any downed pilots, thereby broadening the mission beyond President Obama's pledges; 2. Absent a UNSC or other international mandate (as in Libya, Iraq, Bosnia), the legal basis for a NFZ in Syria is lacking; there is no UN mandate for a NFZ and Russia has voiced its opposition to any NFZ in Syria; 3. President Obama has ruled out cooperation with Assad due to the regime's alleged human rights violations, and allies such as Turkey would oppose any deal that strengthened the Assad regime; and 4. The Syrian Foreign Ministry has voiced official opposition to any NFZ, citing sovereignty and territorial integrity. [Sources referenced here]"

On November 11, Forecaster B views this justification, along with four other justifications on this question. Forecaster B does not know the identity, experimental condition of Forecaster A, nor do they see which CHAMPS KNOW boxes A has checked. Forecaster B does as follows:

- places a rating of 4 ("very useful") to the justification. This rating is included in Forecaster A's process score.
- updates her own probability estimate from 30% to 15%. She indicates that Forecaster A's justification as the source of the update. This update positively impact the measures of persuasiveness and impact used to measure effective communication.

A3. Best Practice Application and Accuracy

When a forecaster submitted a prediction, they were asked to report which CHAMPS KNOW concepts they applied in formulating the prediction. This count ranged from 0 (no concepts used) to 10 (all concepts used in every comment), and averaged 1.6 per forecast explanation, across all conditions. This process-tracing data allows us to answer the question: Do forecasters who report greater use of standard practices achieve better outcomes? If forecasters apply guidelines judiciously, we would expect better performance. On the other hand, if forecasters simply engage in ritualistic box checking, the relationship between process and outcomes would be non-existent.

To answer these questions, we constructed regression models to estimate the relationship between rescaled Brier scores, the adaptive performance measure aggregated at the forecaster level. Predictors included experimental condition, the average number of CHAMPS KNOW concepts used per participant, and their interaction. The sample again included forecasters who attempted at least 5 questions.

Table A3. Accountability and best practice application as predictors of adaptive performance.

	a. Main effects only		b. Main Effects and Interaction	
Intercept	0.046 (0.021)	**	0.047 (0.026)	**
Team	-0.095 (0.018)	**	-0.095 (0.018)	**
Accountability				
Outcome (reference)				
Hybrid	0.037 (0.021)		0.011 (0.038)	
Process	0.098 (0.021)	**	0.104 (0.036)	**
CHAMPS KNOW count	-0.040 (0.010)	**	-0.040 (0.016)	**
Accountability x Count				
Outcome (reference)				
Hybrid			0.019 (0.022)	
Process			-0.010 (0.026)	
R-squared	0.05		0.05	

Table A2 shows the results. Consistent with results shown in Figure 1, outcome group significantly outperformed process group in terms of accuracy. In addition, CHAMPS KNOW application was associated with better accuracy across all conditions. The second model yielded no significant interaction between accountability condition and CHAMPS KNOW use: the benefits of more intensive application of best practices were approximately equal across the three accountability conditions.

A4. Best Practice Application and Comment Usefulness Ratings

Independent forecasters provided N=105,216 ratings of 9,812 key comments provided by participants. The average comment received more than ten ratings (M=10.7, Med=10, SD=6.1).

Process and hybrid accountable forecasters produced comments with significantly higher usefulness ratings than their outcome accountable peers (See Table A3, column a). Can we trace differences in perceived usefulness across conditions to measurable analytic product attributes? In other words, do process accountable forecasters excel only to the extent that they write longer products and check more boxes?

We performed four regression models to distinguish between direct and indirect effects of accountability on usefulness ratings. The first model estimated the relationship between experimental condition and usefulness. The second model also included product attributes: application of training concepts based on simple count of CHAMPS KNOW concept-use reports. We also included analytic product length, the natural logarithm of number of characters per comment. Notably, justification raters were blind to CHAMPS KNOW self-reports, experimental condition and accuracy performance of the product writers. Product length was the only rater-observable product characteristic.

The linear regression models are presented in Table A2 below. In the first model, both hybrid and process conditions registered significantly higher usefulness ratings than the outcome condition. In the full model, both CHAMPS KNOW index and product length were strong and significant predictors of usefulness ratings, while accountability condition was no longer significant. In other words, the advantage in perceived usefulness for process and hybrid conditions was largely traceable to their tendency to write longer products that used more CHAMPS KNOW concepts. The binary correlation between rescaled Brier scores and usefulness ratings was negative and significant (r =-0.27, t =-6.74, p <.001). In other words, more accurate forecasters tended to produce more useful analytical products.

Table A2. Predictors of comment usefulness ratings, aggregated within a subject.

Condition	a. Experimental Condition Only		b. Comment Attributes	
Intercept	2.31	(0.05)**	2.06	(0.06)**
Team	0.19	(0.04)**	0.19	(0.04)**
Accountability				
Outcome (reference)				
Hybrid	0.14	(0.06) *	0.05	(0.05)
Process	0.17	(0.06)**	0.01	(0.05)
CHAMPS KNOW concepts per comment			0.04	(0.02) *
Characters per comment, 1000s			0.10	(0.01)**
R-squared	0.07		0.30	

*p<.05; p<.01.