

ASSESSING CREDIBILITY IN SUBJECTIVE PROBABILITY JUDGMENT

Joshua D. Baker

A DISSERTATION

in

Psychology

and

Marketing

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

---

Jonathan Baron, Professor of Psychology

Graduate Group Chair, Psychology

Graduate Group Chair, Marketing

---

Sara Jaffee, Professor of Psychology

---

Catherine Schrand, Celia Z. Moh  
Professor, Professor of Accounting

Dissertation Committee:

Barbara Mellers, I. George Heyman University Professor  
Eric T. Bradlow, K.P. Chao Professor

ASSESSING CREDIBILITY IN SUBJECTIVE PROBABILITY JUDGMENT

COPYRIGHT

2019

Joshua D. Baker

This work is licensed under the  
Creative Commons Attribution-  
NonCommercial-ShareAlike 3.0  
License

To view a copy of this license, visit

<https://creativecommons.org/licenses/by-nc-sa/3.0/us/>

*for my Mom,  
who believed in me  
  
(and whose Brier score was better than mine)*

## ACKNOWLEDGEMENTS

Completing this dissertation has been one of the greatest challenges of my academic career. It took five and a half years, three years of therapy, two advisors, two decisions to quit, a medical leave of absence, and countless sleepless nights to get the thing done. Yet, here I am. Many thanks are due.

First and foremost, I would like to thank my advisor, Jonathan Baron. The idea for this dissertation blossomed from a conversation that I had with Jon in 2015 at the Annual Meeting of the Society for Judgment and Decision Making. At the time, Jon was just a faculty member that I would run into in the hallway from time to time and whose research made him a useful sounding-board for my early thinking on the evaluation of beliefs. Today, Jon is an admired mentor, a champion of my research, and one of the clearest thinkers I know. Thank you, Jon. Without your thoughtful guidance, illuminating criticism, and endless patience, I would not be where I am today.

I would like to thank my first advisor, Barbara Mellers, for opening the door to more opportunities than I can count. When I first met Barb, I was an enthusiastic but oblivious grocery-store clerk who was hopelessly out of his depth as an academic and a scientist. Now, only six years later, I am an accomplished coder, statistician, and methodologist — and it's all because Barb gave me a shot. Barb, in addition to being my first mentor, you were the first faculty member who believed in me and the first person to invest in my growth as a scientist. Over the years we have worked together, you have routinely encouraged me to push my boundaries as a researcher and hone my craft as a critical thinker. You have also provided me with the equivalent of a small fortune in terms of financial support, access to data, and professional development. For all these things, I am immensely grateful.

I would like to thank my committee chair, Eric Bradlow, for the many insights he has provided over the years and his seemingly inexhaustible enthusiasm for my research. Eric, though you and I only ever touched-base a few times a year, your excitement about my ideas — and the excitement it fostered in me — was an essential reminder that graduate school is supposed to be fun. Thank you for always pushing me to do better and for giving me the perspective I needed to realize that frustration and failure are sometimes the shortest path to success. It is because of this perspective that I have rediscovered the joy of exploration.

Major thanks are also due to the National Science Foundation and the Wharton Risk Management Center's Russell Ackoff Doctoral Student Fellowship, whose generous support made this research possible.

On a personal note, I would also like to thank my incredible friends for sticking with me during this challenging period of my life. Though far from exhaustive, this list includes (in no particular order): Chaz Lively, Bijan Haney, Jer Clifton, Welton Chang, Kelly

Allred, Steve St. Vincent, Kathy Kim, Sebastian Rowland, Kelley Foster, Ben Elder, Charlotte Swavola, Lauren Provini, Nate Robinson, Micah Plante, Jarus Singh, Jared Shenson, Alison Bedford, Charlie Croom, Emily Kearney, Frank Thompson, Alison Feder, Gretchen Stahl, Pablo Kim, and Perdita Kim-Lively.

Finally, I would never have made it this far without the love and support of my family: my mom, Janet Finn, who believed in me from the start; my stepdad, Michael Martel, who taught me to always strive for something better; my brother, Nick Baker, who reminded me to stay true to my roots; his partner Ashley Catchapaw, who is the best sister a guy could ask for; their daughter Alicia Ryan Baker, who is the cutest (and smartest!) monkey I know; and my girlfriend Melissa Ogg, who is the best person in my life and my favorite person in the world.

From the bottom of my heart, thank you all.

## ABSTRACT

### ASSESSING CREDIBILITY IN SUBJECTIVE PROBABILITY JUDGMENT

Joshua D. Baker

Jonathan Baron

*Subjective probability judgments (SPJs) are an essential component of decision making under uncertainty. Yet, research shows that SPJs are vulnerable to a variety of errors and biases. From a practical perspective, this exposes decision makers to risk: if SPJs are (reasonably) valid, then expectations and choices will be rational; if they are not, then expectations may be erroneous and choices suboptimal. However, existing methods for evaluating SPJs depend on information that is typically not available to decision makers (e.g., ground truth; correspondence criteria). To address this issue, I develop a method for evaluating SPJs based on a construct I call credibility. At the conceptual level, credibility describes the relationship between an individual's SPJs and the most defensible beliefs that one could hold, given all available information. Thus, coefficients describing credibility (i.e., "credibility estimates") ought to reflect an individual's tendencies towards error and bias in judgment. To determine whether empirical models of credibility can capture this information, this dissertation examines the reliability, validity, and utility of credibility estimates derived from a model that I call the linear credibility framework. In Chapter 1, I introduce the linear credibility framework and demonstrate its potential for validity and utility in a proof-of-concept simulation. In Chapter 2, I apply the linear credibility framework to SPJs from three empirical sources and examine the reliability and validity of credibility estimates as predictors of*

*judgmental accuracy (among other measures of “good” judgment). In Chapter 3, I use credibility estimates from the same three sources to recalibrate and improve SPJs (i.e., increase accuracy) out-of-sample. In Chapter 4, I discuss the robustness of empirical models of credibility and present two studies in which I use exploratory research methods to (a) tailor the linear credibility framework to the data at hand; and (b) boost performance. Across nine studies, I conclude that the linear credibility framework is a robust (albeit imperfect) model of credibility that can provide reliable, valid, and useful estimates of credibility. Because the linear credibility framework is an intentionally weak model, I argue that these results represent a lower-bound for the performance of empirical models of credibility, more generally.*

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	iv
ABSTRACT.....	vi
LIST OF TABLES.....	xii
LIST OF ILLUSTRATIONS.....	xvi
<b>INTRODUCTION.....</b>	<b>1</b>
Roadmap.....	6
A Conceptual Discussion of Credibility .....	7
References .....	13
<b>CHAPTER 1</b>	
<b>LINEAR REGRESSION CAN PROVIDE VALID AND USEFUL ESTIMATES OF</b>	
<b>CREDIBILITY IN SIMULATED DATA.....</b>	<b>17</b>
<b>Introduction .....</b>	<b>17</b>
Methods for Identifying <i>Estimated Optima</i> .....	19
Methods for Relating <i>Estimated Optima</i> to an Individual's SPJs .....	24
Why Examine Credibility? .....	28
<b>The Linear Credibility Framework.....</b>	<b>30</b>
What to Expect from the Linear Credibility Framework .....	32
Estimating an Individual's Credibility Function .....	33
<b>Study 1: Linear Credibility Estimation with Simulated Data .....</b>	<b>37</b>
Method.....	39
Results .....	45
General Discussion.....	57
<b>Conclusions .....</b>	<b>59</b>
<b>References .....</b>	<b>60</b>
<b>INTERLUDE</b>	
<b>CREDIBILITY ESTIMATION WITH EMPIRICAL DATA: A</b>	
<b>METHODOLOGICAL OVERVIEW OF STUDIES 2A-2C AND 3A-3C .....</b>	<b>65</b>
On the Structure and Organization of Chapters 2 and 3 .....	66

General Method for Studies 2a-2c and 3a-3c.....	68
References .....	86
 CHAPTER 2	
<b>LINEAR CREDIBILITY ESTIMATES ARE RELIABLE AND VALID</b>	
<b>PREDICTORS OF FORECAST ACCURACY .....</b>	<b>88</b>
<b>Introduction .....</b>	<b>88</b>
<b>Study 2a: Reliability and Validity of Credibility Estimates Derived from GJP Data (GJP Reliability/Validity).....</b>	<b>91</b>
Analysis 2a.i: Under what conditions are credibility estimates reliable? (GJP reliability) .....	93
Analysis 2a.ii: What are the predictors of credibility and what does credibility predict? (GJP validity).....	101
Analysis 2a.iii: How effective are credibility estimates at predicting forecast accuracy? (GJP enrichment vs. credibility).....	116
General Discussion .....	121
<b>Study 2b: Reliability and Validity of Credibility Estimates Derived from March Madness Data (MM Reliability/Validity).....</b>	<b>123</b>
Analysis 2b.i: Under what conditions are credibility estimates reliable? (MM reliability) .....	125
Analysis 2b.ii: What are the predictors of credibility and what does credibility predict? (MM validity) .....	130
General Discussion .....	146
<b>Study 2c: Reliability and Validity of Credibility Estimates Derived from Philadelphia Air Temperature Data (PHL Reliability/Validity) .....</b>	<b>149</b>
Analysis 2c.i: Under what conditions are credibility estimates reliable? (PHL reliability) .....	151
Analysis 2c.ii: What are the predictors of credibility and what does credibility predict? (PHL validity).....	157
General Discussion .....	163
<b>Conclusions .....</b>	<b>164</b>
<b>References .....</b>	<b>165</b>

CHAPTER 3

<b>THE LINEAR CREDIBILITY FRAMEWORK IS OFTEN USEFUL AND CREDIBILITY (IN GENERAL) IS WORTH EXAMINING.....</b>	<b>168</b>
<b>Introduction .....</b>	<b>168</b>
<b>Study 3a: Typical Effects of Recalibration with GJP Data (GJP Recalibration) 171</b>	
Method.....	172
Results .....	176
General Discussion .....	190
<b>Study 3b: Empirical Effects of Recalibration with March Madness Data (MM recalibration) .....</b>	<b>191</b>
Method.....	193
Results .....	193
General Discussion .....	208
<b>Study 3c: Empirical Effects of Recalibration with Philadelphia Air Temperature Data (PHL recalibration).....</b>	<b>209</b>
Method.....	211
Results .....	212
General Discussion.....	231
<b>Conclusions .....</b>	<b>232</b>
<b>References .....</b>	<b>234</b>
<b>CHAPTER 4</b>	
<b>ON THE ROBUSTNESS OF EMPIRICAL MODELS OF CREDIBILITY .....</b>	<b>235</b>
<b>Introduction .....</b>	<b>235</b>
<b>Study 4: The Effect of Question Heterogeneity on Reliability (PHL reliability redux).....</b>	<b>239</b>
Method.....	240
Results .....	241
Discussion.....	243
<b>Study 5: The Impact of <i>Estimated Optima</i> Accuracy on the Effects of Recalibration (MM recalibration, ALE redux).....</b>	<b>245</b>
Method.....	247
Results .....	247
Discussion.....	251

<b>Summary .....</b>	<b>252</b>
<b>References .....</b>	<b>254</b>
 <b>GENERAL CONCLUSIONS .....</b>	 <b>256</b>
 APPENDIX A .....	 260
APPENDIX B .....	323
APPENDIX C .....	418

## LIST OF TABLES

### CHAPTER 1

TABLE 1:

[Simulated data]: Simple correlations between simulation parameters and bootstrapped estimates of credibility .....	47
---	----

TABLE 2:

[Simulated data]: Predictors of bootstrapped alpha (i.e., bias; $\alpha i$ ) .....	47
--	----

TABLE 3:

[Simulated data]: Predictors of bootstrapped beta (i.e., expertise; $\beta i$ ) .....	48
---	----

TABLE 4:

[Simulated data]: Predictors of bootstrapped sigma (i.e., consistency; $\sigma i$ ) .....	48
---	----

TABLE 5:

[Simulated data]: Typical effects of credibility-based recalibration, summarized across forecasters .....	55
---	----

TABLE 6

[Simulated data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved (reduced) AJE beyond chance .....	55
--	----

### CHAPTER 2

TABLE 7:

[GJP data]: Simple correlations between credibility estimates and individual difference measures .....	105
--	-----

TABLE 8:

[GJP data]: Predictors of bootstrapped alpha (i.e., bias; $\alpha i'$ ) .....	106
---	-----

TABLE 9:

[GJP data]: Predictors of bootstrapped beta (i.e., expertise; $\beta i'$ ) .....	108
--	-----

TABLE 10:

[GJP data]: Predictors of bootstrapped sigma (i.e., consistency; $\sigma i$ ) .....	109
---	-----

TABLE 11:

[GJP data]: Predictors of average Brier score (i.e., forecast accuracy) .....	110
---	-----

TABLE 12: [GJP data]: A comparison of credibility measures vs. environmental enrichment variables as predictors of average Brier score (i.e., forecast accuracy).....	119
TABLE 13: [MM data]: Simple correlations between credibility estimates and individual difference measures .....	136
TABLE 14: [MM data]: Predictors of bootstrapped alpha (i.e., bias; $\alpha i'$ ).....	137
TABLE 15: [MM data]: Predictors of bootstrapped beta (i.e., expertise; $\beta i'$ ).....	138
TABLE 16: [MM data]: Predictors of bootstrapped sigma (i.e., consistency; $\sigma i$ ).....	140
TABLE 17: [MM data]: Predictors of average Brier score (i.e., forecast accuracy).....	141
TABLE 18: [PHL data]: Simple correlations between credibility estimates and individual difference measures .....	159
TABLE 19 [PHL data]: Predictors of bootstrapped alpha (i.e., bias; $\alpha i'$ ).....	160
TABLE 20: [PHL data]: Predictors of bootstrapped beta (i.e., expertise; $\beta i'$ ).....	160
TABLE 21: [PHL data]: Predictors of bootstrapped sigma (i.e., consistency; $\sigma i$ ).....	161
TABLE 22: [PHL data]: Predictors of average Brier score (i.e., forecast accuracy) .....	161

### **CHAPTER 3**

TABLE 23: [GJP data]: Typical effects of credibility-based recalibration on AJE, summarized across forecasters .....	180
TABLE 24:	

[GJP data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced AJE) beyond chance. ....	180
<b>TABLE 25:</b>	
[GJP data]: Typical effects of credibility-based recalibration on ALE, summarized across forecasters. ....	185
<b>TABLE 26:</b>	
[GJP data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced ALE) beyond chance. ...	185
<b>TABLE 27:</b>	
[GJP data]: Typical effects of credibility-based recalibration on reliability, summarized across forecasters.....	189
<b>TABLE 28:</b>	
[GJP data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced reliability) beyond chance. ....	189
<b>TABLE 29:</b>	
[MM data]: Typical effects of credibility-based recalibration on AJE, summarized across forecasters. ....	197
<b>TABLE 30:</b>	
[MM data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced AJE) beyond chance. ....	197
<b>TABLE 31:</b>	
[MM data]: Typical effects of credibility-based recalibration on ALE, summarized across forecasters. ....	202
<b>TABLE 32:</b>	
[MM data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced ALE) beyond chance. ...	203
<b>TABLE 33:</b>	
[MM data]: Typical effects of credibility-based recalibration on reliability, summarized across forecasters.....	207
<b>TABLE 34:</b>	
[MM data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced reliability) beyond chance. ....	207

TABLE 35: [PHL data]: Typical effects of credibility-based recalibration on AJE vs. crowd aggregates, summarized across forecasters.....	215
TABLE 36: [PHL data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced AJE vs. crowd aggregates) beyond chance.....	216
TABLE 37: [PHL data]: Typical effects of credibility-based recalibration on AJE vs. baserates, summarized across forecasters.....	220
TABLE 38: [PHL data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced AJE vs. baserates) beyond chance .....	221
TABLE 39: [PHL data]: Typical effects of credibility-based recalibration on ALE, summarized across forecasters .....	226
TABLE 40: [PHL data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced ALE) beyond chance. ...	226
TABLE 41: [PHL data]: Typical effects of credibility-based recalibration on reliability, summarized across forecasters.....	229
TABLE 42: [PHL data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced reliability) beyond chance. ....	230

## LIST OF ILLUSTRATIONS

### CHAPTER 1

FIGURE 1:	
[Simulated data]: Typical proportion of judgments for which recalibration improved (reduced) AJE .....	52
FIGURE 2:	
[Simulated data]: Typical pairwise change in AJE (pre – post), due to recalibration.	53
FIGURE 3:	
[Simulated data]: Typical effect-size (Cohen's d) of recalibration on AJE.....	53
FIGURE 4:	
[Simulated data]: Proportion of samples in which recalibration improved (reduced) mean AJE .....	54
FIGURE 5:	
[Simulated data]: Mean AJE, before and after recalibration. ....	54

### CHAPTER 2

FIGURE 6:	
[GJP data]: Reliability of non-bootstrapped estimates of credibility, varying by calibration sample size ( $n_{cal}$ ) .....	96
FIGURE 7:	
[GJP data]: Reliability of bootstrapped estimates of bias ( $\alpha i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ). ....	96
FIGURE 8:	
[GJP data]: Reliability of bootstrapped estimates of expertise ( $\beta i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ).....	97
FIGURE 9:	
[GJP data]: Reliability of bootstrapped estimates of consistency ( $\sigma i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ). ....	98
FIGURE 10:	
[MM data]: Reliability of non-bootstrapped estimates of credibility, varying by calibration sample size ( $n_{cal}$ ). ....	127

FIGURE 11:	
[MM data]: Reliability of bootstrapped estimates of bias ( $\alpha i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ). ....	128
FIGURE 12:	
[MM data]: Reliability of bootstrapped estimates of expertise ( $\beta i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ). ....	128
FIGURE 13:	
[MM data]: Reliability of bootstrapped estimates of consistency ( $\sigma i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ). ....	129
FIGURE 14:	
[PHL data]: Reliability of non-bootstrapped estimates of credibility, varying by calibration sample size ( $n_{cal}$ ). ....	153
FIGURE 15:	
[PHL data]: Reliability of bootstrapped estimates of bias ( $\alpha i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ). ....	153
FIGURE 16:	
[PHL data]: Reliability of bootstrapped estimates of expertise ( $\beta i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ). ....	154
FIGURE 17:	
[PHL data]: Reliability of bootstrapped estimates of consistency ( $\sigma i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ). ....	154
<b>CHAPTER 3</b>	
FIGURE 18:	
[GJP data]: Typical proportion of judgments for which recalibration improved (reduced) AJE. ....	178
FIGURE 19:	
[GJP data]: Typical pairwise change in AJE (pre – post), due to recalibration.....	178
FIGURE 20:	
[GJP data]: Typical effect-size (Cohen's d) of recalibration on AJE. ....	178
FIGURE 21:	
[GJP data]: Proportion of samples in which recalibration improved (reduced) mean AJE.....	179

FIGURE 22: [GJP data]: Mean AJE, before and after recalibration.....	179
FIGURE 23: [GJP data]: Typical proportion of judgments for which recalibration improved (reduced) ALE.....	183
FIGURE 24: [GJP data]: Typical pairwise change in ALE (pre – post), due to recalibration.....	183
FIGURE 25: [GJP data]: Typical effect-size (Cohen's d) of recalibration on ALE.....	184
FIGURE 26: [GJP data]: Proportion of samples in which recalibration improved (reduced) mean ALE.....	184
FIGURE 27: [GJP data]: Mean ALE, before and after recalibration.....	184
FIGURE 28: [GJP data]: Proportion of samples in which recalibration improved (reduced) reliability.....	188
FIGURE 29: [GJP data]: Typical pairwise change in reliability (pre – post), due to recalibration. .....	188
FIGURE 30: [MM data]: Typical proportion of judgments for which recalibration improved (reduced) AJE.....	194
FIGURE 31: [MM data]: Typical pairwise change in AJE (pre – post), due to recalibration.....	195
FIGURE 32: [MM data]: Typical effect-size (Cohen's d) of recalibration on AJE.....	195
FIGURE 33: [MM data]: Proportion of samples in which recalibration improved (reduced) mean AJE.....	196
FIGURE 34:	

[MM data]: Mean AJE, before and after recalibration. ....	196
<b>FIGURE 35:</b>	
[MM data]: Typical proportion of judgments for which recalibration improved (reduced) ALE. ....	200
<b>FIGURE 36:</b>	
[MM data]: Typical pairwise change in ALE (pre – post), due to recalibration. ....	200
<b>FIGURE 37:</b>	
[MM data]: Typical effect-size (Cohen’s d) of recalibration on ALE. ....	201
<b>FIGURE 38:</b>	
[MM data]: Proportion of samples in which recalibration improved (reduced) mean ALE.....	201
<b>FIGURE 39:</b>	
[MM data]: Mean ALE, before and after recalibration.....	202
<b>FIGURE 40:</b>	
[MM data]: Proportion of samples in which recalibration improved (reduced) reliability.....	206
<b>FIGURE 41:</b>	
[MM data]: Typical pairwise change in reliability (pre – post), due to recalibration. ....	206
<b>FIGURE 42:</b>	
[PHL data]: Typical proportion of judgments for which recalibration improved (reduced) AJE vs. crowd aggregates.....	213
<b>FIGURE 43:</b>	
[PHL data]: Typical pairwise change in AJE vs. crowd aggregates (pre – post), due to recalibration. ....	213
<b>FIGURE 44:</b>	
[PHL data]: Typical effect-size (Cohen’s d) of recalibration on AJE vs. crowd aggregates. ....	214
<b>FIGURE 45:</b>	
[PHL data]: Proportion of samples in which recalibration improved (reduced) mean AJE vs. crowd aggregates.....	214
<b>FIGURE 46:</b>	

[PHL data]: Mean AJE vs. crowd aggregates, before and after recalibration. ....	214
<b>FIGURE 47:</b>	
[PHL data]: Typical proportion of judgments for which recalibration improved (reduced) AJE vs. baserates. ....	218
<b>FIGURE 48:</b>	
[PHL data]: Typical pairwise change in AJE vs. baserates (pre – post), due to recalibration. ....	218
<b>FIGURE 49:</b>	
[PHL data]: Typical effect-size (Cohen's d) of recalibration on AJE vs. baserates.	219
<b>FIGURE 50:</b>	
[PHL data]: Proportion of samples in which recalibration improved (reduced) mean AJE vs. baserates. ....	219
<b>FIGURE 51:</b>	
[PHL data]: Mean AJE vs. baserates, before and after recalibration.....	220
<b>FIGURE 52:</b>	
[PHL data]: Typical proportion of judgments for which recalibration improved (reduced) ALE. ....	223
<b>FIGURE 53:</b>	
[PHL data]: Typical pairwise change in ALE (pre – post), due to recalibration.....	224
<b>FIGURE 54:</b>	
[PHL data]: Typical effect-size (Cohen's d) of recalibration on ALE. ....	224
<b>FIGURE 55:</b>	
[PHL data]: Proportion of samples in which recalibration improved (reduced) mean ALE.....	225
<b>FIGURE 56:</b>	
[PHL data]: Mean ALE, before and after recalibration. ....	225
<b>FIGURE 57:</b>	
[PHL data]: Proportion of samples in which recalibration improved (reduced) reliability. ....	229
<b>FIGURE 58:</b>	
[PHL data]: Typical pairwise change in reliability (pre – post), due to recalibration. ....	229

## CHAPTER 4

### FIGURE 59:

[PHL data]: Reliability of non-bootstrapped estimates of bias ( $\alpha_{in}^*$ ), varying by calibration sample size ( $n_{cal}$ ) and data subset..... 242

### FIGURE 60:

[PHL data]: Reliability of non-bootstrapped estimates of expertise ( $\beta_{in}^*$ ), varying by calibration sample size ( $n_{cal}$ ) and data subset..... 242

### FIGURE 61:

[PHL data]: Reliability of non-bootstrapped estimates of consistency ( $\sigma_{in}^*$ ), varying by calibration sample size ( $n_{cal}$ ) and data subset..... 243

### FIGURE 62:

[MM data]: Comparison of the typical proportion of judgments for which recalibration improved (reduced) ALE when estimated optima are derived from the SPJs of all forecasters vs. top performers..... 248

### FIGURE 63:

[MM data]: Comparison of the typical pairwise change in ALE (pre – post), due to recalibration when estimated optima are derived from the SPJs of all forecasters vs. top performers..... 249

### FIGURE 64:

[MM data]: Comparison of the typical effect-size (Cohen's d) of recalibration on ALE when estimated optima are derived from the SPJs of all forecasters vs. top performers..... 250

### FIGURE 65:

[MM data]: Comparison of the proportion of samples in which recalibration improved (reduced) mean ALE when estimated optima are derived from the SPJs of all forecasters vs. top performers..... 250

## INTRODUCTION

When the probabilities of uncertain events are unknown, subjective probability judgments (SPJs) play an important role in decision making. In medicine, for example, doctors rely on SPJs to identify the most likely cause of a patient’s symptoms and prescribe an effective treatment. In finance, investors rely on SPJs to predict the most favorable stocks and build a profitable portfolio. And in military intelligence, analysts rely on SPJs to anticipate threats and protect our national interest. Indeed, in any domain characterized by incomplete information, decision makers have little choice but to rely on intuitive, empirical, or otherwise bottom-up assessments of uncertainty. When expressed as SPJs, these assessments are assumed to reflect an individual’s *degree of belief* in uncertain events (Ramsey, 1926) and should — in principle— provide the same “type” or “class” of information as mathematical probabilities. According to most modern theories of choice (e.g., Savage 1954), therefore, SPJs are an integral component of decision making under uncertainty and can be incorporated into a decision maker’s rational calculus in cases where she must form (cognitive or behavioral) expectations in the absence of well-defined probabilities (Baron, 2008; Elster, 1986).

Critically, however, research on judgment under uncertainty has demonstrated that SPJs are an imperfect source of probabilistic information. In medicine, Eddy (1982) has shown that physicians tend to overestimate the likelihood of rare diseases (e.g., breast cancer), given positive test results (e.g., a positive mammogram). In finance, Barber & Odean (2001) have demonstrated that amateur investors tend to be overconfident in their

stock choices. And in military intelligence, Mandel & Barnes (2014) have shown that analysts tend to be *underconfident*, despite their high level of expertise. Indeed, after more than a century of descriptive research, studies of judgment under uncertainty have widely concluded that SPJs (a) are sensitive to context cues (e.g., Windschitl & Weber, 1999; Wallsten, Fillenbaum, & Cox, 1986) and modes of elicitation (Morgan, 2014); (b) are often formulated on the basis of risky (albeit efficient) heuristics (Kahneman, 2011; Kahneman, 2003; Tversky & Kahneman, 1974); (c) frequently fail to weight and incorporate evidence appropriately (e.g., Christensen-Szalanski & Bushyhead, 1981; Doherty, Mynatt, Tweney, & Schiavo, 1979; Kahneman & Tversky, 1973); (d) tend towards overconfidence in estimates of both binary and continuous criteria (e.g., Soll & Klayman, 2004; Soll, 1996); and (e) under certain conditions, violate the laws of mathematical probability (e.g., Birnbaum & Schmidt, 2008; Tversky & Fox, 1995; Tversky & Kahneman, 1983).

From a practical perspective, therefore, decision making under uncertainty presents a dilemma. On the one hand, SPJs represent a simple, inexpensive, and widely-available source of information about the probability of uncertain events. Indeed, in cases where a decision maker is concerned with one-off, unprecedented, or otherwise “unique” events, SPJs are likely to be her *only* source of probabilistic information, as it is impossible to draw inferences about the relative frequency of an event that will only occur once (for a seminal discussion of this topic, see: Ellsberg, 1961). On the other hand, research on judgment under uncertainty has demonstrated that SPJs are vulnerable to a variety of errors and biases (for an overview, see: Kahneman, 2011; Tversky &

Kahneman, 1974). In some cases, researchers have argued that these vulnerabilities are immaterial, as errors in judgment need not lead to errors decision making — and, in support of this position, Gigerenzer and colleagues have made a strong case for the “ecological rationality” of heuristic modes of judgment (e.g., Todd & Gigerenzer, 2007; see also: Brunswik, 1955). From a normative perspective, however, the simple fact of the matter is that suboptimal SPJs expose a decision maker to risk.

To put this in perspective, consider the neoclassical concept of rationality. According to this view — and therefore according to most normative theories of choice (e.g., Expected Utility Theory: von Neumann & Morgenstern, 1945; Bernoulli, 1738) — the goal of a rational decision maker is to maximize her expected utility (EU). In the face of uncertainty, however, a decision maker does not have access to (precise) mathematical probabilities and therefore cannot calculate EU directly. As discussed above, decision makers can compensate for this lack of “true” or “objective” probability information by substituting SPJs (i.e., intuitive, empirical, or otherwise bottom-up *estimates* of probability) into their rational calculus (Savage, 1954). In doing so, however, a decision maker runs the risk that the results of this calculus — i.e., her *subjective* expected utilities (SEU) — will fail to preserve the rank-order of her most strongly preferred alternatives (as measured by “objective” expected utilities) and cause her to choose suboptimally. In the interest of preventing this, real-world decision makers have an incentive to ensure that their SEUs are as valid as possible — especially in cases where errors in decision making are likely to be pervasive or costly. Because a decision maker’s utilities are unlikely to change between the calculation of EU and SEU, however, this incentive can be stated

more simply. Namely, decision makers have an incentive to ensure that their *beliefs* about an event's likelihood (i.e., their SPJs) are valid representations of an event's “true” probability, as dictated by the (perhaps latent) causal and/or stochastic structure of the decision environment. If they are, then a decision maker’s expectations and choices will be rational; if they are not, then she runs the risk that her expectations will be erroneous and her choices suboptimal.

Despite the inherent risk of decision making under uncertainty, however, research in the decision sciences has yet to develop practical tools for evaluating SPJs. In large part, this is because research on judgment under uncertainty can only provide direct insight into the validity of an SPJ when an event’s “true” or “objective” probability is available for comparison — or, at the very least, when strong claims about such values can be made (for seminar arguments to this effect, see: Savage, 1954; Ramsey, 1926). In some cases, this difficulty can be ameliorated by assessing the validity of SPJs indirectly — such as by examining the *coherence* of SPJs with truthful propositions (e.g., logical constraints, axioms, established facts); or the *correspondence* of SPJs with the outside world (e.g., historical tendencies, observed outcomes) (Dunwoody, 2009; Hammond, 2007). However, because these types of information are rarely available outside the lab, standards of this sort are of limited use to real-world decision makers when attempting to identify their best course of action, *ex ante*.

To address this issue, the present research will develop a method for evaluating SPJs based on their relative epistemic defensibility — a construct I call *credibility*. As a foundation for this research, I will leverage the empirical literature on forecast

aggregation (for an overview, see: Armstrong, 2001) to argue that high-quality approximations of “true” or “objective” probabilities (i.e., *estimated optima*) can often be identified with an ecologically realistic amount of data. By using these quasi-normative criteria as a standard for “good” judgment, I will then argue that decision makers can use simple statistical methods to identify an individual’s relative tendencies towards error and bias in judgment. These tendencies—captured by mathematical coefficients that describe the relationship between an individual’s SPJs and *estimated optima* (i.e., captured by empirical *credibility estimates*)—are the essence of what I call credibility. In practice, of course, the information provided by an empirical model of credibility will be constrained by (a) the latent agreement between a decision maker’s *estimated optima* and “objective probabilities;” and (b) the descriptive fit of the model used to relate *estimated optima* to an individual’s SPJs. However, as long as neither of these constraints is prohibitive, there is a reasonable theoretical basis for expecting that information about credibility will provide decision makers with insight into the “quality” or relative validity of an individual’s SPJs.

To determine whether this information can be extracted from real-world judgments, the purpose of this dissertation will be to examine the reliability, validity, and utility of empirical models of credibility. Because credibility is a *concept* rather than a specific scale or measure, however, I will not stake my claim in any one method for estimating credibility. Instead, I will attempt to build a case for examining credibility, in general. To do so, I will propose a specific model of credibility that I call the *linear credibility framework* and examine its performance across a wide variety of decision

environments. By design, the linear credibility framework will be a relatively weak model that is (a) minimally reliant on *ex post* data; and (b) among the most rigid and/or least informative that a decision maker might apply. Though this approach will limit the validity of the results presented in this dissertation, it will also help to mitigate concerns about the robustness of credibility information across heterogeneous (and often noisy or sparse) decision environments. Thus, while the “best” method for estimating credibility may vary from one case to another, the performance of the linear credibility framework can likely be treated as a lower bound for the performance of empirical models of credibility, more generally.

## Roadmap

In this dissertation, I will demonstrate that (a) simple statistical models fit to ecologically realistic amounts of data can be used to model credibility; and (b) that the information contained in these models can be used to identify an individual’s relative tendencies towards error and bias in subjective probability judgment (i.e., the “quality” or relative validity of her SPJs). To frame this contribution, I will begin with a brief discussion of credibility at the conceptual level, and its relationship to the notion of epistemic defensibility. In Chapter 1, I will then introduce the linear credibility framework and explore the applicability of this model to complex decision environments in a proof-of-concept simulation. In Chapter 2, I will apply the linear credibility framework to SPJs drawn from three empirical sources and examine the reliability and validity of linear credibility estimates (decomposed into components of *bias*, *expertise*,

and *consistency*) as predictors of performance in subjective probability judgment. In Chapter 3, I will use linear credibility estimates from the same three sources to recalibrate and improve SPJs (i.e., increase accuracy and reduce errors and biases) in an out-of-sample prediction task. Finally, in Chapter 4, I will discuss the robustness to empirical models of credibility and present two studies in which I use exploratory research methods to (a) tailor the linear credibility framework to the data at hand; and (b) boost empirical performance.

## A Conceptual Discussion of Credibility

To be useful to real-world decision makers, methods for evaluating SPJs must bridge the gap between top-down, *ex post* definitions of “good” judgment (e.g., “good” judgment is that which allows a decision maker to maximize expected utility), and the bottom-up, *ex ante* decision environments in which choices usually occur (e.g., deciding whether to undergo a risky medical procedure). In service of this goal, I will examine the performance of empirical models of *credibility*—a construct that is intended to describe the “quality” or relative validity of an individual’s SPJs.

Because credibility is built upon the notion of SPJ “quality,” however, defining credibility requires that I first define “good” judgment. From a normative perspective, I have already touched-upon this definition: according to rational theories of choice (e.g., Expected Utility Theory: von Neumann & Morgenstern, 1945), SPJs are typically defined as “good” or “rational” to the extent that they help decision makers (a) form valid expectations about the consequences of their actions; and/or (b) select the behaviors that

are best suited to reaching their goals (Baron, 2008; Elster, 1986; see also: Arkes et al., 2016). In practice, however, these standards can be difficult to apply because they essentially boil down to “an SPJ is ‘good’ to the extent that it resembles an ‘objective’ probability,” and decision makers typically do not have access to these values for comparison. When applying this standard prescriptively, therefore, it is useful to recall that normative theories of choice rely on probabilities because they are central to the mathematical definition of *expectation*. From the perspective of a real-world decision maker, this means that the absence of (precise, mathematical) probability information is a problem with the *operationalization* of normative theories of choice—not a problem with the theories themselves. Thus, if a decision maker (or a decision scientist) is willing to divorce the *concept* of expectation from its mathematical expression, then he or she can begin to get some traction on the idea of an “objective” probability, even in cases where it is an abuse of the mathematical term.

For the purposes of this dissertation, therefore, I will define an “objective” probability as an expression of the “true” likelihood of an uncertain event, as dictated by the (perhaps latent) causal and/or stochastic structure of the decision environment. Or, in slightly simpler terms, an “objective” probability is the likelihood that a rational actor would ascribe to an event, given complete information about the true state of nature. In practice, of course, it is unlikely that a decision maker would be able to observe such a likelihood. However, the prevailing neoclassical assumption is that uncertain events are fundamentally characterizable objects that *could, in principle*, be described by a number that contains the same information as a mathematical probability (i.e., a number that

describes the long-run relative frequency that one *would* observe if counterfactual instances of the event ran to infinity). Though practically and philosophically fraught, the value of this definition is purely conceptual. Namely, it makes clear the assumptions that (a) there is a *true* answer to the question (e.g.) “what is the probability that Candidate X will win the presidential election?” and (b) that it makes sense for a decision maker to think about this “true answer” (and therefore, plan her future behaviors) *as if* it were a mathematical probability — even if it can never be expressed as a relative frequency. Thus, while it may often be difficult to *identify* “objective” probabilities, it is reasonable to assume that such a criterion exists, and that SPJs can be regarded as “better” or “worse” to the extent that they agree with this standard.

In practice, of course, the problem with defining an SPJ’s validity in terms of “agreement with the truth” is that decision makers generally don’t know what the truth *is*. In some cases, this may be because the “true state of nature” is not adequately reflected in the full scope of information that is observable at a given time (i.e., the diagnostic cues necessary for estimating an event’s “true” probability are absent, weak, and/or biased); in others, it may be because a judge does not have access to the full scope of extant information. In either scenario, however, real-world decision makers are unlikely to have access to complete and unbiased information about the events that bear on their decisions. Thus, while a conceptual definition of “objective” probability is useful in that it makes clear the *sort of information* that a “good” judgment should provide, it is still unlikely to serve as a useful criterion by which to evaluate SPJs.

Critically, however, if a rational actor *were* privy to the full scope of extant information, he or she could use inferential statistical procedures (e.g., Bayes' rule; see: Edwards & Fasolo, 2001) to identify the most *epistemically defensible* belief that an individual could hold, given the information in the decision environment. Though it is possible that the corresponding SPJ would depart from the event's "true" probability (especially if the pool of extant information were biased in some way), this judgment would be "optimal" in the sense that it maximizes "agreement with the truth" to the greatest extent that evidence and reason allow. From a conceptual perspective, therefore, this kind of "optimal" judgment represents a ceiling on the validity of real-world judgments and can be used as a reference point for examining the "quality" or *relative validity* of other judgments (i.e., the absence of errors and biases, relative to this reference point). Thus, if "optimal" judgments can be estimated empirically, then the resulting *estimated optima* would be a desirable benchmark by which to evaluate SPJs.

Fortunately, research in the decision sciences has identified a variety of methods for calculating (or eliciting) *estimated optima*. Indeed, depending on the situation, a decision maker might opt to use (a) Bayesian updating (see: Edwards & Fasolo, 2001); (b) belief aggregation or crowdsourcing (e.g., Mellers, Ungar, Baron, Ramos, Gürçay, Fincher, Scott, Moore, Atanasov, Swift, Murray, Stone, & Tetlock, 2014; see also: Armstrong, 2001); (c) prediction markets (e.g., Atanasov, Rescober, Stone, Swift, Servan-Schreiber, Tetlock, Ungar, & Mellers, 2017; Cowgill & Zitzewitz, 2015; Spann & Skiera, 2003); (d) actuarial or statistical models (e.g., Meehl, 1959; Goldberg, 1968; see also: Dawes, Faust, & Meehl, 1989); (e) machine learning algorithms (Breiman, 1996;

Freund & Schapire, 1996; see also: Kaelbling, Littman, & Moore, 1996); or (f) statistical optimizations, combinations, or ensembles of any of the above (e.g., Baron, Mellers, Tetlock, Stone, & Ungar, 2014; Grushka-Cockayne, Jose, & Lichtendal Jr., 2017; Flowerdew, 2014; see also: Sawyer, 1966) — all of which I discuss in greater detail in Chapter 1. Regardless of how a decision maker arrives at these judgments, however, the implication is the same. If the judgments produced by these methods can be generally assumed to be “better” (i.e., less biased and/or prone to error) than the judgments one wishes to evaluate, then decision makers can use these *estimated optima* as an informative reference point by which to evaluate the *relative* degree of error and/or bias associated with SPJs.

For the purposes of this dissertation, therefore, I will define *credibility* as a conceptual construct that relates an individual’s SPJs (i.e., “what she said”) to the most epistemically defensible judgments that one could have provided, given the information in the decision environment (i.e. “optimal” judgments, or what she “would have said” if she had had access to more or better information). In practice, of course, empirical models of credibility are unlikely to capture this relationship perfectly and their validity will be limited by factors such as (a) the descriptive fit of the model relating an individual’s SPJs to *estimated optima*; (b) the stability and generalizability of this model’s fit; and (c) the latent validities that relate *estimated optima* to “optimal” judgments; and “optimal” judgments to “objective probabilities.” Assuming that none of these factors is egregiously lacking, however, a well-specified model of credibility ought to provide decision makers with (a) descriptive insight into individual  $i$ ’s “quality” (i.e.,

relative validity) as a source of probabilistic information; and (b) a statistical procedure for shrinking individual  $i$ 's judgments towards *estimated optima* (i.e., recalibrating individual  $i$ 's judgments to account for observed errors and biases), even when *estimated optima* are no longer available (for more information about why the latter is valuable, see the section titled “Why Examine Credibility?” in Chapter 1).

For the remainder of this dissertation, I will examine the extent to which empirical models of credibility can deliver these types of information.

## References

- Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision*, 3(1), 20-39.
- Armstrong, J. S. (2001). Combining forecasts. *Principles of forecasting: a handbook for researchers and practitioners*. J. S. Armstrong (Ed.). Norwell, MA: Kluwer Academic Publishers.
- Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3), 691-706.
- Barber, B. & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment, *Quarterly Journal of Economics*, 116(1), 261–291.
- Baron, J. (2008). *Thinking and deciding*. New York: Cambridge University Press.
- Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2), 133-145.
- Bernoulli, D. (1738 / 2011). Exposition of a new theory on the measurement of risk. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, 11-24.
- Birnbaum, M. H., & Schmidt, U. (2008). An experimental investigation of violations of transitivity in choice under uncertainty. *Journal of Risk and Uncertainty*, 37(1), 77.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140.
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193.
- Christensen-Szalanski, J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, 7(4), 928-935.
- Cowgill, B., & Zitzewitz, E. (2015). Corporate prediction markets: Evidence from Google, Ford, and Firm X. *The Review of Economic Studies*, 82(4), 1309-1341.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899), 1668-1674.

- Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica*, 43(2), 111-121.
- Dunwoody, P. T. (2009). Theories of truth as assessment criteria in judgment and decision making. *Judgment and Decision Making*, 4(2), 116.
- Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.
- Edwards, W., & Fasolo, B. (2001). Decision technology. *Annual Review of Psychology*, 52(1), 581-606.
- Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4), 643-669.
- Elster, J. (Ed.). (1986). *Rational Choice*. Washington Square, NY: New York University Press.
- Flowerdew, J. (2014). Calibrating ensemble reliability whilst preserving spatial structure. *Tellus A: Dynamic Meteorology and Oceanography*, 66(1), 1-20.
- Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In *Machine learning: proceedings of the thirteenth international conference* (pp. 325-332). San Francisco: Morgan Kauffman.
- Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 23(7), 483.
- Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendal Jr., K. C. (2017). Ensembles of overconfident and overfit forecasts. *Management Science*, 63(4), 1110-1130.
- Hammond, K. R. (2007) *Beyond rationality: The search for wisdom in a troubled time*. New York: Oxford University Press.
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 237-285.
- Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, 58(9), 697-720.

- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Strauss and Giroux.
- Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences*, 111(30), 10984-10989.
- Meehl, P. E. (1959). A comparison of clinicians with five statistical methods of identifying psychotic MMPI profiles. *Journal of Counseling Psychology*, 6(2), 102.
- Mellers, B., Ungar, L., Baron, J., Ramos, J., Gürçay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5), 1106-1115.
- Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, 111(20), 7176-7184.
- Ramsey, F. P. (1926). Truth and probability. *The foundations of mathematics and other logical essays*. New York: Harcourt Brace.
- Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
- Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, 66(3), 178-200.
- Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, 65(2), 117-137.
- Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 299-314.
- Spann, M., & Skiera, B. (2003). Internet-based virtual stock markets for business forecasting. *Management Science*, 49(10), 1310-1326.
- Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, 16(3), 167-171.
- Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, 102(2), 269.

- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4), 293.
- Von Neumann, J., & Morgenstern, O. (1945). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
- Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language*, 25(5), 571-587.
- Windschitl, P. D., & Weber, E. U. (1999). The interpretation of "likely" depends on the context, but "70%" is 70%—right? The influence of associative processes on perceived certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(6), 1514-1533.

## CHAPTER 1

### LINEAR REGRESSION CAN PROVIDE VALID AND USEFUL ESTIMATES OF CREDIBILITY IN SIMULATED DATA

#### **Abstract:**

As a conceptual construct, *credibility* is intended to provide two types of information: (a) a description of an individual’s “quality” as a source of probabilistic information; and (b) a statistical procedure for recalibrating her subjective probability judgments (SPJs). The goal of this dissertation is to determine whether empirical models of credibility can provide these types of information. To establish a preliminary lower-bound for the performance of empirical models of credibility, this chapter will examine the performance of an intentionally weak model of credibility that I call the *linear credibility framework*. Across two analyses, I examine the validity and utility of credibility estimates derived from 3150 simulated forecasters. These analyses demonstrate that linear credibility estimates are (a) predictive of “true” levels of credibility; and (b) can be used to improve the accuracy of SPJs out-of-sample. Based on these results, I conclude that examining credibility may often be beneficial and cost-effective.

#### **Introduction**

As a conceptual construct, credibility describes the relationship between person  $i$ ’s subjective probability judgments (SPJs) and the most epistemically defensible beliefs that one could hold, given the information in the decision environment (i.e. “optimal” judgments, which are empirically approximated by *estimated optima*; for a detailed discussion of these terms, see the introduction to this dissertation). Though necessarily less informative than a model that relates an individual’s SPJs to “true” or “objective” probabilities, models of credibility are intended to provide insight into an individual’s relative tendency towards error and bias in judgment. When used to describe the performance of a specific judge (or to compare the performance of several judges), however, it is often useful to frame this relationship in terms of the *absence* of errors and

biases. Thus, it is also reasonable to interpret credibility estimates as empirical measures of “skill” or “proficiency” in subjective probability judgment. Depending on how this relationship is framed, therefore, a well-specified model of credibility ought to provide one of two types of information: either (a) a description of an individual  $i$ ’s “quality” (i.e., relative validity) as a source of probabilistic information; or (b) a statistical procedure for recalibrating an individual  $i$ ’s SPJs by “undoing” or “correcting for” historical tendencies towards error and bias.

As described in the introduction to the dissertation, the goal of the present research is to determine whether empirical models of credibility can deliver these types of information. To accomplish this goal, however, it is first necessary to operationalize credibility as an empirical construct. Depending on (a) a decision maker’s goals; and (b) the pragmatic constraints of the decision environment (e.g., access to computational resources; the scope and validity of available information), there are a variety of ways that a decision maker might do this. In one case, a decision maker might model credibility by fitting a power function to the relationship between an individual’s SPJs and *estimated optima* derived from an online crowdsourcing platform. In another, a decision maker might get better results by fitting a series of splines to *estimated optima* derived from an online prediction market. And in a third case, a decision maker might find that the most parsimonious model of credibility is one that fits a hierarchical linear regression to actuarial estimates of risk. Indeed, because credibility is a broad, conceptual construct rather than a specific set of empirical procedures, there are endless variations in how credibility might be modeled. Setting aside the specifics, however, operationalizing

credibility requires a decision maker to commit to two empirical procedures: (a) a method for modeling, calculating, or identifying *estimated optima* (i.e., empirical approximations of “optimal” judgments); and (b) a method for relating *estimated optima* to an individual’s SPJs.

### **Methods for Identifying *Estimated Optima***

Over the course of the past century, decision scientists have developed a wide variety of methods for eliciting, combining, and optimizing SPJs (for an overview, see: Armstrong, 2001). In large part, these methods were developed to aid in organizational decision making, where individual decision makers (e.g., managers, policy-makers, government officials) are responsible for synthesizing “optimal” judgments from the beliefs and opinions of their associates (e.g., colleagues, analysts, subject-matter experts). Depending on (a) the field from which a method was drawn (e.g., forecasting vs. risk assessment); and (b) the technology that was available at the time (e.g., basic computing vs. modern machine learning), each of these methods takes a slightly different approach to the combination and/or optimization of noisy, disparate, and sometimes correlated beliefs. In all cases, however, the purpose of these methods is to help decision makers identify the “best” possible SPJ (or distribution of SPJs), given the information that is available at the time (e.g., Budescu & Rantilla, 2000; Clemen & Winkler, 1999; Wallsten, Budescu, Erev, & Diederich, 1997). For the purposes of estimating credibility, therefore, the decision science literature provides a variety of ready-made methods for

identifying *estimated optima* (for a comprehensive summary/review of these methods, see: e.g., Armstrong, 2001; Clemen & Winkler, 1999).

Among these methods, perhaps the most well-known is Bayesian updating. As a mathematical procedure, Bayes' rule can be used to calculate the probability of an event  $A$ , given information about an event (or state of the world)  $B$ , that bears on its outcome. When used several times in succession, Bayes' rule can be used to “update” one’s beliefs about the likelihood of event  $A$ , given new information about the conditions under which  $A$  might occur. Thus, Bayesian updating is widely regarded as a normative method for belief formation (i.e., combining information about uncertain events to arrive at an SPJ) (Edwards & Fasolo, 2001) and Bayesian probability estimates are often used as a benchmark of “rational” judgement (see also: Laming, 2010; Edwards, Lindman, & Savage, 1963). Under ideal conditions, therefore, Bayesian updating represents a strongly defensible approach to identifying *estimated optima*.

Critically, however, Bayesian updating relies on (at least approximate) knowledge of (a) the baserates with which events  $A$  and  $B$  occur; and (b) the degree to which event  $B$  lends support to event  $A$ . Thus, in many cases, decision makers do not have access to the necessary information to arrive at *estimated optima* via Bayesian updating. To circumvent this problem in real-world decision making, practitioners often have little choice but to rely on some form of intersubjective agreement (i.e., “pooled” or “common” knowledge) to identify *estimated optima*. Fortunately, research on probabilistic forecasting has demonstrated that even elementary applications of “crowdsourcing,” or the large-scale belief aggregation can produce surprisingly accurate

*estimated optima*. Indeed, the general conclusion from this line of work is that one of the most robust methods for identifying “optimal” SPJs is to take a simple, unweighted average of beliefs across individuals (e.g., Armstrong, 2001; Rantilla & Budescu, 1999; Clemen, 1989; Wallsten et al., 1997; Genest & Zidek, 1986). In some cases, researchers have demonstrated that the accuracy of this type of *unweighted linear opinion pool* can be improved by weighting SPJs according to an individual’s (e.g.) statistical dependence with other judges (e.g., Hora & Kardeş, 2015); tendency towards probabilistic coherence (e.g., Karvetski, Olson, Mandel, & Twardy, 2013); and other measures of judgmental performance. In most cases, however, comparisons of weighted vs. unweighted aggregates have shown that weighting procedures rarely improve the accuracy of unweighted averages by more than 20% (Armstrong, 2001).

In addition to simple SPJ aggregation (or, as is sometimes common, the combination of subjective probability *distributions*; for reviews, see: Gneiting & Ranjan, 2013; Jacobs, 1995), researchers in the decision sciences have also developed a variety of methods for debiasing crowdsourced aggregates. In most cases, these techniques are intended to correct for biases that emerge in the aggregation process (e.g., as the result of correlated beliefs across forecasters), though there is at least one case where researchers have attempted to debias SPJs at the individual level (Turner, Steyvers, Merkle, Budescu, & Wallsten, 2014).<sup>1</sup> Though far from exhaustive, examples of such

---

<sup>1</sup> While similar, in spirit, to the recalibration analyses I conduct in Chapters 1 and 3, it is worth noting that Turner et al. (2014) rely on an approach to recalibration that is considerably less feasible for real-world decision makers. For additional arguments to this effect, see the section below titled “Why examine credibility?”

optimization techniques include those proposed by Chen, Fine, & Huberman (2004), who correct for public knowledge biases among small groups of judges; Budescu & Yu (2007), who account for latent correlations among judges and informational cues (for a conceptually similar procedure in a different context, see: Basili & Pratelli, 2014); Budescu & Chen (2014), who weight individual's judgments according to their historical accuracy, relative to the "crowd;" and Baron, Mellers, Tetlock, Stone, & Ungar (2014), who statistically correct for individuals' tendency to shrink their SPJs towards 0.50 when missing key pieces of information (resulting in underconfident crowd aggregates; for related methods, see: Cross, Ramos, Mellers, Tetlock, & Scott, 2018; Satopää, Baron, Foster, Mellers, Tetlock, & Ungar, 2014).

In a conceptually similar vein, the decision science literature also highlights an economic method for identifying *estimated optima*: large-scale prediction markets. In general, prediction markets allow participants to buy and sell fixed-payout contracts related to the outcome of an uncertain event. Drawing on economic theory, observers of these markets typically assume that free-market forces will drive the price of such contracts towards a (per dollar) value that closely approximates the event's underlying probability (Atanasov, Rescober, Stone, Swift, Servan-Schreiber, Tetlock, Ungar, & Mellers, 2017; Cowgill & Zitzewitz, 2015; Spann & Skiera, 2003). Beginning with the research of Hayek (1945), studies of prediction markets have shown that free-market structures can be used to aggregate information across a diverse group of individuals, and that the resulting "SPJs" (i.e., prices) are often more accurate than both expert judgments (e.g., Gürkaynak & Wolfers, 2006; Plott & Chen, 2002; Pennock, Lawrence, Nielson, &

Giles, 2001) and simple linear averages (e.g., Rothschild, 2009; Berg, Forsythe, Nelson, & Rietz, 2001). Consequently, while prediction markets are less common than survey-style aggregation, they tend to be a robust method for eliciting *estimated optima* (Atanasov et al. 2017).

Finally, thanks to recent advances in computer technology, research in the decision sciences has begun to explore methods for optimizing SPJs though purely statistical means. For example, work in areas such as risk assessment, medicine, and finance have demonstrated that machine learning algorithms can be used to estimate the likelihood of uncertain events such as groundwater contamination (e.g., Sajedi-Hosseini, Malekian, Choubin, Rahmati, Cipullo, Coulon, & Pradhan, 2018); medical outcomes (e.g., Choo, Uhmnn, Kim, Han, Kim, Kim, & Lee, 2018); and financial crashes (Chatzis, Siakoulis, Petropoulos, Stavroulakis, & Vlachgiannakis 2018). Furthermore, research in the rapidly expanding literatures on model aggregation (for an introduction, see: Hoeting, Madigan, Raftery, & Volinsky, 1999), model ensembling (e.g., Grushka-Cockayne, Jose, & Lichtendahl Jr., 2017; Flowerdew, 2014), and the synthesis of clinical and actuarial judgments (e.g., Nagar, 2013; Blattberg & Hoch, 1990; Peterson & Pitz, 1986; for an overview, see: Sawyer, 1966) suggests that there is a degree of truth to the neoclassical adage that (with enough information) “*all* uncertainties can be reduced to *risks*” (Ellsberg, 1961, p. 645; emphasis in the original; attributed by the source to Ramsay, 1926). Indeed, while the relative level of performance demonstrated by these approaches depends on a variety of factors (e.g., the outcome being predicted; the exact model or procedure that was used), the general conclusion from these lines of work is that highly

accurate SPJs can be extracted from extremely noisy decision environments, given the right tools and types of information (e.g., Kruppa, Ziegler, & König, 2012; Gerds, Cai, & Schumacher, 2008).

### **Methods for Relating *Estimated Optima* to an Individual's SPJs**

From a statistical perspective, credibility describes the relationship between two sets of judgments. Thus, it is reasonable to assume that credibility might be best described by a statistic that measures covariation (e.g., Pearson correlation), agreement (e.g., intraclass correlation), or correspondence (i.e., accuracy; e.g., mean absolute error; strictly proper scoring rules). Indeed, these sorts of measures are frequently used in the social, ecological, and behavioral sciences to describe the relationship between noisy sets of judgments. In studies of coding and classification, for example (e.g., automated object recognition: da Silva & Garcia, 2006; content analysis: Shannon, Hsieh, & Shannon, 2005; behavioral coding: Funder & Sneed, 1993), researchers often rely on measures of inter-rater agreement to determine whether there is consensus among various sources of judgment (e.g., one judge vs. all others; machine classifications vs. human judgments; etc.). Similarly, in domains such as meteorology (e.g., Wang, Ding, Fu, Kang, Jin, Shukla, & Doblas-Reyes, 2005), public health (e.g., Hrust, Klaic, Krizan, Antonic, & Hercog, 2009) and climate science (e.g., Massonet, Bellprat, Guemas, & Doblas-Reyes, 2016), correlations between predictive models and empirical measurements are often used as evidence for model validity. And finally, in research on ecological judgment (see: Brunswik, 1955), expertise (e.g., Tetlock, 2005; Olsen, 1997; Ebbesen & Konecni, 1975),

and clinical assessment (see: Dawes, Faust, & Meehl, 1989), decision scientists have often used measures of correspondence to assess the accuracy of intuitive judgments.

In practice, however, measures of covariation, agreement, and correspondence only capture the “big picture” when it comes to credibility. In other words, while these types of measures can provide insight into an individual’s “quality” (i.e., relative validity) as a source of probabilistic information (relative to *estimated optima*), they generally cannot provide a detailed description of *how* and *why* an individual’s judgments tend to err. Thus, in many cases, a decision maker may gain more insight by using a curve with a known functional form (e.g., a line, power-function, spline, sinusoid, etc.) to estimate credibility. Though it is likely that the most descriptive functional form will vary from one context to another, a curve-fitting approach to credibility estimation is valuable because it provides a concrete, mathematical expression of the relationship between an individual’s SPJs and *estimated optima*. Thus, to the extent that an empirical model of credibility is well-specified (i.e., statistically appropriate and reasonably descriptive), modeling credibility as a functional relationship can provide decision makers with (a) a quantitative basis for estimating an individual’s tendencies towards errors and bias in judgment (or, at the very least, a quantitative basis for forming expectations about her accuracy); and (b) a statistical procedure for shrinking her SPJs towards *estimated optima*, thereby reducing errors and biases.

For an empirical model of credibility to provide these types of information, however, it is necessary for that model to provide a reasonably descriptive fit to the data. Thus, examining credibility will often require decision makers to conduct some degree of

statistical exploration to determine which of several models provides the best fit to their data (for a more in-depth discussion of this topic, see: Chapter 4). Depending on the context, this could mean that the “best” model of credibility might range from a simple line to a highly parameterized and/or deeply nested hierarchical regression. As a result, this means that (a) the costs of examining credibility (i.e., the time, data, and computational resources necessary to estimate the model) might vary widely across decision environments; and (b) that the benefits of extracting credibility information might not outweigh the costs.

Fortunately, there are at least two reasons to be optimistic about the robustness — and therefore, value — of simple models of credibility. First, existing research suggests that *source credibility* in expert judgment (i.e., the line of work that inspired the use of the term *credibility* in this dissertation) can be usefully conceptualized in terms of simple statistical models. For example, Birnbaum & Stegner (1979) suggest that source credibility can be reasonably operationalized in terms of linear regression, where:

“The *expertise* of a source refers to the perceived correlation between the source’s report and the outcomes of empirical verification...the *bias* of the source refers to factors that are perceived to influence the expected algebraic difference between the source’s report and the true state of nature...[and] the distinction between expertise and bias is like the distinction between regression slope and intercept.” (Birnbaum & Stegner, 1979, p. 48; emphasis in the original).

Though it is unclear whether Birnbaum & Stegner (1979) intend for this analogy to be interpreted as an *argument* for linear regression as model of source credibility, they then go on to use linear regression as a basis for formulating (and testing) hypotheses

about the effects of *bias* and *expertise* on the interpretation of a source’s “reports,” given various theories about how judges (i.e., those asked to interpret the source’s “reports”) weight and combine information. Consequently, while linear regression has not been used to model credibility in previous research *per se*, there appears to be at least basic consensus (as provided by the peer review process) that something as simple as a linear equation can be used to describe the “quality” of an individual’s judgments.

Perhaps more importantly, however, the second reason to be optimistic about the robustness of simple models of credibility is that simple statistical models tend to be robust, in general. Indeed, as long as the relationship between two variables is (essentially) monotonic, basic descriptive techniques such as linear regression can often provide a reasonable summary of their covariation—even in cases where the “true” relationship is known to be non-linear. Furthermore, a large body of research on clinical vs. actuarial judgment has demonstrated that (a) standard linear regression; and (b) even intentionally agnostic/atheoretical models (e.g., equal-weights regression) can be used to combine noisy and poorly-behaved data to arrive at surprisingly accurate judgments (Dawes, 1979; Meehl, 1954). Thus, decision makers are likely to see benefits from even the most rudimentary models of credibility and—when their resource-constraints allow—can use any variety of (more sophisticated) techniques to identify a model that suits their needs. In general, therefore, while it is unlikely that (e.g.) linear regression will be the “best” way to examine credibility, there is reason to expect that (with a little exploration) simple statistical tools can be used to extract a useful degree of information about credibility.

## **Why Examine Credibility?**

Given that the decision science literature provides a variety of techniques for identifying *estimated optima* (i.e., judgments that are typically expected to be more accurate than an individual’s SPJs), it is natural to wonder why decision makers would be interested in examining credibility. In general, the answer to this question is threefold:

1. First, while the methods discussed above may help decision makers identify highly accurate SPJs, they typically do not provide any information about the “quality” (i.e., relative validity) of an individual’s judgments. Consequently, in cases where a decision maker is interested in assessing the “skill” or “proficiency” of an individual judge— perhaps for the purposes of accountability, performance evaluation, or targeted interventions (e.g., training or debiasing)— he or she can only do so by examining (something like) credibility.
2. Second, a non-trivial proportion of the methods described above rely on costly information-gathering strategies. Thus, while *estimated optima* from an online prediction market might be considerably more accurate than the judgments provided by a handful of individuals, “crowdsourcing” these types of *estimated optima* typically requires (a) a significant up-front investment (e.g., recruitment efforts, website coding, data-handling protocols); and (b) non-

trivial ongoing expenditures (e.g., participant retention efforts; participant compensation).

3. Finally, many of the methods described above depend on (a) uncharacteristically rich data-sets; (b) information that is generally not available in real-world decision environments; and/or (c) information that a decision maker can only observe *ex post*. Thus, even in cases where it is feasible to rely on “crowdsourced” aggregation methods, many such approaches are most effective when they can be optimized against external criteria (e.g., historical baserates; observed outcomes). In general, therefore, many of the methods discussed above are of limited use to real-world decision makers.

Of course, because credibility estimation relies on the identification of *estimated optima*, the latter two conditions also limit the applicability of empirical models of credibility. Critically, however, if a model of credibility is well-specified (i.e., if it provides genuine insight into an individual’s tendencies towards error and bias in judgment), then the relationship captured in one set of judgments ought to generalize to other SPJs made by the same judge in the same domain. In principle, therefore, it should be possible for a decision maker to use an ecologically realistic amount of data to (a) identify a small set of *estimated optima*; and (b) use these judgments to build an empirical snapshot of credibility. Though far from foolproof, the curve captured in this

snapshot should allow decision makers to “estimate” or “reproduce” *estimated optima* from an individual’s noisy SPJs—or, more realistically, to shrink her SPJs towards *estimated optima* when *estimated optima* are no longer available. Thus, if fit to a suitably representative sample of data, empirical models of credibility should allow a decision maker to “correct-for” or “undo” (at least some degree of) an individual’s historical tendencies towards error and bias in judgment. To test whether empirical models of credibility can deliver on this promise, I will carry out this exact sort of recalibration analysis in the chapters that follow.

### The Linear Credibility Framework

Because decision making is not a single task with a single set of constraints (e.g., risks; error tolerances; costs and benefits), it is difficult to make the case that decision makers would benefit from examining credibility, in general. As a starting-point for this conversation, however, it is instructive to consider that—at a bare minimum—an empirical model of credibility must provide decision makers with a positive expected utility to be deployed. In other words, to make the case for a *specific* model of credibility, it is necessary to convince a decision maker that deploying it is “cost-effective” in the sense that it is expected to provide a large informational “return” on his or her statistical (and perhaps logistical) “investment.” In operational terms, this provides three criteria for a (minimally) well-specified model of credibility. Specifically, to be worth examining, an empirical model of credibility must: (i) be applicable to the decision environment at hand (i.e., must not rely on ecologically unrealistic amounts of types of data); (ii) produce

generally reliable, valid, and useful estimates; and (iii) provide a generally favorable cost-benefit ratio.

Given these criteria, one way to make a case for the value of credibility information, *in general* is to “stack the deck” against my own research and examine the performance of a model whose cost-benefit ratio is intentionally low. If, despite these adverse conditions, the value of the information provided by such a model still outweighs its costs, then it is largely untenable to argue that a decision maker should not examine credibility. Critically, however, because the “costs” of examining credibility are largely determined by factors that are within a decision maker’s control (e.g., the complexity of one’s statistical model; the resource-intensiveness of one’s method for identifying *estimated optima*), examining the performance of an arbitrarily costly model (vs. an arbitrarily weak model) is unlikely to be informative. Indeed, because decision makers have an inherent incentive to minimize costs, the most representative model of credibility is likely to be the simplest (i.e., “cheapest”) model that will serve a decision maker’s goals. Thus, to make a case for the value of examining credibility *in general*, the most informative model I can test is the simplest, least costly, and least-likely to be informative model that a decision maker might reasonably apply.

Following from this line of reasoning, the research presented in this dissertation will test the performance of a credibility model that uses (a) the aggregation method developed by Baron et al. (2014) to identify *estimated optima*; and (b) a simple, main-effects-only approach to linear regression to relate *estimated optima* to an individual’s SPJs. For simplicity, I will refer to this model as the *linear credibility framework*, and to

the measures it produces (i.e., the empirical estimates of this model’s intercept, slope, and standard error) as *linear credibility estimates*. In practice, this model is among the simplest that a decision maker could devise. However, thanks to the *estimated optima* used in this model, it is not necessarily an unrealistic one in terms of either (a) the types and amount of data that are necessary to deploy it; or (b) the expected validity of its *estimated optima*.

Indeed, in simulation studies, Baron et al.’s (2014) aggregation method (and others like it: e.g., Satopää et al., 2014) have been shown to produce remarkably precise crowd forecasts. In Baron et al.’s (2014) study, this was accomplished by statistically correcting for the effects of individual-level *regression* on simple crowd averages: i.e., the empirical tendency for individuals to shrink their estimates towards 0.50 when missing key pieces of information. By adjusting for this tendency, Baron et al.’s model (drawn from a sample of 100 simulated judgments) was able to produce fitted-values (i.e., *estimated optima*) that were never more than 0.005 units away from the simulation’s “true” probability values, as measured on the probability scale. Thus, while unlikely to capture nuanced differences in credibility, the linear credibility framework is likely to provide an informative lower-bound for the performance of empirical models of credibility.

## **What to Expect from the Linear Credibility Framework**

Due to its simplicity, the linear credibility framework is unlikely to be an especially descriptive (or ecologically valid) model of credibility. Indeed, from a

theoretical perspective, the linear credibility framework presents a variety of weaknesses that decision makers may be reluctant to ignore. In most cases, for example, decision makers have little basis for speculating about the nature of the relationship between an individual’s judgments and the corresponding *estimated optima* (i.e., in general, there is little reason to expect one functional form to be more likely than another). Furthermore, if such a relationship exists, there is no strong reason to believe that it will be stable or generalizable. And finally, even if such a relationship is stable, there is no strong reason to think that it can be estimated from an ecologically realistic sample of data. Thus, even if the *theory* of credibility is sound, the value of examining it with simple statistical tools may not be.

Despite these challenges, however, the linear credibility framework presents a useful starting place for examining the performance of empirical models of credibility. Specifically, because the “true” relationship between an individual’s judgments and *estimated optima* is unlikely to be linear, any benefits that I uncover by adopting this approach can be treated as an empirical lower bound for the performance of more sophisticated (or, at the very least, better-specified) models of credibility. Thus, while I generally do not expect the linear credibility framework to provide veridical, face-value, or even particularly accurate measures of credibility, it is my hope that these measures will serve as useful predictors of “skill” or “proficiency” in subjective probability judgment—and thus, be useful to decision makers.

## **Estimating an Individual’s Credibility Function**

To estimate an individual's credibility, I will regress a small sample of her SPJs on a corresponding set of optimized crowd aggregates, calculated using the method developed by Baron et al. (2014; see below for details). For simplicity, I will call this regression a *credibility function*, as it describes the relationship between an individual's judgments (i.e., what she "said") and a set of model judgements or *estimated optima* (i.e., what an observer might infer she "should have said," or "would be justified in believing," given access to additional information). In an ideal scenario, both sets of judgments would contain the same information, and an individual's credibility function would closely approximate identity. Thus, if fit on a suitably representative sample of data, the various features of this model (and their departures from identity) should correspond to the essential features of the individual's credibility as a judge of subjective probability.

From an operational perspective, this provides us with three predictions about a maximally credible judge. If an individual's judgments contain the same information as *estimated optima*, then his or her credibility function should have an intercept of zero, a slope of one, and a standard error of zero. Or, to extend the analogy provided by Birnbaum & Stegner (1979), the judgments of a maximally credible judge should not exhibit any algebraic bias, relative to *estimated optima*; should indicate expertise in the sense that they are appropriately scaled, relative to *estimated optima* (and thus, exhibit a perfect correlation); and should be consistent in the sense that their relationship to *estimated optima* is both stable and descriptive at the judge  $\times$  domain level. For the purposes of the present research, therefore, an individual  $i$  will be said to be more *consistent* in subjective probability judgment to the extent that the standard error of her

credibility function ( $\hat{\sigma}_i$ ) is close to zero; less *biased* to the extent that the estimated intercept ( $\hat{\alpha}_i$ ) is close to zero; and more *expert* to the extent that the estimated slope ( $\hat{\beta}_i$ ) is close to 1.<sup>2</sup>

As the term credibility implies, each of these measures is intended to represent a different aspect of the extent to which an individual's judgments should be taken at face value. Due to the properties of the probability scale, however, several statistical transformations will be necessary to ensure that estimates of *bias* can be meaningfully interpreted. Specifically, because the probability scale is bounded at 0 and 1, linear regression is (a) an inappropriate descriptive model for probabilistic data; and (b) unlikely to provide face-valid estimates of *bias* because its intercept reflects an unrepresentative case — namely, the difference between an individual's SPJs and *estimated optima* when *estimated optima* are equal to (a probability of) zero. In addition, there are many cases where SPJs are coded to reflect an individual's beliefs about the (subjectively) more likely outcome of a binary event. When dealing with such data, it is impossible for SPJs to fall below 0.5 and therefore nonsensical to discuss the case where *estimated optima* are equal to zero.

To resolve these issues, it will be necessary to (a) ensure that all SPJs are coded on a 0-1 scale (i.e., that they reflect an individual's beliefs about the likelihood of a *given* outcome rather than the *most likely* outcome of a binary event); and (b) estimate *bias*, *expertise*, and *consistency* after converting all SPJs (and *estimated optima*) to log-odds.

---

<sup>2</sup> Note that the use of the terms *bias* and *expertise* in this research are borrowed from the existing literature on source credibility (e.g., Birnbaum & Stegner, 1979).

By applying these transformations, SPJs ranging from 0-1 probability can be mapped onto a continuous, unbounded scale with a log-odds of zero corresponding to a probability of 0.5. As such, estimating credibility in log-odds space confers several advantages. First, linear regression is now an appropriate model for describing log-odds data; and second, the intercept of such a regression now describes the difference between an individual's SPJs and *estimated optima* at a value corresponding to a probability of 0.5 — i.e., the average value along the 0-1 probability scale. As such, estimates of *bias* derived from linear regression in log-odds space can be directly (and intuitively) interpreted as face-valid descriptions of the expected arithmetic difference between an individual's judgments and the corresponding *estimated optimum*, and estimates of *expertise* and *consistency* can be interpreted in the same manner as before (with the single exception that all three components of credibility are now measured on the log-odds scale).

Despite these advantages, however, it is still an empirical question whether linear regression can serve as an adequate model of credibility. Indeed, because the linear credibility framework is built on several rather fragile assumptions (e.g., stability; linearity; the validity of *estimated optima*), it is essential to demonstrate that linear regression can provide valid and/or informative estimates of credibility. As a proof-of-concept, therefore, I will begin with a series of simulations. Following the procedures described below, I will first evaluate the validity<sup>3</sup> of credibility estimates in simulated

---

<sup>3</sup> I will not examine the reliability of credibility estimates in Study 1 because reliability is strongly influenced by the *a priori* error parameters used in simulation. Because the ranges of these errors are

data, and then examine the degree to which credibility-based recalibration can be used to improve (i.e., increase the accuracy and/or validity of) simulated judgments.

### **Study 1: Linear Credibility Estimation with Simulated Data**

To determine whether the linear credibility framework can provide useful and/or valid estimates of credibility in complex decision environments, I conducted a series of simulations. In each simulation, I generated 100 sets of “subjective probability judgments” by applying four types of error to a set of known (i.e., *a priori*) probability values. Across simulations, I varied the magnitude of these errors according to a  $[10 \times 5 \times 9 \times 7]$  factorial design (see below). In total, this yielded 3150 unique simulations, each of which utilized a different combination of error parameters. For simplicity, I will refer to each simulation as a “forecaster,” as each combination of error parameters was intended to simulate the SPJs of a unique individual in a given decision environment.

The purpose of these simulations was to examine the typical performance of the linear credibility framework across all 3150 forecasters. For simplicity, I will break my discussion of this study’s results into two *analyses*, each of which examines a different aspect of the credibility estimation procedure and answers different questions about its effects.

---

arbitrary, it is likely that some or all of the simulated decision environments in Study 1 are unrealistically noisy. In such environments, it is uninteresting to note that credibility estimates are unreliable, as *all* information in the environment is unreliable. Thus, I will postpone my examinations of reliability until Chapter 2.

1. In Analysis 1a, I examined the validity of bootstrapped credibility estimates ( $\hat{\alpha}_i$ ,  $\hat{\beta}_i$ , and  $\hat{\sigma}_i$ ) as measures of the *a priori* error parameters used in simulation. Given the complexity of Study 1's decision environment, it was unlikely that linear credibility estimates would recover an individual's "true" credibility values in this analysis (i.e., it was unlikely that linear credibility estimates would be especially accurate or face-valid measures of *bias*, *expertise*, and *consistency*). In principle, however, the strength of the covariation between each pair of variables (i.e.,  $\hat{\alpha}_i$  and *a priori bias*;  $\hat{\beta}_i$  and *a priori expertise*;  $\hat{\sigma}_i$  and *a priori noise*, respectively) should provide a preliminary indication of the linear credibility framework's robustness. Thus, the purpose of Analysis 1a was to shed light on the extent to which simple statistical models can be used to model credibility.
  
2. In Analysis 1b, I used credibility information to recalibrate each forecaster's SPJs, out-of-sample. To do so, I leveraged the mathematical relationship captured by each forecaster's credibility function as a means of "correcting-for" or "undoing" the errors and biases it purportedly describes. If this procedure can be used to systematically improve the accuracy of a forecaster's SPJs out-of-sample, then it is reasonable to conclude that the linear credibility framework captured genuine information about individual-level tendencies towards error and bias. Thus, the purpose of Analysis 1b was to shed light on the strength and practical value of the information provided by the linear credibility framework.

Taken together, these two analyses were intended to provide a preliminary, empirical basis for determining whether linear regression can provide valid and/or useful estimates of credibility, even in decision environments where it is too simple to measure errors and biases veridically.

## Method

**Design.** To simulate the judgments of 3150 hypothetical forecasters, four types of error were applied to a set of known (i.e., *a priori*) probabilities according to a  $[10 \times 5 \times 9 \times 7]$  factorial design:  $\text{noise} = \{0, 1, \dots, 9\}$ , which corresponded to the standard deviation of a normally distributed, additive, random shock applied to each judgment;  $\text{regression} = \{0.2, 0.4, 0.6, 0.8, 1\}$ , which corresponded to a multiplicative constant  $c$  describing each forecaster's tendency to shrink his or her judgments toward 0.50 in the face of incomplete information;  $\text{bias} = \{-1.0, -0.75, \dots, 1\}$ , which corresponded to a systematic, additive error in the numerical expression of a forecaster's beliefs in log-odds space; and  $\text{expertise} = \{0.25, 0.50, \dots, 1.75\}$ , which corresponded to the validity of a forecaster's judgments as a predictor of aggregate judgments in log-odds space, expressed as a multiplicative constant.

**Procedure.** In each of the 3150 simulations, I began by generating a set of optimized aggregate judgments (i.e., *estimated optima*) from a noisy “crowd” of 100

simulated forecasters. To do so, I followed the exact procedure<sup>4</sup> developed by Baron et al. in their 2014 paper:

1. Generate 100 “signals,” corresponding to a set of “optimal” probability judgments, given the information in the decision environment (for a discussion of the assumptions underlying this assertion, see: Baron et al., 2014). For the purposes of simulation, these judgments correspond to values ranging from 0.500 to 0.995 in increments of 0.005.
2. Transform these values to log-odds using the standard logit link function. These values now range from 0 to 5.29.<sup>5</sup>
3. Replicate this vector of signals 100 times, yielding a 100-by-100 matrix. Each column of this matrix is identical to the original vector of signals, and each row comprises 100 instances of a given signal value.
4. Add *noise* to each row (i.e., to each signal value). The basic noise vector in each case consists of 100 normal quantiles ranging from 0.005 to 0.995 in increments of 0.01. This vector is normally distributed, and ranges from -2.58 to 2.58. For

---

<sup>4</sup> Note, the steps for generating *estimated optima* in this section are largely paraphrased from Baron et al. (2014).

<sup>5</sup> This transformation is necessary to ensure that the credibility estimation procedure does not violate the assumption of symmetric errors. Because probabilities are strictly bounded by 0 and 1, estimating credibility in probability space would limit the size of allowable errors at the extremes (likely resulting in a violation of the assumption of symmetry). This problem can be solved by estimating credibility in log-odds space, where non-certain probability values are mapped onto the real numbers (Baron et al., 2014; Satopää et al., 2014).

each simulation, this vector is multiplied by the *noise* parameter (i.e., a constant ranging from 0 to 9, which represents the standard deviation of the noise distribution for each simulation) and added to each row of the signal matrix. The entries in this matrix now represent the noisy judgments of 100 forecasters in log-odds space, prior to any *regression* (see: Baron et al., 2014).

5. Multiply the entire matrix by a constant  $c$ , representing the amount of *regression*. For different simulations,  $c$  takes values of  $\{0.2, 0.4, 0.6, 0.8, \text{ and } 1.0\}$ ,<sup>6</sup> where a value of 1.0 corresponds to no regression.
6. Transform these judgments<sup>7</sup> back to probabilities.
7. Aggregate the judgments in each row by averaging.
8. Find the squared deviation of the mean of each row from its corresponding signal value.
9. Use the following extremization function to estimate the optimal transformation constant  $a$ , that minimizes the sum of these squared deviations (this is done to

---

<sup>6</sup> Baron et al. (2014) list the values of  $c$  as  $\{0, 0.2, 0.4, 0.6, \text{ and } 0.8\}$  (p. 138, emphasis added). However, the inclusion of zero is likely a typo, as this would yield a matrix in which all predictions are zero. Given that Baron et al. later discuss a condition in which judgments are subjected to “no regression,” (p. 139), we assume that Baron et al. intended to present the values of  $c$  as  $\{0.2, 0.4, 0.6, 0.8, \text{ and } 1\}$ . This assertion is corroborated by their R code (p. 144).

<sup>7</sup> Baron et al. (2014) lists this step as “Transform these *aggregates* back to probabilities” (p. 138, emphasis added). However, both their procedure and their R code suggest that this may be a typo, as the judgments in the 100-by-100 matrix are not aggregated until step 7.

optimally correct for judgmental regression. For a detailed discussion of regression and its origins, see Baron et al., 2014).

$$t(p) = \frac{p^a}{p^a + (1-p)^a}, \quad (1)$$

Where  $p$  is one of the aggregate probability judgments calculated in step 7, and  $t(p)$  is an optimized aggregate.

10. As an additional step not discussed in Baron et al. (2014), apply this optimal transformation to the aggregate probabilities calculated in step 7, and re-transform to log-odds to arrive at a set of *estimated optima*.

I take the estimates calculated in step 10 to represent the optimized wisdom of a noisy and heterogeneous “crowd” of 100 forecasters, given parameters for *noise* and *regression* in the decision environment. For each simulation (i.e., for each forecaster), these values were used as the near-objective criteria against which I assessed credibility.

For each forecaster, I then generated a calibration sample for his or her credibility function according to the following procedure:

11. Generate a vector of 100 “signals” and transform to log-odds, as in steps 1 and 2 above.

12. Adjust each judgment in this vector by a random additive shock, drawn from a normal distribution with mean zero and standard deviation equal to the *noise* parameter of the relevant simulation.
13. Multiply each judgment by a constant corresponding to the forecaster's *expertise*.
14. Adjust each judgment by an additive constant, corresponding to the forecaster's *bias*.<sup>8</sup>
15. And finally, multiply each judgment by a constant corresponding to the forecaster's degree of *regression*.

For each forecaster  $i$ , these calibration data were then regressed on the corresponding *estimated optima* calculated in step 10, yielding a credibility function with estimated coefficients corresponding to *bias* ( $\hat{\alpha}_i^*$ ) and *expertise* ( $\hat{\beta}_i^*$ ), and a standard error corresponding to *consistency* ( $\hat{\sigma}_i^*$ ):

$$t(p) = \hat{\alpha}_i^* + \hat{\beta}_i^* y_i^* + \varepsilon, \quad \varepsilon \sim N(0, \hat{\sigma}_i^{*2}), \quad (2)$$

Where  $i$  is a forecaster; the asterisk symbol (\*) indicates that an estimate was derived from a forecaster's calibration sample;  $y_i^*$  is a vector of log-odds transformed judgments that correspond to forecaster  $i$ 's beliefs about

<sup>8</sup> Note that the effect of *bias* was intentionally applied *after* the effect of *expertise* to preserve the independence of the two effects on a forecaster's final judgments. From a psychological perspective, this type of bias corresponds to a systematic error in response formation rather than in perception.

the events associated with the original vector of signals;  $t(p)$  corresponds to the vector of *estimated optima* calculated in step 9; and  $\hat{\alpha}_i^*$ ,  $\hat{\beta}_i^*$ , and  $\hat{\sigma}_i^*$  are the estimated parameters of participant  $i$ 's credibility function, as defined above.

After estimating credibility, I then generated 100 additional samples of judgments (i.e., 100 “prediction samples”) for each forecaster and used bootstrap estimation procedures to evaluate the typical performance of the linear credibility framework:

16. Using the process outlined in steps 11 through 15, generate one-hundred prediction samples for each forecaster  $i$ .<sup>9</sup>
17. For each forecaster, fit a new credibility function to the judgments in each prediction sample  $n$  and record the resulting credibility estimates (i.e.,  $\hat{\alpha}_{in}$ ,  $\hat{\beta}_{in}$ , and  $\hat{\sigma}_{in}$ ).
18. For each forecaster, recalibrate judgments in each prediction sample  $n$  according to the linear relationship captured by his or her *original* credibility function,

---

<sup>9</sup> Note that while each of these 100 predictions samples is not technically a “subset” of a forecaster’s SPJs, they can be thought of as a subset of the SPJs that a given set of simulation parameters can plausibly produce. As such, this step is designed to simulate the conventional bootstrapping procedure of random sampling within a set of observations (with replacement across trials).

estimated prior to step 16 (i.e., “plug” a forecaster’s vector of judgments  $y_{in}$  into Eq. 2 to estimate  $t(p)$ ).<sup>10</sup>

19. After recalibration, record the impact of step 18 by comparing the accuracy of judgments before and after recalibration (for a specific list of outcome measures, see Analysis 1b).

20. After completing steps 16-19 for all forecasters, average across the sampling distributions generated in steps 17 and 19 to calculate (a) bootstrapped credibility estimates for each forecaster (for use in Analysis 1a); and (b) bootstrapped summary-statistics representing the typical effects of recalibration for each forecaster (for use in Analysis 1b).

## Results

**Fit of *estimated optima*.** Similar to the results of Baron et al. (2014), the *estimated optima* calculated in each simulation demonstrated a remarkable correspondence to the original 100-item vector of signals generated in step 1 (prior to re-transforming these estimates to log-odds). Across all 3150 simulations, the mean sum of squared deviations between *estimated optima* and the original vector of signals was  $4.43 * 10^{-5}$  ( $Mdn. = 9.34 \times 10^{-6}$ ;  $SD = 8.96 \times 10^{-5}$ ;  $Max = 4.17 \times 10^{-4}$ ). Given this fit, I

---

<sup>10</sup> Note, only one credibility function was estimated for each forecaster, and all 100 samples were recalibrated using the same parameters,  $\hat{\alpha}_i^*$  and  $\hat{\beta}_i^*$  (see: Eq. 2). This was done to ensure that all instances of recalibration were conducted on out-of-sample judgments without having to partition each sample into a calibration sample and a hold-out sample.

concluded that these *estimated optima* would serve as a suitable standard for estimating credibility.

**Analysis 1a: validity of credibility estimates.** In most cases, *bias*, *expertise*, and *consistency* are unlikely to be independent components of credibility. Indeed, in Study 1, the additive effects of *noise* and *bias* were explicitly moderated by the multiplicative effects of *regression* and *expertise*. As a result, linear regression is unlikely to be a perfect tool for modeling credibility and should not be expected to produce credibility estimates (i.e., estimated coefficients corresponding to the credibility function's intercept, slope, and standard error) that closely approximate the *a priori* error parameters (*bias*, *expertise*, and *noise*) used in simulation.

As a practical matter, however, linear credibility estimates (i.e.,  $\hat{\alpha}_i^*$ ,  $\hat{\beta}_i^*$ , and  $\hat{\sigma}_i^*$ ) do not need to be scalar measures of “skill” or “proficiency” to be useful. Instead, because uncertain decision environments provide so little basis for defining “good” judgment, credibility estimates are useful to the extent that they allow decision makers to distinguish between “better” vs. “worse” judges of subjective probability. In general, therefore, the utility of the linear credibility framework is defined by its ability to predict “skill” or “proficiency” in subjective probability judgment, rather than its ability to measure the same. In Study 1, I evaluated this predictive validity by examining the extent to which bootstrapped credibility estimates were correlated with and/or uniquely predicted by the complementary error parameters used in simulation (i.e., a forecaster’s “true” degree of credibility). Table 1 shows simple correlations between these two sets of

variables, and Tables 2-4 show the results of exploratory linear regressions where simulation parameters were used to predict bootstrapped measures of *bias*, *expertise*, and *consistency*, respectively.

Table 1

*[Simulated data]: Simple correlations between simulation parameters and bootstrapped estimates of credibility.*

Credibility Estimate (Measured)	Simulation Parameter (Manipulated)			
	Noise	Regression	Bias	Expertise
Bootstrapped Alpha: $\hat{\alpha}_i$ (Bias)	0.54***	0.00	-0.33***	0.00
Bootstrapped Beta: $\hat{\beta}_i$ (Expertise)	-0.12***	-0.42***	0.00	-0.49***
Bootstrapped Sigma: $\hat{\sigma}_i$ (Consistency)	-0.33***	0.02	0.00	0.00

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: because credibility estimates are intended to describe the linear transformations required to “undo” the errors applied to judgments in simulation, negative correlations are to be expected between credibility estimates and complementary simulation parameters.

Table 2

*[Simulated data]: Predictors of bootstrapped alpha (i.e., bias;  $\hat{\alpha}_i$ ).*

Simulation Parameters (Manipulated)	Est. Coefficient	SE	t-value	p-value
(Intercept)	0.55	0.03	18.82	<.001***
Noise	0.11	0.00	38.96	<.001***
Regression	0.00	0.03	0.11	0.92

Bias	-0.32	0.01	-24.21	<.001***
Expertise	0.00	0.02	0.00	1.00

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: because credibility estimates are intended to describe the linear transformations required to “undo” the errors applied to judgments in simulation, negative relationships are to be expected between credibility estimates and complementary simulation parameters.

Table 3

*[Simulated data]: Predictors of bootstrapped beta (i.e., expertise;  $\hat{\beta}_i$ ).*

Simulation Parameters (Manipulated)	Est. Coefficient	SE	t-value	p-value
(Intercept)	0.26	0.00	60.23	<.001***
Noise	0.00	0.00	-9.00	<.001***
Regression	-0.14	0.00	-31.24	<.001***
Bias	0.00	0.00	-0.08	0.94
Expertise	-0.09	0.00	-36.07	<.001***

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: because credibility estimates are intended to describe the linear transformations required to “undo” the errors applied to judgments in simulation, negative relationships are to be expected between credibility estimates and complementary simulation parameters.

Table 4

*[Simulated data]: Predictors of bootstrapped sigma (i.e., consistency;  $\hat{\sigma}_i$ ).*

Simulation Parameters (Manipulated)	Est. Coefficient	SE	t-value	p-value
(Intercept)	$3.24 \times 10^{-2}$	0.00	39.14	<.001***
Noise	$-1.64 \times 10^{-3}$	0.00	-19.68	<.001***

Regression	$1.07 \times 10^{-3}$	0.00	1.26	0.21
Bias	$8.88 \times 10^{-6}$	0.00	0.02	0.98
Expertise	$-7.39 \times 10^{-5}$	0.00	-0.16	0.89

Significance levels: \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$

Note: because credibility estimates are intended to describe the linear transformations required to “undo” the errors applied to judgments in simulation, negative relationships are to be expected between credibility estimates and complementary simulation parameters.

**Discussion.** The results of Analysis 1a demonstrate that a simple, main-effects-only approach to linear regression (i.e., the linear credibility framework) can provide decision makers with useful information about who is likely to be a “better” vs. “worse” judge of subjective probability. Specifically, Tables 1-4 indicate that the linear credibility framework can produce estimates that are predictive of “skill” or “proficiency” in subjective probability judgment, and that all covary with the expected simulation parameters. In practice, the strengths of these relationships were insufficient to conclude that credibility estimates should be interpreted as face-value measures of credibility. However, because Study 1’s decision environment was considerably more complex than the method used to model it (i.e., Analysis 1a’s simulated decision environment imposed two- and three-way interactions between *a priori* error parameters, whereas credibility was estimated with a main-effects-only model), the existence of these relationships suggests a robust predictive validity between linear credibility estimates and the constructs they are intended to represent.

When examined in terms of simple correlations (Table 1), these predictive relationships can be seen in (a) the unique covariation between *a priori bias* and estimated *bias* ( $\hat{\alpha}_i$ ); (b) the unique covariation between *a priori expertise* and estimated *expertise* ( $\hat{\beta}_i$ ); and (c) the moderate, appropriately-signed relationship between estimated *consistency* ( $\hat{\sigma}_i$ ) and *a priori noise*. In the corresponding regression analyses (Tables 2-4), similar relationships can be observed in the separable effects of *a priori bias* and *a priori expertise* in the prediction of estimated *bias* ( $\hat{\alpha}_i$ ) and estimated *expertise* ( $\hat{\beta}_i$ ), respectively (Tables 2 and 3); and the fact that estimated *consistency* ( $\hat{\sigma}_i$ ) is uniquely predicted by *a priori noise* (Table 4).

Taken together, these effects suggest that linear regression can provide useful information about an individual's relative degree of *bias*, *expertise*, and *consistency* in subjective probability judgment. Strictly speaking, of course, the strength and separability of these effects was diminished by the fact that the linear credibility framework was too simple to capture the interactive features of Study 1's decision environment. However, the presence of systematic, sensible relationships between linear credibility estimates and the complementary simulation parameters (i.e.,  $\hat{\alpha}_i$  and *a priori bias*;  $\hat{\beta}_i$  and *a priori expertise*;  $\hat{\sigma}_i$  and *a priori noise*, respectively) suggests that the linear credibility framework is likely to be a robust—albeit oversimplified—tool for examining the validity of SPJs.

**Analysis 1b: empirical effects of recalibration.** To assess the practical value of the linear credibility framework, I examined the impact of credibility-based recalibration,

both within- and between-subjects. To do this, I recorded a variety of outcome measures after each instance of recalibration (step 19, above), most of which pertained to the absolute difference between a forecaster’s judgments and the corresponding *estimated optima*. For simplicity, I will call this difference *absolute judgment error*, or AJE. In each prediction sample (for each forecaster), I recorded the following summaries of AJE:

- The proportion of individual judgments for which recalibration improved (reduced) AJE;
- The mean pairwise difference in AJE due to recalibration;<sup>11</sup>
- The effect-size (Cohen’s  $d$ ) associated with pairwise changes in AJE due to recalibration;
- And a binary indicator of whether recalibration improved (reduced) the sample’s mean AJE.<sup>12</sup>

At the conclusion of all 3150 simulations, these values were averaged across each forecaster’s 100 prediction samples, yielding four summary-statistics per forecaster. The four statistics were as follows, each of which represents a different aspect of the *typical*<sup>13</sup> or expected effect of recalibration on a given forecaster’s judgments:

---

<sup>11</sup> Note that this measure (i.e., the mean difference) is mathematically equivalent to the difference in mean AJE, due to recalibration. Thus, I will only discuss the former and not the latter when presenting results.

<sup>12</sup> In all cases, binary indicators were coded as “1” if the stated event occurred, and “0” if it did not.

<sup>13</sup> I will use the word “typical” in this chapter to indicate bootstrapped averages, calculated over each forecaster’s 100 prediction samples. This is done to prevent confusion in instances where “typical” values are bootstrapped averages of sample-level means.

- The typical proportion of forecaster  $i$ 's judgments for which recalibration improved (reduced) AJE.
- The typical pairwise change in forecaster  $i$ 's AJE due to recalibration;
- The typical effect-size (Cohen's  $d$ ) of recalibration on forecaster  $i$ 's AJE;
- And proportion of prediction samples in which recalibration improved (reduced) forecaster  $i$ 's mean AJE.

These summary-statistics were then tabulated across all 3150 forecasters and served as the primary dependent variables (DVs) for the tests that follow. Figures 1-4 show the empirical distributions of these DVs across forecasters, and Figure 5 shows a visual comparison of mean AJEs before and after recalibration. Table 5 provides descriptive statistics for each of the distributions represented in Figures 1-4, and Table 6 shows the results of Wilcoxon signed-rank tests examining the null hypothesis that the median of each distribution does not differ from chance (represented by the red dotted lines in Figures 1-4).

Figure 1

*[Simulated data]: Typical proportion of judgments for which recalibration improved (reduced) AJE.*

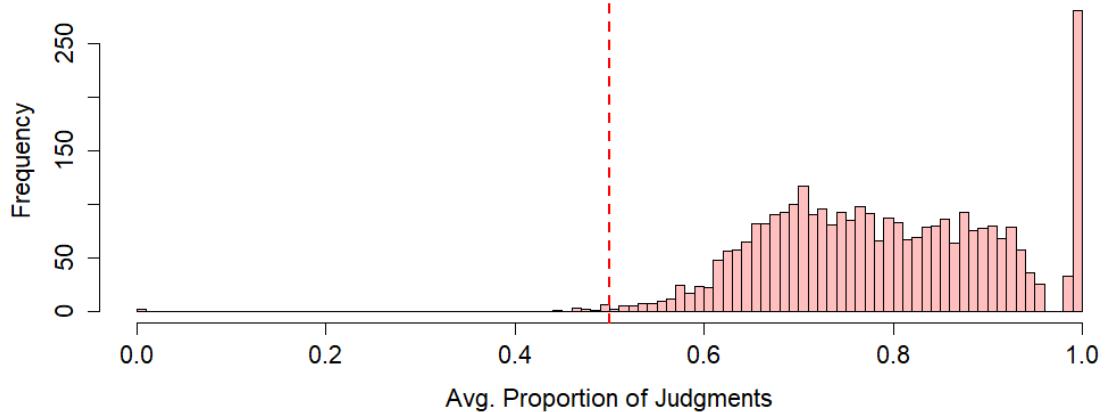
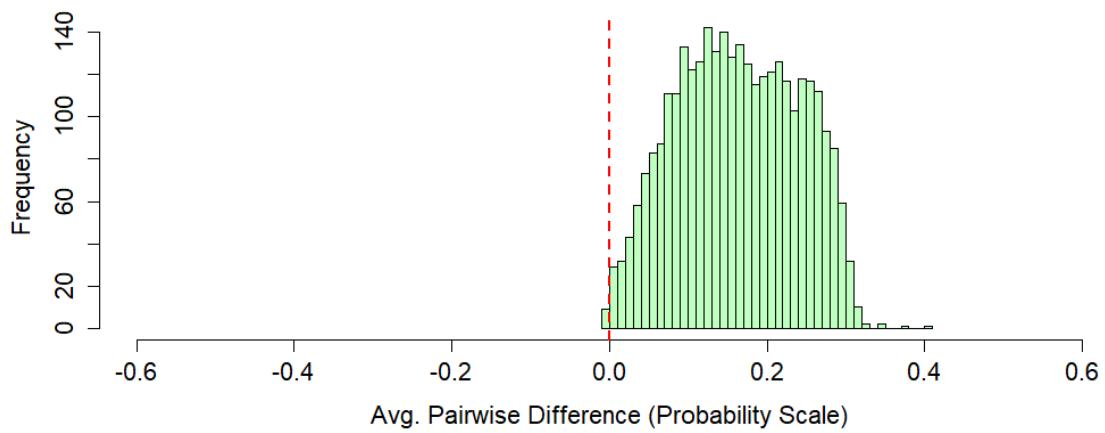


Figure 2

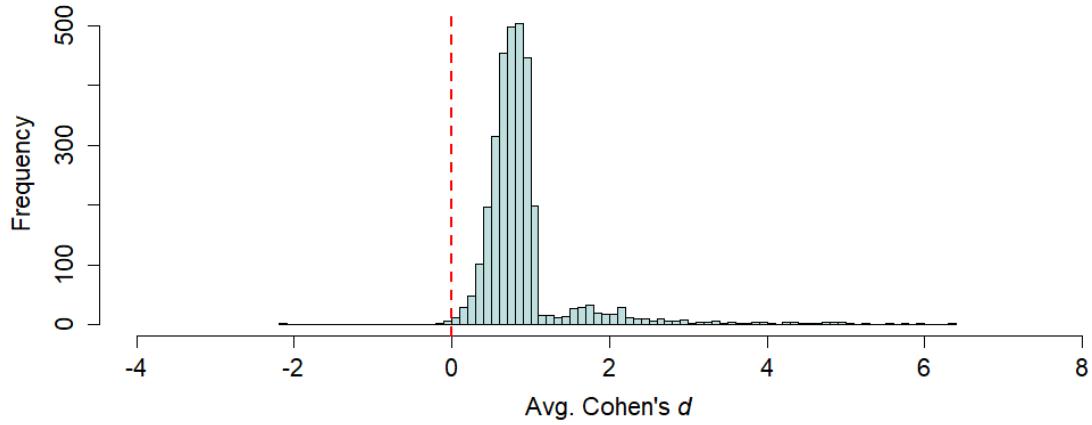
*[Simulated data]: Typical pairwise change in AJE (pre – post), due to recalibration.*



Note: positive values indicate an improvement (reduction) in AJE.

Figure 3

*[Simulated data]: Typical effect-size (Cohen's d) of recalibration on AJE.*



*Note:* positive values indicate an improvement (reduction) in AJE.

Figure 4

[*Simulated data*]: Proportion of samples in which recalibration improved (reduced) mean AJE.

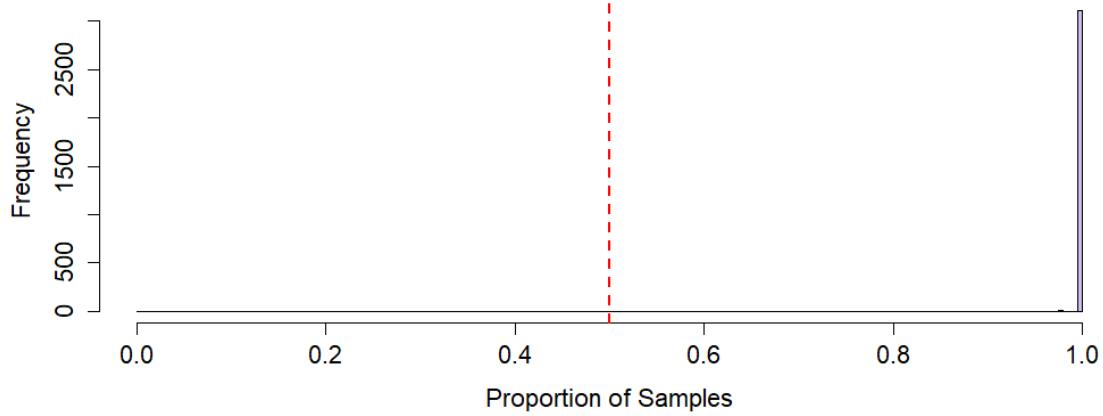
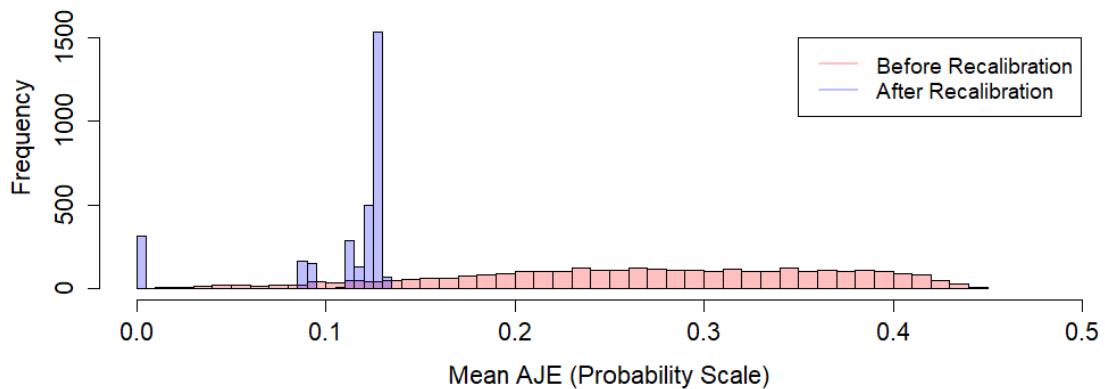


Figure 5

[*Simulated data*]: Mean AJE, before and after recalibration.



*Note:* smaller values indicate more accurate judgements (smaller errors), on average.

Table 5

*[Simulated data]: Typical effects of credibility-based recalibration, summarized across forecasters.*

Outcome Measure	Between-Subjects Summary Statistics				
	Mean	Mdn.	SD	Min.	Max.
Typical proportion of judgments for which recalibration improved AJE.	79%	78%	12%	0%	100%
Typical pairwise change in AJE (pre - post), due to recalibration.	16.25 $\times 10^{-2}$	16.13 $\times 10^{-2}$	7.71 $\times 10^{-2}$	-0.32 $\times 10^{-2}$	40.50 $\times 10^{-2}$
Typical effect (Cohen's $d$ ) of recalibration on AJE.	0.90	0.78	0.63	-2.19	6.37
Proportion of samples in which recalibration improved mean AJE.	100%	100%	5%	0%	100%

Table 6

*[Simulated data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved (reduced) AJE beyond chance.*

Outcome Measure	$H_0$	$Prop. Mass > H_0$	Stat. ( $V$ )	p-value
Typical proportion of judgments for which recalibration improved AJE.	Mdn. = 0.5	100%	$4.96 \times 10^6$	<.001***
Typical pairwise change in AJE (pre - post), due to recalibration.	Mdn. = 0	100%	$4.96 \times 10^6$	<.001***
Typical effect (Cohen's $d$ ) of recalibration on AJE.	Mdn. = 0	100%	$4.96 \times 10^6$	<.001***
Proportion of samples in which recalibration improved mean AJE.	Mdn. = 0.5	100%	$4.96 \times 10^6$	<.001***

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

**Discussion.** The results of Analysis 1b demonstrate that credibility-based recalibration can be used to increase the accuracy of SPJs across a wide variety of decision environments. Across 3150 forecasters, the typical effect of credibility-based recalibration was to significantly improve (reduce) AJE, regardless of whether this outcome was defined in terms of (a) the typical proportion of judgments for which AJE improved; (b) the typical pairwise change in AJE; (c) the typical effect-size (Cohen's  $d$ ) associated with pairwise changes in AJE; or (d) the proportion of samples in which mean AJE improved. Indeed, as can be seen in Figures 1-4, there were very few cases in which the effects of recalibration were negative, with less than 0.5% of observations falling below chance in each distribution (see also: Table 6).

In addition to being widely beneficial, the results of Analysis 1b also indicate that the expected effect-size of credibility-based recalibration is likely to be substantial. When considered on a study-wide level, Analysis 1b suggests that decision makers who employ

credibility-based recalibration under Study-1-like conditions can expect an average of 79% of an individual's judgments to improve; for the AJE of each judgment to improve by an average of 16.25 points on the probability scale; for mean AJE to improve in 100% of samples; and for AJE to improve by an average of 0.9 standard deviations, overall (for additional descriptive statistics, see: Table 5). In practice, of course, the generalizability of these results will depend on the degree to which the decision environments simulated in Study 1 are representative of the real world. However, because the positive effects of credibility-based recalibration were observed across a broad range of simulated environments, Analysis 1b suggests that the linear credibility framework may serve as a widely applicable tool for improving SPJs.

## General Discussion

In Study 1, I employed a simulation procedure that purposefully failed to model *noise*, *regression*, *bias*, and *expertise* as independent effects on judgment. As a result (and as expected), the linear credibility framework produced suboptimal results in two ways: (a) the bootstrapped credibility estimates examined in Analysis 1a did not uniquely covary their complementary simulation parameters; and (b) the strength of the covariation between each pair of variables ( $\hat{\alpha}_i$  and *a priori bias*;  $\hat{\beta}_i$  and *a priori expertise*;  $\hat{\sigma}_i$  and *a priori noise*, respectively) did not suggest that linear credibility estimates were pure, face-valid measures of *bias*, *expertise*, or *consistency*. Nevertheless, the results of Study 1 demonstrate that linear regression can provide both valid and useful estimates of credibility.

In Analysis 1a, the predictive validity of linear credibility estimates was established by observing that estimated *bias* ( $\hat{\alpha}_i$ ), estimated *expertise* ( $\hat{\beta}_i$ ), and estimated *consistency* ( $\hat{\sigma}_i$ ) systematically covaried with the complementary error parameters used in simulation. Though the strength of the covariation between each pair was only moderate, the presence (and relative independence) of these relationships demonstrate the robustness of the linear credibility framework as a tool for examining credibility. Despite an intentionally too-complex decision environment, the results of Analysis 1a indicate that the bare-bones assumptions of the linear credibility framework were sufficient to identify valid, first-order approximations of credibility in simulated data. As a result, it stands to reason that linear credibility estimates might help decision makers evaluate the validity of SPJs (or sources of SPJs) across a wide variety of ecological conditions.

In Analysis 1b, the utility of the linear credibility framework was demonstrated by using credibility information to recalibrate SPJs. Across 3150 simulated forecasters, credibility-based recalibration significantly reduced absolute judgment error (AJE) in nearly all cases. This effect was evident across four outcome measures, each of which represented a different approach to summarizing AJE. Indeed, regardless of whether one chose to look at the average proportion of judgments improved, the average pairwise improvement in judgments, or the average improvement in mean AJE, the results of Analysis 1b were unanimous in demonstrating that credibility-based recalibration can be used to improve SPJs. In addition, Analysis 1b demonstrated that the expected effects of credibility-based recalibration can be substantial. Though it remains to be seen whether the effect-sizes observed in Study 1 are representative of the real world, the sheer breadth

of decision environments examined in Study 1 provides grounds for optimism that credibility-based recalibration might serve as a general tool for improving SPJs.

## Conclusions

The results of Study 1 demonstrate that a main-effects-only approach to linear regression can be used to model credibility in simulated data. It is evident from this study that the linear credibility framework is unlikely to provide face-valid measures of *bias*, *expertise*, and *consistency* in complex, ecologically-representative decision environments. However, it is also clear that the linear credibility framework can provide decision makers with: (a) general information about individual-level errors and biases in subjective probability judgment; and (b) concrete indicators of who is likely to be a “better” vs. “worse” judge of subjective probability. As a proof-of-concept, therefore, the results of Study 1 suggest that it is plausible for simple statistical models to provide decision makers with insight about a judge’s “quality” (i.e., relative validity) as a source of probabilistic information— even in the absence of an objective standard. Thus, practical applications of the linear credibility framework should be explored.

## References

- Armstrong, J. S. (2001). Combining forecasts. *Principles of forecasting: a handbook for researchers and practitioners*. J. S. Armstrong (Ed.). Norwell, MA: Kluwer Academic Publishers.
- Armstrong, J. S. (Ed.). (2001). *Principles of forecasting: a handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic Publishers.
- Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3), 691-706.
- Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2), 133-145.
- Basili, M., & Pratelli, L. (2015). Aggregation of not independent experts' opinions under ambiguity. *Structural Safety*, 52, 144-149.
- Berg, J., Forsythe, R., Nelson, F., & Rietz, T. (2001). Results from a dozen years of election futures markets research. C. Plott & V. Smith (Eds.). *Handbook of experimental economic results*. Amsterdam: North Holland.
- Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology*, 37(1), 48-74.
- Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, 36(8), 887-899.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193.
- Budescu, D. V., & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2), 267-280.
- Budescu, D. V., & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, 104(3), 371-398.

- Budescu, D. V., & Yu, H. T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, 20(2), 153-177.
- Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogiannakis, N. (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Systems with Applications*, 112, 353-371.
- Chen, K. Y., Fine, L. R., & Huberman, B. A. (2004). Eliminating public knowledge biases in information-aggregation mechanisms. *Management Science*, 50(7), 983-994.
- Choo, M. S., Uhm, S., Kim, J. K., Han, J. H., Kim, D. H., Kim, J., & Lee, S. H. (2018). A prediction model using machine learning algorithm for assessing stone-free status after single session shock wave lithotripsy to treat ureteral stones. *The Journal of Urology*, 200(6), 1371-1377.
- Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4), 559-583.
- Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2), 187-203.
- Cowgill, B., & Zitzewitz, E. (2015). Corporate prediction markets: Evidence from Google, Ford, and Firm X. *The Review of Economic Studies*, 82(4), 1309-1341.
- Cross, D., Ramos, J., Mellers, B., Tetlock, P. E., & Scott, D. W. (2018). Robust forecast aggregation: Fourier L2E regression. *Journal of Forecasting*, 37(3), 259-268.
- Da Silva, B. N., & Garcia, A. C. B. (2006). A hybrid method for image taxonomy: Using captcha for collaborative knowledge acquisition. In *Proceedings of the AAAI 2006 Fall Symposium on Semantic Web for Collaborative Knowledge Acquisition*, 17-23.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571-582.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899), 1668-1674.
- Ebbesen, E. B., & Konecni, V. J. (1975). Decision making and information integration in the courts: The setting of bail. *Journal of Personality and Social Psychology*, 32(5), 805-821.

- Edwards, W., & Fasolo, B. (2001). Decision technology. *Annual Review of Psychology*, 52(1), 581-606.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3), 193-242.
- Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4), 643-669.
- Flowerdew, J. (2014). Calibrating ensemble reliability whilst preserving spatial structure. *Tellus A: Dynamic Meteorology and Oceanography*, 66(1), 1-20.
- Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology*, 64(3), 479-490.
- Genest, C., & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1), 114-135.
- Gerds, T. A., Cai, T., & Schumacher, M. (2008). The performance of risk prediction models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(4), 457-479.
- Gneiting, T., & Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7, 1747-1782.
- Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendal Jr., K. C. (2017). Ensembles of overconfident and overfit forecasts. *Management Science*, 63(4), 1110-1130.
- Gürkaynak, R., & Wolfers, J. (2006). *Macroeconomic derivatives: An initial analysis of market-based macro forecasts, uncertainty, and risk*. Cambridge, MA: National Bureau of Economic Research.
- Hayek, F. (1945). The use of knowledge in society. *American Economic Review*, 35(4), 519-530.
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 382-401.
- Hora, S. C., & Kardeş, E. (2015). Calibration, sharpness and the weighting of experts in a linear opinion pool. *Annals of Operations Research*, 229(1), 429-450.
- Hrust, L., Klaić, Z. B., Križan, J., Antonić, O., & Hercog, P. (2009). Neural network

- forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. *Atmospheric Environment*, 43(35), 5588-5596.
- Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277-1288.
- Jacobs, R. A. (1995). Methods for combining experts' probability assessments. *Neural Computation*, 7(5), 867-888.
- Karvetski, C. W., Olson, K. C., Mandel, D. R., & Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis*, 10(4), 305-326.
- Kruppa, J., Ziegler, A., & König, I. R. (2012). Risk estimation and risk prediction using machine-learning methods. *Human Genetics*, 131(10), 1639-1654.
- Laming, D. (2010). Statistical information and uncertainty: A critique of applications in experimental psychology. *Entropy*, 12(4), 720-771.
- Massonet, F., Bellprat, O., Guemas, V., & Doblas-Reyes, F. J. (2016). Using climate models to estimate the quality of global observational data sets. *Science*, 354(6311), 452-456.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.
- Nagar, Y. (2013). *Combining human and machine intelligence for making predictions*. Doctoral dissertation, Massachusetts Institute of Technology.
- Olsen, R. A. (1997). Desirability bias among professional investment managers: Some evidence from experts. *Journal of Behavioral Decision Making*, 10(1), 65-72.
- Pennock, D. M., Lawrence, S., Nielsen, F. Å., & Giles, C. L. (2001, August). Extracting collective probabilistic forecasts from web games. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*.
- Peterson, D. K., & Pitz, G. F. (1986). Effect of input from a mechanical model on clinical judgment. *Journal of Applied Psychology*, 71(1), 163-167.
- Plott, C. R., & Chen, K. Y. (2002). *Information aggregation mechanisms: Concept*,

- design and implementation for a sales forecasting problem.* Pasadena, CA: California Institute of Technology.
- Rantilla, A. K., & Budescu, D. V. (1999). Aggregation of expert opinions. In *Proceedings of the 32<sup>nd</sup> Annual Hawai'i International Conference on Systems Sciences*.
- Rothschild, D. (2009). Forecasting elections: Comparing prediction markets, polls, and their biases. *Public Opinion Quarterly*, 73(5), 895-916.
- Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., & Pradhan, B. (2018). A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Science of the Total Environment*, 644, 954-962.
- Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2), 344-356.
- Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, 66(3), 178-200.
- Spann, M., & Skiera, B. (2003). Internet-based virtual stock markets for business forecasting. *Management Science*, 49(10), 1310-1326.
- Tetlock, P. E. (2005). *Expert political judgment. How good is it? How can we know?* Princeton, NJ: Princeton University Press.
- Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, 95(3), 261-289.
- Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, 10(3), 243-268.
- Wang, B., Ding, Q., Fu, X., Kang, I. S., Jin, K., Shukla, J., & Doblas-Reyes, F. (2005). Fundamental challenge in simulation and prediction of summer monsoon rainfall. *Geophysical Research Letters*, 32(15), 1-4.

## INTERLUDE

### **CREDIBILITY ESTIMATION WITH EMPIRICAL DATA: A METHODOLOGICAL OVERVIEW OF STUDIES 2A-2C AND 3A-3C**

For the remainder of this dissertation, I will discuss the applicability of the linear credibility framework to three sets of real-world data. As in Study 1, the purpose of this research will be to (a) determine whether simple, linear models of credibility can be used to identify meaningful indicators of “skill” or “proficiency” in subjective probability judgment; and (b) to test whether this information can be used to improve the accuracy of subjective probability judgments (SJPs) out-of-sample. Because these aims represent two sides of the same coin — i.e., because both depend on the extent to which empirical models of credibility can capture information about errors and biases in human judgment — both can be probed with similar analytic procedures. Much like Chapter 1, therefore, the research in Chapters 2 and 3 will involve a single, general procedure but will be presented as a series of discrete analyses for the sake of clarity.

To help the reader navigate these chapters, this interlude will be dedicated to a methodological overview of the studies presented in Chapters 2 and 3 — namely, Studies 2a-2c and 3a-3c. In the sections that follow, I will begin with a brief discussion of the structure and organization of the research presented in Chapters 2 and 3. After this, I will describe the general method used for Studies 2a-2c and 3a-3c, including detailed descriptions of the data-sets I analyzed and the protocols used to collect them. Finally, I will provide a detailed description of the General Procedure used for Studies 2a-2c and

will conclude with a brief, conceptual discussion of the ways in which these studies differ from Study 1. After discussing these topics, I will proceed to Chapter 2, where I examine the reliability and validity of linear credibility estimates derived from empirical data.

### **On the Structure and Organization of Chapters 2 and 3**

In Chapters 2 and 3, research findings will be organized according to three main criteria: the broad research question to which they pertain (a criterion that also corresponds to the chapter in which they appear); the data-set from which they were drawn; and the narrow, empirical question that they address. To help readers orient themselves to these criteria, section headings in Chapters 2 and 3 will include a three-level naming convention that encodes all three types of information:

1. At the first level, broad research questions will be indicated by Arabic numerals.

Because broad research questions are separated into chapters, this level can also be understood as a signifier of chapter, with all results and discussions related to Chapter 2 (reliability and validity of credibility estimates) being indicated by the number “2” and all results and discussions related to Chapter 3 (the typical effects of credibility-based recalibration) being indicated by the number “3.”

2. At the second level, data-sets will be indicated by lowercase English letters. With the help of this information, research within chapters will be organized into *studies* (i.e., treatments that use specific data-sets to address a broad research

question), each of which will be designated by a two-character combination of one Arabic numeral and one English letter (e.g., “2b”). In Chapters 2 and 3, studies will be conducted across three empirical data-sets (i.e., data-sets “a,” “b,” and “c”), each of which I describe in the “Source” sections of the General Method, below.

3. At the third level, empirical research questions will be indicated by lowercase Roman numerals. With the help of this information, studies in Chapters 2 and 3 will be subdivided into *analyses* (i.e., discrete empirical investigations that leverage a specific data-set to address a broad research question), each of which will be designated by a three-component combination of an Arabic numeral, an English letter, and a Roman numeral (e.g., “2b.iii”). Because the numerical designation of each analysis is arbitrary, the title of each analysis will also contain (a) a plain English description of the empirical question at hand; and (b) a brief, descriptive parenthetical. In all cases where an analysis (or study) is referenced out of context, this parenthetical will be included to remind the reader of the treatment’s substantive content.

Taken together, the three components of this naming convention (and the accompanying parenthetical) can be used to remind the reader which study or analysis is currently being discussed. Consider, for example, the designation “2c.i.” Based on this heading, a reader can infer that this section pertains to the first analysis (“i”) of the

reliability and/or validity (“2”) of credibility estimates derived from data-set “c.” Or, in simpler terms, this section refers to the first analysis of Study 2c, which — as the title of the analysis indicates — is concerned with a narrow, empirical question: *Under what conditions are credibility estimates reliable? (PHL reliability)*. For the reader’s convenience, this title also includes the parenthetical designation “(PHL reliability),” which indicates that the analysis concerns the reliability of the “PHL” data-set — a descriptive shorthand that links the arbitrary designation “c” to its substantive content (here, judgments from the Philadelphia air temperature study; for more information on data-sets, see below).

### **General Method for Studies 2a-2c and 3a-3c**

**Data.** Data for Studies 2a-2c and 3a-3c were drawn from three sources: (a) the Good Judgment Project, which gathered over a million SPJs during the course of four year-long geopolitical forecasting tournaments; (b) an online forecasting tournament that asked participants to make predictions about the outcomes of games in the 2017 NCAA Division-I Men’s Basketball Championship (i.e., the 2017 “March Madness” tournament); and (c) a brief, online survey that asked participants to express their beliefs about the likelihood of various air temperatures in Philadelphia during the months of January and July.

#### ***Source A: The Good Judgment Project.***

*Overview.* The Good Judgment Project (GJP) consisted of four year-long forecasting tournaments sponsored by the Intelligence Advanced Research Projects Activity (i.e., IARPA; the research and development wing of the U.S. Intelligence Community). In each tournament, amateur forecasters were recruited from sources such as professional societies, university alumni associations, research centers, science blogs, popular media sources (e.g., magazines and news websites), and by word of mouth. To participate, forecasters were required to have a bachelor's degree and to complete a series of psychological and political knowledge tests. After fulfilling these requirements, participants were randomly assigned to experimental conditions that varied by *teaming* (individuals vs. collaborative groups), *training* (training in probabilistic reasoning vs. no training), *forecast elicitation method* (survey vs. prediction market), and several other factors (for a complete summary of the Good Judgment Project's protocol, see: Mellers, Stone, Atanasov, Rohrbaugh, Metz, Ungar, Bishop, Horowitz, Merkle, & Tetlock, 2015; Mellers, Ungar, Baron, Ramos, Gürçay, Fincher, Scott, Moore, Atanasov, Swift, Murray, Stone, & Tetlock, 2014).

After being assigned to conditions, forecasters were given access to an online portal where they could (a) view and respond to geopolitical forecasting questions; (b) interact with teammates (if applicable); and (c) view a variety of leaderboards displaying the most accurate forecasters (or teams) in each condition. An example question from the fourth year of the tournament was “Will Syria’s president Bashar al-Assad vacate office before 10 June, 2015?” In the survey conditions, forecasters responded to this sort of question with numerical probabilities, where a response of 1.0 indicated certainty in the

answer “yes.” While questions remained open, participants were encouraged to update their predictions as often as they desired. When a question resolved (i.e., when an outcome was observed), the accuracy of a forecaster’s judgment was assessed using the Brier score — a strictly proper scoring rule that incentivizes forecasters to provide their true beliefs (Brier, 1950). For additional details about the design, procedures, and findings of the Good Judgment Project, see: Mellers, Stone, Murray, Minster, Rohrbaugh, Bishop, Chen, Baker, Hou, Horowitz, Ungar, & Tetlock (2015b); Mellers et al., 2015a; Mellers et al., 2014. For related research, see: Cross, Ramos, Mellers, Tetlock & Scott (2018); Friedman, Baker, Mellers, Tetlock, & Zeckhauser (2018); Mellers, Baker, Chen, Mandel, & Tetlock (2017); Atanasov, Rescober, Stone, Swift, Servan-Schreiber, Tetlock, Ungar, & Mellers (2017); Baron, Mellers, Tetlock, Stone, & Ungar (2014); and Satopää, Baron, Foster, Mellers, Tetlock, & Ungar, (2014).

*Sample.* In Studies 2a-2c and 3a-3c, I used a sample of judgments drawn from all four years of the Good Judgment Project. This sample was nearly identical to that used by Friedman et al. (2018) and includes all forecasts from individuals who addressed at least 25 forecasting questions in a given year.<sup>14</sup> All told, this sample consisted of 444,164 numerical forecasts drawn from 1,832 individuals over four years of GJP tournaments and includes forecasts for only those questions with binary outcomes, comprising 380

---

<sup>14</sup> The one difference between this sample and the sample used by Friedman et al. (2018) is that Friedman et al.’s sample included separate SPJs for the probability that an event will occur and its complement. The present sample excludes these complements, thereby halving the number of observations.

questions. Within this sample, 53% of forecasters were trained in probabilistic reasoning, 44% were assigned to work in collaborative teams, and 4% were *superforecasters* — a title given to the top 2% of forecasters from each tournament year (as measured by average Brier score; for a detailed summary of superforecasters' remarkable performance in the GJP, see: Mellers et al., 2017; Tetlock & Gardner, 2016; Mellers et al., 2015b).

The average age of forecasters in this sample was 39.2 ( $Mdn. = 34.0$ ;  $SD = 13.4$ ), 85% of forecasters were male, and 57% of forecasters had some level of advanced degree. For clarity, all references to these data (i.e., figures, tables, and references to studies/analyses) will be accompanied by the designation *GJP*.

***Source B: March Madness data.***

*Sample.* Beginning in late February 2017, participants were recruited to participate in an online study in which they could earn up to \$85 for making predictions about the outcomes of basketball games in the 2017 NCAA Division-I Men's Basketball Championship (i.e., the "March Madness" tournament). Students at the University of Pennsylvania were alerted to this study through a series of flyers, some of which were posted in public places around campus, and other of which were delivered directly to student organizations (e.g., fraternities, sororities, sports teams, clubs). Based on the large number of non-student responses, it is evident that interest in this study soon spread to other members of the community (e.g., students' friends, coworkers, and family members) via word of mouth. As a result, participants in this study were largely — but

not exclusively — students from the University of Pennsylvania and individuals from the surrounding Philadelphia area.

Among the 140 participants who provided demographic information in this study (total number of participants = 143), the mean age was 24.3 ( $Mdn. = 23.0$ ;  $SD = 5.6$ ) and 44% of participants self-reported as having completed “some undergraduate” education (0% reported no college education; 24% reported an associate’s or bachelor’s degree; 13% reported “some post-baccalaureate” education; and 19% reported an advanced degree). In addition, 49% of participants self-identified as female (51% male; 0% other/neither) and 50% self-identified as “white or Caucasian” (20% black or African American; 12% East Asian; 7% Hispanic or Latinx; 5% South Asian; 5% Other; 1% Middle Eastern; 0% Native American; 0% Pacific Islander; 0% Other Indigenous People). All participants who completed this study were paid a baseline of \$20 and were given the opportunity to win an additional \$15 for correctly predicting the outcome of a randomly selected game. To incentivize accuracy, the three participants with the best Brier scores (i.e., the most accurate forecasts) at the end of the study also received performance bonuses of \$50, \$25, and \$10 respectively. For clarity, all references to these data (i.e., figures, tables, and references to studies/analyses) will be accompanied by the designation *MM*.

*Materials.* From March 4<sup>th</sup> to April 3<sup>rd</sup>, 2017, participants completed 10 online surveys. The first survey — which was sent to participants immediately after registering — included additional information about the study, informed consent, and a short section

asking for details related to online payment (first name, last name, and email address). The second survey was sent to participants several days prior to the first games of the 2017 March Madness tournament (i.e., the “play-in” games). This survey included (a) reminders about the guidelines for study participation; (b) a brief tutorial on how to provide subjective probability judgments (for a printout of this tutorial, see Appendix B); and (c) a list of additional resources, including: links to information about the structure and rules of the NCAA tournament, news outlets with the latest results and projections for the tournament, and a printable version of the tournament bracket. After reading through these materials, participants were asked to provide predictions about the four “play-in” games of the NCAA tournament.

When making their predictions, participants were presented with basic information about each of the two teams that would be playing (name, tournament seeding, date of game, regular season conference, Division-I win/loss record, and strength of schedule according to espn.com). With this information displayed, participants were asked to provide a prediction about which of the two teams would win and a numerical SPJ describing their confidence in the predicted outcome. For each game, participants also provided (a) the minimum price at which they would sell a lottery ticket that would pay \$15 if their preferred team won (and \$0 otherwise); and (b) their choice between keeping the ticket or selling it at this price, with the understanding that one of their choices (selected at random) would be paid out for real money.<sup>15</sup> This same

---

<sup>15</sup> Note: the present research will only examine participants’ binary predictions and SPJs. It does not address the latter two types of responses.

basic format was used for the remaining eight surveys. At the beginning of each survey, participants were given the opportunity to refresh their memory about any (or all) of the resources included in the “play-in game” survey. Then, participants were asked to make predictions about upcoming games.

Because later-round matchups were determined in real time, each of the remaining eight surveys was sent to participants with as much lead time as possible (i.e., as soon as the match-ups were determined) and asked for predictions about all determined match-ups that had not yet been played. In the third overall survey (which included predictions about the 32 first-round games), these predictions were followed by a battery of cognitive tests similar to those used in the GJP (for additional details, see the Detailed Procedure section of Analysis 2a.ii (GJP validity)). In all other cases, surveys consisted entirely of participant predictions. Over the course of the study, participants were asked to provide predictions for all 67 games in the 2017 March Madness tournament. On average, participants provided 59 of these predictions ( $Mdn. = 67$ ;  $SD = 16.4$ ). Full printouts of each of the ten surveys can be found in Appendix B.

***Source C: Philadelphia air temperature data.***

*Sample.* Participants were recruited to participate in a third study through an online panel curated by Jonathan Baron (the principal advisor on this thesis). To expedite the data-gathering process, this study did not include an extended battery of cognitive tests or demographic items and was designed to be completed in about 20 minutes. Participants were alerted to the availability of the study through a private email list and

were offered \$4 for participation. In response to this email, 76 people completed the study, of whom 73 provided usable data.<sup>16</sup> Among the 76 participants who completed the study (the SPJs from the three individuals who provided “unusable” data were still included in the calculation of *estimated optima*; see below), the mean age was 49.0 (*Mdn.* = 49.5; *SD* = 12.3) and 36% self-identified as male. For clarity, all references to these data (i.e., figures, tables, and references to studies/analyses) will be accompanied by the designation *PHL*.

*Materials.* After agreeing to participate in the study, participants were redirected to an online survey consisting of three sections. In the first section, participants were provided with a brief overview of the study and instructions on how to provide SPJs. On the same page, participants were also asked to provide their sex, age, and email address. In the second section of the study (beginning on the next page), participants were presented with 40 sets of forecasting questions, each of which was displayed on a separate page. At the top of each page, participants were provided with (a) a probe temperature; (b) the month and year during which the probe temperature was recorded; and (c) a statement indicating that the probe temperature was the 5<sup>th</sup> [warmest/coldest] temperature on record for the stated month (in the stated year). Across pages, probe temperatures varied according to a 2 × 2 factorial design that manipulated *rank* (5<sup>th</sup>

---

<sup>16</sup> Participants were excluded from analysis if their SPJs yielded inestimable credibility functions on more than 10 bootstrap trials (for more details, see the General Method section below). In the three cases that were excluded, this occurred because the participants in question had zero variance in their responses.

warmest vs. 5<sup>th</sup> coldest) and *month* (January vs. July), as recorded at the Philadelphia International Airport during the 10-year span between 2008 to 2017.

For each probe temperature, participants were asked to provide three numerical SPJs. The first asked participants to estimate the probability of observing a more extreme temperature on the following day (i.e., cooler in January and warmer in July); the second asked participants to estimate the probability of observing a temperature 5-degrees warmer on the following day; and the third asked participants to estimate the probability of observing a temperature 5-degrees cooler on the following day. For example, one set of questions was:

The 5<sup>th</sup> lowest temperature in January 2008 was 35 degrees F.

What is the probability (in %) that the next day's temperature was lower than 35?  
What is the probability that the next day's temperature was 40 or higher?  
What is the probability that the next day's temperature was 30 or lower?

Questions were presented in calendar order, with low probe temperatures preceding high probe temperatures for each month. After providing estimates for each probe temperature, participants completed the final section of the study, which consisted of the eleven-item Actively Open-Minded Thinking scale used by Baron (*in press*) in his research on actively open-minded thinking in American politics (this scale is based on previous scales developed by: Baron, Scott, Fincer, & Metz, 2015; and Haran, Ritov, & Mellers, 2013; see also: Baron, 2008).

**General procedure.** As discussed above, Studies 2a-2c and 3a-3c all examined the real-world performance of the linear credibility framework with parallel procedures. In general, those procedures were as follows:

1. Select a calibration sample size ( $n_{\text{cal}}$ ); a minimum prediction sample size ( $n_{\text{pred}}$ ); and a number of bootstrap trials ( $n_{\text{boot}}$ ) that are appropriate to the data-set and analysis at hand. For additional details on how this was done in Studies 2a-2c and 3a-3c, see the Detailed Procedure sections of Analyses 2a.i (GJP reliability), 2b.i (MM reliability), and 2c.i (PHL reliability), respectively.
2. For a given data-set, extract a subset of observations that is consistent with the parameters selected in step 1. In most cases, this will involve excluding participants who did not provide a sufficient number of judgments to estimate credibility and the effects of recalibration on the same trial.
3. Within the working data-set extracted in step 2, generate aggregate judgments for each forecasting question by averaging SPJs across participants (within-question). In cases where participants provided more than one SPJ for the same forecasting question, first average these SPJs (within-participant) to arrive at participant-level aggregates.
4. Use the following extremization function to estimate the optimal transformation constant  $a$ , that minimizes the Brier scores of the aggregates calculated in step 3

(this is done to optimally correct for judgmental regression. For a detailed discussion of regression and its origins, see Baron et al., 2014).

$$t(p) = \frac{p^a}{p^a + (1-p)^a}, \quad (1)$$

Where  $p$  is one of the aggregate probability judgments calculated in step 3, and  $t(p)$  is an optimized aggregate.

5. Apply this optimal transformation to the aggregate probabilities calculated in step 3 to arrive at a set of optimized aggregates, or *estimated optima*.
6. Transform all SPJs extracted in step 2 and all *estimated optima* calculated in step 5 to log-odds using the standard logit link function.

After preparing the data, I then estimated credibility for each participant 100 times, selecting a new calibration sample, estimating a new credibility function, and recalibrating judgments on each trial. Similar to Study 1, the purpose of these procedures was to estimate two sets of outcome measures for each participant: (a) bootstrapped credibility estimates (for use in Studies 2a-2c); and (b) bootstrapped summary-statistics representing the typical effects of recalibration (for use in Studies 3a-3c).

7. For participant  $i$  on trial  $n$ , select a calibration sample by randomly sampling  $n_{\text{cal}}$  observations from his or her pool of SPJs, where  $n_{\text{cal}}$  is the calibration sample-size selected in step 1.
8. Regress this calibration sample on the corresponding *estimated optima* calculated in step 6 to arrive at a credibility function with estimated coefficients corresponding to *bias* ( $\hat{\alpha}_{in}^*$ ) and *expertise* ( $\hat{\beta}_{in}^*$ ), and a standard error corresponding to *consistency* ( $\hat{\sigma}_{in}^*$ ):

$$t(p) = \hat{\alpha}_{in}^* + \hat{\beta}_{in}^* y_{in}^* + \varepsilon, \quad \varepsilon \sim N(0, \hat{\sigma}_{in}^{*2}), \quad (2)$$

Where  $i$  is a participant; the asterisk symbol (\*) indicates that an estimate was derived from a participant's calibration sample (here, a random subset of his or her judgments rather than a stand-alone sample);  $y_{in}^*$  is a vector of log-odds transformed judgments corresponding to participant  $i$ 's beliefs about the  $n_{\text{cal}}$  events associated with her calibration sample;  $t(p)$  corresponds to the vector of log-odds transformed *estimated optima* calculated in step 6; and  $\hat{\alpha}_{in}^*$ ,  $\hat{\beta}_{in}^*$ , and  $\hat{\sigma}_{in}^*$  are the estimated parameters of participant  $i$ 's credibility function, as defined above.

9. Record the values of  $\hat{\alpha}_{in}^*$ ,  $\hat{\beta}_{in}^*$ , and  $\hat{\sigma}_{in}^*$  for the current trial,  $n$ .

10. For each participant, recalibrate judgments in the prediction sample (i.e., all SPJs that were not included in the calibration sample) according to the linear relationship captured by the credibility function estimated in step 8 (i.e., “plug” a participant’s prediction sample  $y_{in}^*$  into Eq. 2 to estimate  $t(p)$ ).
11. After recalibration, examine (and record) the impact of step 10 by comparing the accuracy of judgments before and after recalibration (for a specific list of outcome measures, see the Detailed Procedure sections for the recalibration analyses presented in Studies 3a-3c).
12. Conduct steps 7-11 a total of 100 times for each participant  $i$ , selecting a new, random subset of her SPJs for use as a calibration sample on each trial.
13. Finally, average across the sampling distributions generated in steps 9 and 11 for each participant to calculate (a) bootstrapped credibility estimates for each participant (for use in Studies 2a-2c); and (b) bootstrapped summary-statistics representing the typical effects of recalibration for each participant (for use in Studies 3a-3c).

**Procedural contrasts with Study 1.** For each of the data-sets used in Studies 2a-2c and 3a-3c, the performance of the linear credibility framework was examined using the General Procedure, described above. As in Study 1, these analyses were intended to shed light on the typical results of credibility estimation and credibility-based

recalibration across forecasters. However, because these analyses involved empirical sets of SPJs and latent probabilities (rather than simulated judgments and known “signal” values), carrying them out required several modifications to Study 1’s procedure. Because these modifications may not have been salient in the General Procedure, I discuss them in more detail, below.

***Optimizing aggregates by minimizing Brier scores.*** In each of the data-sets described above, participants provided SPJs about actual, uncertain events in the world. As such, it was difficult to identify “true” or “objective” probability values against which to compare their judgments. In the absence of this information, it was no longer possible to optimize crowd aggregates by minimizing the sum of squared deviations between crowd aggregates and the corresponding “objective” values.<sup>17</sup> To rectify this problem, Studies 2a-2c and 3a-3c optimized crowd aggregates by finding the transformation constant  $a$  that minimized empirical *Brier scores* (i.e., maximized empirical accuracy: see step 4 of the General Procedure above; for a detailed discussion of the rationale behind this modification, see: Baron et al., 2014) —a procedure that can be carried out in any case where at least some outcomes are known.

---

<sup>17</sup> Notably, detailed historical records allowed for the estimation of historical baserates concerning Philadelphia air-temperatures. However, normative criteria of this sort are not widely available to real-world decision makers. In the interest of making a general case for the value and applicability of the linear credibility framework, therefore, I do not consider the case where baserates are available for use as *estimated optima*.

***Partitioning judgments into calibration samples and prediction samples.***

Because participants in Studies 2a-2c and 3a-3c provided only a single set of judgments, examining the effects of credibility-based recalibration required that I partition each participant’s SPJs into a *calibration sample* and a *prediction sample*. By doing so, I was able to estimate a participant’s credibility function on a small subset of his or her judgments (the calibration sample) and examine the effects of recalibration on the remaining, “out-of-sample” judgments (the prediction sample). Within each data-set, the size of the calibration sample was held constant across participants and was determined by a series of reliability experiments that I report in Chapter 2. In the results sections for Study 2a-2c, these reliability analyses are designated by the Roman numeral “i” (i.e., Analyses 2a.i, 2b.i, and 2c.i), and are all accompanied by the parenthetical designation “[data-set] reliability.”

***Limiting analyses to forecasters with a sufficient number of judgments.*** For all analyses associated with the GJP and Philadelphia air temperature study (i.e., Studies 2a, 2c, 3a, and 3c), forecasters were excluded from analysis if they did not provide a sufficient number of judgments. In each of these cases, a “sufficient number” was defined as the calibration sample size plus thirty ( $n_{\text{cal}} + 30$ ) to ensure that each forecaster’s prediction sample would be sufficiently large to estimate the effects of recalibration (i.e., the minimum prediction sample size, or  $n_{\text{pred}} = 30$ ). However, because forecasters were limited to a maximum of 67 judgments in the March Madness data, this restriction was not applied in Studies 2b and 3b, where calibration sample sizes were often  $> 37$ .

***Estimating a new credibility function for each bootstrap trial.*** To examine the typical performance of the linear credibility framework, it is necessary to examine credibility estimates and the effects of credibility-based recalibration across a large number of trials. However, because each participant provided only a single set of judgments, there is no way to run multiple trials with the same participant while holding their calibration sample (and credibility function) constant — as was done in Study 1. To solve this problem, Studies 2a-2c and 3a-3c used random sampling procedures to select a small subset of each participant’s SPJs (the calibration sample) on each bootstrap trial. Across trials, credibility functions were fit to only those SPJs in the calibration sample, allowing for variance in the performance of the linear credibility framework from trial-to-trial.

***Examining additional standards of “improved” judgment.*** Finally, because it was difficult to identify normative standards for judgmental “improvement” in Studies 3a-3c, tests of recalibration in these studies tracked a wider variety of outcome measures than those examined in Study 1. As in Study 1, these measures fell into four (generic) categories for each prediction sample: (a) the average proportion of judgments that improved; (b) the mean pairwise difference in judgments; (c) the average effect-size (Cohen’s  $d$ ) associated with recalibration; and (d) the proportion of samples in which the mean outcome measure improved.

However, because “objective” probabilities were not available in Studies 3a-3c, it was no longer possible to evaluate judgments in terms of the absolute difference between judgments and the “truth” (i.e., absolute judgment error, or AJE). To account for this, Studies 3a-3c measured outcomes in terms of (a) a variant of AJE that measures the absolute difference between an SPJ and the corresponding *estimated optimum* (here, an optimized crowd aggregate); and (b) a measure that I call absolute linear error (ALE), which corresponds to the absolute difference between an SPJ and the event’s empirical outcome. In Study 3c (PHL recalibration), the existence of detailed historical records made it possible to estimate the baserates of various air temperatures in January and July (for details, see the Detailed Procedure section of Study 3c). Thus, in Study 3c, I also examined a variant of AJE that measures the absolute difference between SPJs and estimated baserates.

Finally, in all three studies, I also examined the effects of credibility-based recalibration on *reliability* — one component of Murphy’s (1973) three-component decomposition of the Brier score (Brier, 1950). As a summary statistic, *reliability* is defined as the weighted sum of squared-differences between an individual’s SPJs and within-sample baserates (here, the relative frequencies of event occurrence when SPJs are separated into 101 percentage-point “bins”). As such, *reliability* bears a close mathematical relationship to the forecasting term *calibration* (for an overview, see: Lichtenstein, Fischhoff, & Phillips, 1982) and can generally be understood as a measure of agreement between SPJs and empirical baserates (though it is defined in such a way that lower values indicate better agreement). While less interpretable than AJE and ALE,

my main reason for examining *reliability* is that it can help provide insight into *why* recalibration influences accuracy, rather than simply summarizing its effects.

## References

- Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3), 691-706.
- Baron, J. (2008). *Thinking and deciding*. New York: Cambridge University Press.
- Baron, J. (in press). Actively open-minded thinking in politics. *Cognition*.
- Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2), 133-145.
- Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2), 133-145.
- Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, 4(3), 265-284.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3.
- Cross, D., Ramos, J., Mellers, B., Tetlock, P. E., & Scott, D. W. (2018). Robust forecast aggregation: Fourier L2E regression. *Journal of Forecasting*, 37(3), 259-268.
- Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, 62(2), 410-422.
- Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, 8(3), 188-201.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. D. Kahneman, P. Slovic, and A. Tversky (Eds.) *Judgement under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How

- generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Making*, 12(4), 369-382.
- Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015a). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, 21(1), 1-14.
- Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015b). Identifying and cultivating superforecasters as a method of improving probabilistic prediction. *Perspectives on Psychological Science*, 10(3), 267-281.
- Mellers, B., Ungar, L., Baron, J., Ramos, J., Gürçay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5), 1-10.
- Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 595-600.
- Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2), 344-356.
- Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. New York: Random House.

## CHAPTER 2

### LINEAR CREDIBILITY ESTIMATES ARE RELIABLE AND VALID PREDICTORS OF FORECAST ACCURACY

#### **Abstract:**

At the conceptual level, measures of credibility are intended to provide information about “skill” or “proficiency” in subjective probability judgment. However, it is an empirical question whether these types of information can be extracted from empirical models of credibility. To test whether they can, this chapter examines the extent to which “credible” judgment is related to “good” judgment in three real-world contexts. Specifically, in Studies 2a-2c, I examine (a) the conditions under which linear credibility estimates can be reliably estimated; and (b) the convergent validity of linear credibility estimates with other individual-level measures of performance in probabilistic prediction. The results of these studies demonstrate that linear credibility estimates are often reliable and valid predictors of forecast accuracy and generally occupy sensible positions within the larger nomological network related to “good” judgment. Thus, I conclude that empirical credibility estimates may often be indicative of “skill” or “proficiency” in subjective probability judgment.

#### **Introduction**

In the introduction to this dissertation, I arrived at the notion of credibility by working backwards from a conceptual definition of “objective” probability. Though philosophically fraught, this definition helped to clarify the concept of “good” judgment and allowed me to define “better” vs. “worse” judgment in terms of “agreement with the truth.” With this definition in mind, I then argued that in any given decision environment, there must always be a judgment that best describes the true state of nature, given the information that is available at the time. In some cases, a decision maker may fail to identify this “optimal” judgment because he or she does not have access to the full scope of extant information. In others, this “optimal” judgment may stray from the “truth”

because the full scope of extant information is biased or incomplete with respect to the true state of nature. In all cases, however, the defining feature of this “optimal” judgment is that it reflects the most epistemically defensible belief that one could hold, given the information in the decision environment (i.e., it maximizes “agreement with the truth” to the greatest extent that evidence and reason allow). In principle, therefore, this sort of “optimal” judgment represents an appealing benchmark by which to evaluate the “quality,” or relative validity, of subjective probability judgments (SPJs).

To formalize this approach to evaluation, I then defined credible judgment in terms of the relationship between an individual’s SPJs and empirical approximations of “optimal” judgments, which I call *estimated optima*. At the conceptual level, I did not concern myself with a specific method for identifying *estimated optima* or modeling the relationship between *estimated optima* and an individual’s SPJs. Instead, I simply defined credibility as an abstract construct that describes the relationship between an individual’s judgments and (estimates of) the “best” judgments that one could hold, given the information that was available at the time. With this definition in mind, I then went on to argue that decision makers might benefit from examining credibility in real-world judgment. Specifically, if credibility can be modeled in an informative way, then access to credibility information should help decision makers (a) evaluate an individual  $i$ ’s relative “quality” as a source of probabilistic information; and (b) reverse-engineer what an individual  $i$  “should have said” or “would be justified in believing,” given nothing more than her SPJs.

In practice, however, the costs associated with extracting this information play a key role in determining whether decision makers will benefit from examining credibility. If simple (and therefore, “cheap”) models of credibility can deliver the types of information that the concept of credibility promises, then examining credibility may be useful across a wide variety of domains. If instead, simple models of credibility fail to deliver this information, then (a) the decision environment may not be amenable to examining credibility (i.e., the credibility relationship may be prohibitively weak or difficult to identify); or (b) the benefits of doing so may not outweigh the costs. From an empirical perspective, therefore, the performance of simple models of credibility (e.g., the linear credibility framework) can serve as a useful litmus test for the value of empirical models of credibility, more generally.

Following from this line of reasoning, Studies 2a-2c will test whether the linear credibility framework appears to capture information about SPJ “quality.” To determine whether they do, I will leverage the fact that credibility estimates can be interpreted as measures of “skill” or “proficiency” in subjective probability judgment (see: the introduction Chapter 1). If linear credibility estimates can be shown to be meaningfully correlated with other measures of “skill” or “proficiency” in subjective probability judgment (e.g., forecast accuracy), then it is reasonable to conclude that even simple models of credibility can provide valid information about SPJ “quality.” Indeed, because the linear credibility framework is one of the most rudimentary models that a decision maker might apply, demonstrating that linear credibility estimates predict “good” judgment (i.e., forecast accuracy) would suggest that (a) there is non-trivial amount of

information to be gained by examining credibility; and (b) this information can be extracted with simple statistical tools. Thus, if the linear credibility framework can be shown to provide even a small degree of traction on the question of who tends to be a “better” vs. “worse” judge of subjective probability, then credibility information may represent a widely available (yet generally untapped) resource for evaluating beliefs.

### **Study 2a: Reliability and Validity of Credibility Estimates Derived from GJP Data (GJP Reliability/Validity)**

Throughout its four-year lifespan, the Good Judgment Project (GJP) produced a remarkable set of results. Though too extensive to detail here (for summaries, see: Mellers, Ungar, Baron, Ramos, Gürçay, Fincher, Scott, Moorse, Atanasov, Swift, Murray, Stone, & Tetlock, 2014; Mellers, Stone, Atanasov, Rohrbaugh, Metz, Ungar, Bishop, Horowitz, Merkle, & Tetlock, 2015; and Mellers, Stone, Murray, Minster, Rohrbaugh, Bishop, Chen, Baker, Hou, Horowitz, Ungar, & Tetlock, 2015), three of these results are particularly relevant to the study of credibility. First, researchers in the GJP discovered that forecast accuracy can be improved by simple interventions such as training participants in probabilistic reasoning or having participants work in teams (Mellers et al., 2014). Second, exploratory analyses showed that the GJP’s most accurate forecasters (i.e., the top 2% of participants, known as *superforecasters*) were not dramatically different from other forecasters in terms of intelligence or education. Instead, these individuals were distinguished by their tendency towards a flexible

cognitive-style that emphasized (e.g.) the enjoyment of difficult problems and puzzles; cognitive reflection; and actively open-minded thinking (Mellers et al., 2015b). Finally, comparisons of superforecasters with other forecaster in the GJP indicated that the effects of environmental enrichment (i.e., teaming, training, and superforecaster status) tended to *increase* over time, rather than regressing toward the mean (Mellers et al., 2015b). Taken together, these results led researchers in the GJP to conclude that forecast accuracy was not the product of inborn talent or domain-specific expertise, but instead the result of a cultivatable set of “skills” — many of which encouraged a flexible, elaborative, and unbiased approach to reasoning (Mellers et al., 2015a; Mellers et al., 2015b; Mellers et al., 2014).

Given these findings, data from the Good Judgment Project provide a rich nomological backdrop against which to explore the relationships between linear credibility estimates and other indicators of “good” judgment. Indeed, across four years of tournaments, the GJP (a) provided participants with a large number of questions that varied across a wide variety of geopolitical topics; (b) observed reliable and non-trivial differences in forecast accuracy; and (c) were able to attribute these differences to a succinct (and sensible) set of individual-difference variables. As a result, data from the GJP provide a unique opportunity to compare the predictive validity of linear credibility estimates to that of the most extensive profile of “skillful” or “proficient” forecasting yet reported in the literature.

### **Analysis 2a.i: Under what conditions are credibility estimates reliable? (GJP reliability)**

As discussed in the General Procedure, a prerequisite for credibility estimation and credibility-based recalibration is the selection of three parameters: (a) a calibration sample size,  $n_{cal}$ ; (b) a minimum prediction sample size,  $n_{pred}$ ; and (c) the number of bootstrap trials over which to examine outcomes,  $n_{boot}$ .

To determine appropriate values for these parameters, it is important to consider how each will influence the credibility estimation and/or credibility-based recalibration procedures. In the case of prediction sample size ( $n_{pred}$ ), selecting an appropriate value is simple, as varying this parameter has only two consequences: (a) to change the number of predictions available for out-of-sample recalibration; and (b) under some circumstances, to change the overall number of forecasters included in the analysis (in step 2 of the General Procedure, forecasters are excluded if they have not provided at least  $n_{cal} + n_{pred}$  SPJs). Because forecasters in the GJP typically provided a large number of forecasts ( $Mean = 141.52$ ;  $Mdn. = 119.5$ ;  $SD = 62.01$ ), their likelihood of exclusion on this basis was low. Thus, the minimum prediction sample size for Study 2a was set at an *a priori* value of 30 to ensure that each forecaster would have an adequate number of observations for estimating the typical effects of recalibration.

In contrast to  $n_{pred}$ , selecting a calibration sample size ( $n_{cal}$ ) and a number of bootstrap trials ( $n_{boot}$ ) is often less straightforward. From an empirical perspective, setting either of these parameters too low could compromise the reliability and validity of the resulting estimates, while setting them too high could result in wasted computational

resources and prohibitively long run-times (especially for large data-sets, such as that drawn from the GJP). To get the most out of the linear credibility framework, therefore, it is important to select appropriate values for each of these parameters. However, because the reliability and validity of linear credibility estimates depend on the degree of noise and/or complexity in the informational environment, it is often difficult to identify an appropriate set of values *a priori*. To address this issue, I began Study 2a with an analysis of the parameter ranges under which it is possible to extract reliable credibility estimates from GJP data.

### **Method.**

**Detailed Procedure.** To identify an appropriate set of analytic parameters for Study 2a, I conducted an experiment. Because there are several ways that reliability can impact the performance of the linear credibility framework, this experiment was divided into two arms, each of which addressed a different aspect of reliability.

In the first arm, I examined the reliability of non-bootstrapped credibility estimates (i.e., non-aggregated estimates of *bias*, *expertise*, and *consistency* drawn from individual bootstrap trials), varying by calibration sample size ( $n_{cal}$ ). The purpose of this arm was to shed light on the variability of the recalibration transformation from trial-to-trial, and to identify an  $n_{cal}$  at which its effects would be relatively consistent. To gather data for this portion of the experiment, I used the random sampling procedures described in steps 7-9 of the General Procedure to record 30 estimates of *bias*, *expertise*, and

*consistency* (i.e.,  $\hat{\alpha}_{in}^*$ ,  $\hat{\beta}_{in}^*$ , and  $\hat{\sigma}_{in}^*$ )<sup>18</sup> for each forecaster at each level of  $n_{cal} = \{10, 20, \dots, 50\}$ . For each combination of  $n_{cal} \times$  component of credibility = {*bias*, *expertise*, *consistency*}, I then calculated the intraclass correlation (ICC) across these 30 observations.

In the second arm of the experiment, I examined the reliability of bootstrapped credibility estimates while varying calibration sample size ( $n_{cal}$ ) and number of bootstrap trials ( $n_{boot}$ ) according to a  $5 \times 12$  factorial design. The purpose of this arm was to identify a set of parameter values that would ensure bootstrapped credibility estimates were well-suited to the analyses of validity that follow. To gather data for this portion of the experiment, I followed the full set of steps outlined in the General Procedure. For each combination of  $n_{cal} = \{10, 20, \dots, 50\} \times n_{boot} = \{10, 20, \dots, 100, 200, 250\}$ , I conducted three full iterations of the general credibility estimation procedure and recorded bootstrapped credibility estimates ( $\hat{\alpha}_i$ ,  $\hat{\beta}_i$ , and  $\hat{\sigma}_i$ ) for each forecaster. After completing these iterations, I then calculated the intraclass correlation across these three observations (within each experimental cell).

---

**Results.** The results of Analysis 2a.i can be seen in Figures 6-9. In Figure 6, the reliability of non-bootstrapped estimates of *bias*, *expertise*, and *consistency* are plotted in

<sup>18</sup> An *a priori* sample size of 30 was selected for this experiment to ensure that the random sampling procedures described in steps 7-9 of the General Procedure would produce representative samples of credibility estimates across trials. To ensure that this sample size did not introduce undue bias, this analysis was also conducted with sample sizes of 10 and 20, neither of which had a substantive impact on the results.

a single graph, with the y-axis representing reliability (i.e., intraclass correlation, or ICC); the x-axis representing the calibration sample size; and separate curves representing different components of credibility. In Figures 7-9, the reliabilities of bootstrapped credibility estimates are plotted in separate graphs, with the y-axis of each graph representing reliability (ICC); the x-axis of each graph representing the bootstrap sample size; and separate curves representing different calibration sample sizes.

Figure 6

*[GJP data]: Reliability of non-bootstrapped estimates of credibility, varying by calibration sample size ( $n_{cal}$ ).*

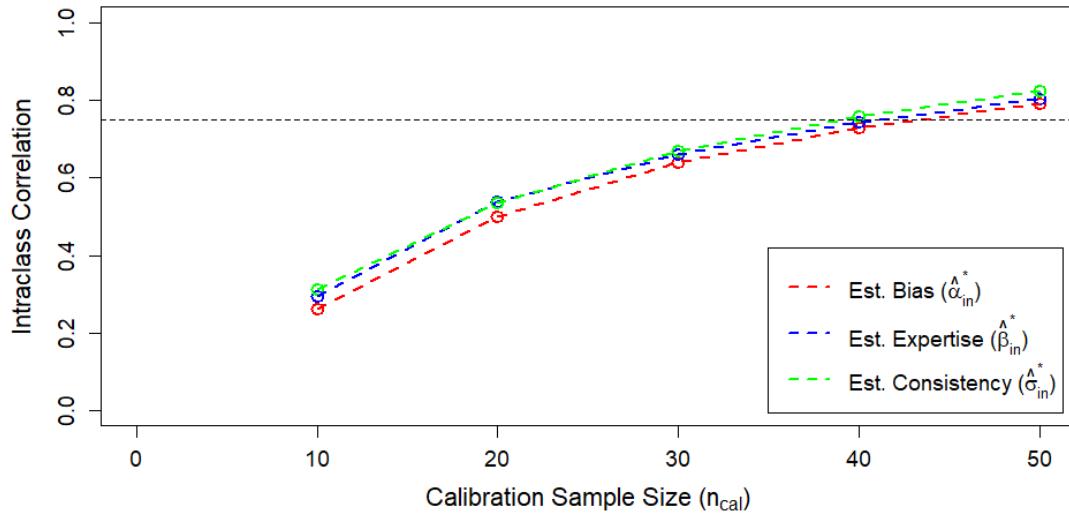


Figure 7

*[GJP data]: Reliability of bootstrapped estimates of bias ( $\hat{\alpha}_i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ).*

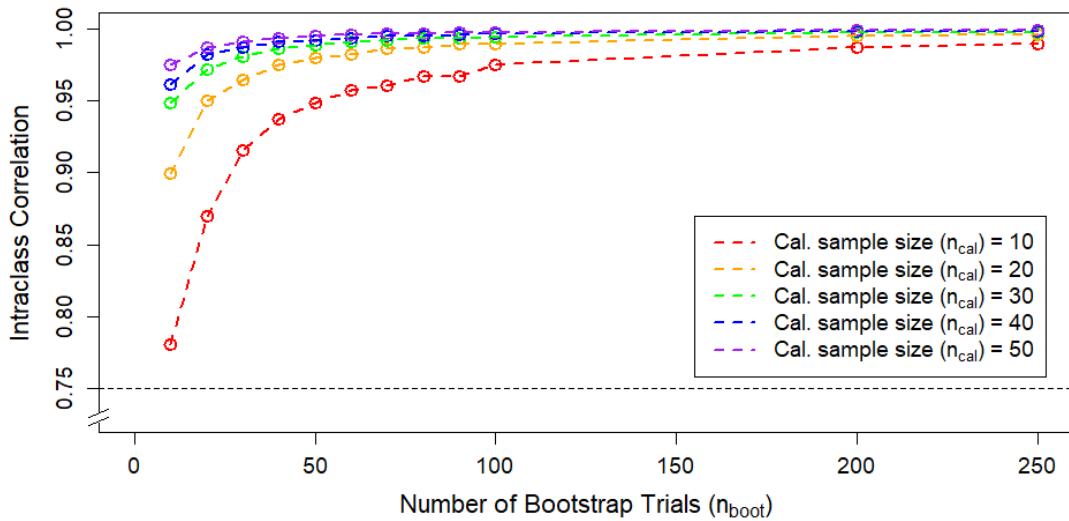


Figure 8

[GJP data]: Reliability of bootstrapped estimates of expertise ( $\hat{\beta}_i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ).

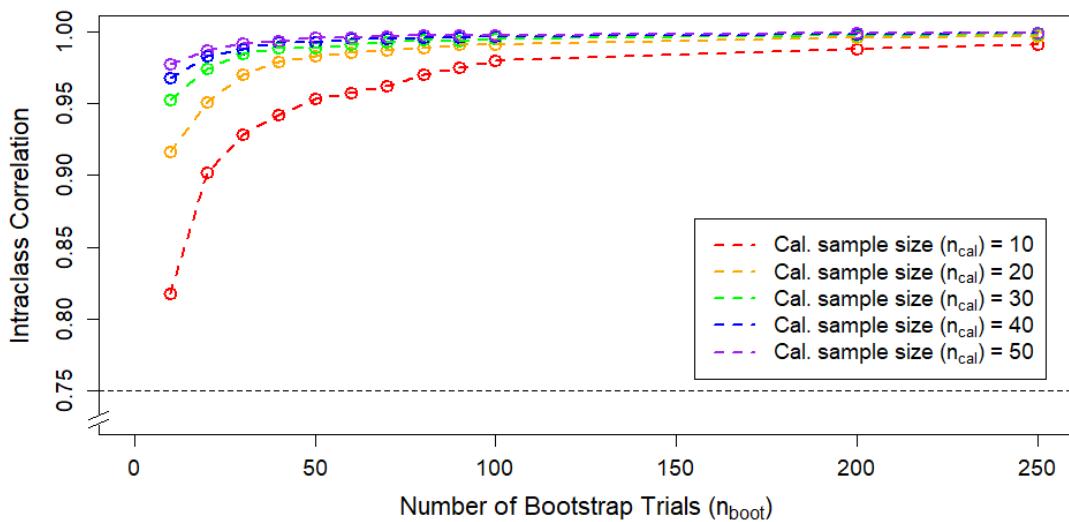
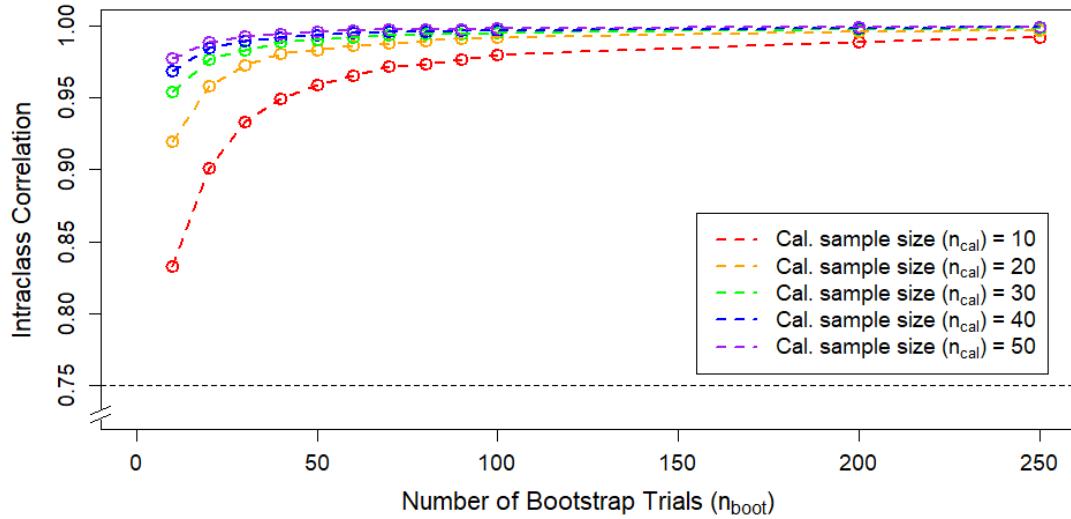


Figure 9

[GJP data]: Reliability of bootstrapped estimates of consistency ( $\hat{\sigma}_i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ).



**Discussion.** For the linear credibility framework to be useful, the estimates it produces must provide insight into forecaster's relative degree of "skill" or "proficiency" in subjective probability judgment. To achieve this goal, credibility estimates must be reliable enough to ensure that estimation errors are unlikely to be mistaken for genuine between-subjects differences.

Practically speaking, however, it is difficult to know what constitutes a suitably reliable estimate of credibility. In the absence of specific error tolerances, the only reasonable criterion is that credibility estimates should be "reliable enough" to allow forecasters to be correctly ordered according to their "skill" or "proficiency." In statistical terms, this means that the error variance in an individual's credibility estimates should be considerably smaller than the between-subjects variance in the same measure of credibility. If this condition is met, then it is unlikely that the effects of measurement

error will result in forecasters being sorted into the wrong order. To determine whether credibility estimates from the GJP met this condition, Analysis 2a.i examined intraclass correlation — a statistic that directly compares within-group variance to between-group variance. In Analysis 2a.i, the “groups,” or classes being examined were the forecaster-level credibility estimates (either bootstrapped or non-bootstrapped) produced by different iterations of the same estimation procedure. If the estimation procedure was reliable, then the within-class variance in credibility estimates (i.e., the “measurement error” associated with the estimation procedure) will be small relative to the variance between classes (i.e., the total variance, across forecasters), and the intraclass correlation will be high.

If we take the conventional threshold for “excellent” intraclass correlation as our benchmark ( $ICC \geq 0.75$ ; Cicchetti, 1994), then the results on Analysis 2a.i indicate that the linear credibility framework is capable of producing reliable credibility estimates when fit to data from the GJP. In the case of non-bootstrapped credibility estimates (Figure 6), the reliability of all three components of credibility reached the threshold for “excellent” at a calibration sample size of  $n_{cal} = 40$  and continued to climb thereafter. Thus, it is likely that linear credibility estimates can be used to sort GJP forecasters into the correct skill-order (and for the recalibration transformation to be relatively consistent from trial-to-trial) with as few as 40 observations per forecaster.

In addition, the results of Analysis 2a.i also indicate that bootstrapping can be used to dramatically improve the reliability of credibility estimates in the GJP. Indeed, across all 180 observations (3 components of credibility  $\times$  5 levels of calibration sample

size  $\times$  12 levels of bootstrap sample size), there were no cases in which the ICC of bootstrapped credibility estimates fell below the conventional threshold for “excellent.” Furthermore, an examination of all cases where  $n_{\text{cal}} \geq 40$  demonstrates that the ICC of bootstrapped credibility estimates never fell below 0.95 — a cut-off indicating that the between-group variance (i.e., the total variance associated with individual differences in credibility) was nineteen times larger than the within-group variance (i.e., the variance attributable to “estimation error”). Thus, by applying a small amount of computational power to the same sample size as before (40 observations per forecaster), it is exceedingly likely that bootstrapped credibility estimates could be used to correctly rank-order forecasters in the GJP — a sign that bodes well for later examinations of validity.

With the above results in mind, I selected the following analytic parameters for Study 2a: a calibration sample size of  $n_{\text{cal}} = 50$ ; a minimum prediction sample size of  $n_{\text{pred}} = 30$ ; and a bootstrap sample size of  $n_{\text{boot}} = 100$ . In making these selections, calibration sample size was set to  $n_{\text{cal}} = 50$  to maximize the likelihood of reliable estimates while still maintaining a plausible real-world sample size. As discussed above, the minimum prediction sample size was set at an *a priori* value of  $n_{\text{pred}} = 30$  to ensure that forecasters would have a sufficient number of SPJs to examine the effects of recalibration out-of-sample. And finally (despite being more computationally demanding than was strictly necessary), the number of bootstrap trials was set to a value of  $n_{\text{boot}} = 100$  to ensure that the effects of credibility-based recalibration could be observed across a sufficient number of trials.

## **Analysis 2a.ii: What are the predictors of credibility and what does credibility predict? (GJP validity)**

As discussed above, a key indicator of the linear credibility framework's performance is the extent to which credibility estimates demonstrate convergent validity with other measures of "skill" or "proficiency" in subjective probability judgment. To examine these relationships in GJP data, I conducted a series of exploratory analyses, each of which was intended to shed light on the position of credibility within the larger nomological network related to probabilistic prediction.

### **Method.**

**Detailed procedure.** Using the analytic parameters selected in Analysis 2a.i (GJP reliability) (i.e.,  $n_{cal} = 50$ ;  $n_{pred} = 30$ ;  $n_{boot} = 100$ ), I conducted a single run of the General Procedure to arrive at bootstrapped credibility estimates for each forecaster (i.e.,  $\hat{\alpha}_i$ ,  $\hat{\beta}_i$ , and  $\hat{\sigma}_i$ ). After excluding participants who provided too few forecasts ( $\leq 80$ ), this analysis was conducted on a working data-set of 337,919 forecasts provided by 754 forecasters across all 380 questions. Within this sample, 57% of forecasters were trained in probabilistic reasoning, 47% were assigned to work in collaborative teams, and 6% were *superforecasters* (a designation given to the top 2% of forecasters from each tournament year). The average age of forecasters in this sample was 40.2 ( $Mdn. = 36.0$ ;  $SD = 13.6$ ), 87% of forecasters were male, and 70% of forecasters had some level of advanced degree.

As in Study 1, I did not expect Analysis 2a.ii to produce veridical, face-valid measures of *bias*, *expertise*, or *consistency*. Instead, I expected bootstrapped credibility estimates in this analysis to predict a forecaster’s relative degree of “skill” or “proficiency” in subjective probability judgment. To ensure that my exploratory analyses matched the spirit of this prediction, I conducted Analysis 2a.ii with transformed measures of *bias* and *expertise*, each of which corresponded to the absolute difference between a forecaster’s bootstrapped average and the value that one would expect if an individual’s credibility function were equal to identity (i.e.,  $\alpha = 0$  and  $\beta = 1$ ). To distinguish these measures from their untransformed counterparts, I will indicate each with the superscript “prime,” i.e.,  $\hat{\alpha}'_i$  and  $\hat{\beta}'_i$ , where  $\hat{\alpha}'_i = |\hat{\alpha}_i|$  and  $\hat{\beta}'_i = |\hat{\beta}_i - 1|$ , respectively.

To examine convergent validity in the GJP dataset, I then explored the covariation of bootstrapped credibility estimates ( $\hat{\alpha}'_i$ ,  $\hat{\beta}'_i$ , and  $\hat{\sigma}_i$ ) with a subset of the most strongly explanatory individual difference measures captured by the GJP. For the purposes of Analysis 2a.ii, the main correlate of interest was a forecaster’s *average Brier score*, which is widely used as a measure of forecast accuracy (where lower Brier scores indicate more accurate predictions). In the interest of completeness, however, these analyses also examined variables that fell into categories of: experimental condition (i.e., manipulations designed to “enrich” the forecasting environment such as teaming and training); motivation and engagement; numerical fluency; cognitive ability; cognitive style; and demographics.

In detail (and in the tables below), these variables are as follows: *Experimental condition* refers to a set of effects-coded indicators which describe (a) whether a forecaster was working alone or in a collaborative team; (b) whether the forecaster had received training in probabilistic reasoning; and (c) whether the individual had been designated a superforecaster (all of whom were trained in probabilistic reasoning and worked in collaborative teams). *Number of questions addressed* and *average number of updates per question* serve as proxies for a forecaster's motivation and engagement with the GJP. *Proportion of fine-grained forecasts* refers to the proportion of a forecaster's SPJs which were not multiples of 0.05 or 0.10 (which I also consider a measure of motivation and engagement). *Threshold of estimative precision* is a measure that describes a forecaster's ability to make fine-grained distinctions on the probability scale (for additional details on the measure, see: Friedman, Baker, Mellers, Tetlock, & Zeckhauser, 2018). *Composite Berlin Numeracy* is a forecaster's average score on the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) measured over all four years of the GJP tournaments. *Composite Raven's* is a forecaster's average score on the Raven's progressive matrices test (Bors & Stokes, 1998), measured during years 3 and 4. *Composite CRT* is a forecaster's average score on the extended Cognitive Reflection Task (Baron et al., 2015; see also: Frederick, 2005), measured in years 3 and 4. *Composite Need for Cognition* is a forecaster's average score on the Need for Cognition questionnaire (Cacioppo & Petty, 1982), measured in years 1, 2, and 4. *One question Fox-Hedgehog* is a single-item scale measuring the extent to which a forecaster's approach to problem solving tends to rely on knowing "one big thing" rather

than “many little things” (for additional details on this measure, see: Tetlock, 2005).

*AOMT* is a forecaster’s score on Baron’s 11-item Actively Open-Minded Thinking scale (*in press*; see also: Baron, Scott, Fincer, & Metz, 2015; Haran, Ritov, & Mellers, 2013; Baron, 2008). And *Age*, *Male*, and *Education* are standard descriptors of a forecaster’s demographics (with education being split into effects-coded indicators).

In addition to simple correlations, I also conducted exploratory linear regressions to examine the convergent validity of bootstrapped credibility estimates with individual difference variables from the GJP. Specifically, I examined (a) which GJP measures predicted the three components of credibility; and (b) whether bootstrapped credibility estimates were meaningful predictors of forecast accuracy (as measured by average Brier score). In all cases, continuous variables in these regressions were standardized and categorical variables were effects-coded. Finally, because the GJP data-set presented a non-trivial likelihood of redundancy, multi-collinearity, and/or suppression effects among credibility estimates and GJP predictors, each of these regressions was conducted under three approaches to variable selection.

In the first approach, which I call the *kitchen sink* approach, I included all predictors that were neither the variable being predicted nor the other two components of credibility (if a component of credibility were being predicted). In the *reduced* approach, I included only those predictors that were statistically significant in the kitchen sink model. And in the *ridge-ISE* approach, I used the R function cv.glmnet (from the package *glmnet*: Friedman, Hastie, & Tibshirani, 2010) to conduct penalized ridge regression and select the most parsimonious, statistically justified model from the

available set of predictors (i.e., the model with the most regularized fit within one standard error of the minimum penalized score, lambda). By comparing these models, I was able to identify instances where credibility estimates were competing with other predictors for overlapping segments of predictive variance, and to weakly examine which had the stronger claim.

**Results.** The results of Analysis 2a.ii can be seen in the tables below. Similar to Study 1, pairwise correlations between bootstrapped credibility estimates and GJP individual difference measures can be seen in Table 7; exploratory regressions examining the predictors of credibility can be seen in Tables 8-10; and an exploratory regression examining the degree to which bootstrapped credibility estimates predict forecast accuracy (average Brier scores) can be seen in Table 11.

Table 7

*[GJP data]: Simple correlations between credibility estimates and individual difference measures.*

Ind. Diff. Measure	Credibility Estimate		
	$\hat{\alpha}'_i$ (Bias)	$\hat{\beta}'_i$ (Expertise)	$\hat{\sigma}_i$ (Consist.)
<i>Forecast accuracy</i>			
Average Brier score	0.74***	0.25***	0.76***
<i>Experimental condition</i>			
Individual, w/training	-0.10**	-0.05	-0.12***
Grouped, no training	-0.28***	-0.12**	-0.30***
Grouped, w/training	-0.33***	-0.05	-0.35***

Superforecaster	-0.41***	-0.09*	-0.47***
<i>Motivation and engagement</i>			
Number of questions addressed	-0.15***	-0.01	-0.20***
Avg. num. updates per question	-0.16***	-0.01	-0.21***
Prop. fine-grained forecasts	-0.04	-0.02	-0.13***
<i>Numerical fluency</i>			
Threshold of estimative precision	-0.40***	-0.13***	-0.44***
Composite Berlin Numeracy	-0.30***	-0.05	-0.29***
<i>Cognitive ability</i>			
Composite Raven's	-0.25***	-0.03	-0.29***
Composite CRT	-0.32***	-0.05	-0.36***
<i>Cognitive style</i>			
Composite Need for Cognition	-0.10**	-0.05	-0.14***
One question Fox-Hedgehog	-0.02	-0.02	-0.01
AOMT	-0.16***	-0.04	-0.15***
<i>Demographics</i>			
Age	-0.01	0.06	-0.02
Male	-0.10**	0.03	-0.09*
Education = bachelor's degree	0.07	0.00	0.05
Education = master's degree	-0.03	-0.02	-0.01
Education = doctorate	-0.05	0.03	-0.06

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

Table 8

[GJP data]: Predictors of bootstrapped alpha (i.e., bias;  $\hat{\alpha}'_i$ ).

Ind. Diff. Measure	Variable Selection Approach		
	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	0.02 (0.08)	-0.09 (0.03)***	-0.06 (0.02)*

	Model A	Model B	Model C
<i>Forecast accuracy</i>			
Average Brier score	0.61 (0.03)***	0.64 (0.03)***	0.64 (0.03)***
<i>Experimental condition</i>			
Individual, w/training	0.17 (0.05)***	0.16 (0.05)***	
Grouped, no training	0.00 (0.06)		
Grouped, w/training	-0.04 (0.05)		
Superforecaster	-0.40 (0.08)***	-0.42 (0.07)***	-0.26 (0.04)***
<i>Motivation and engagement</i>			
Number of questions addressed	-0.05 (0.02)*	-0.06 (0.02)**	-0.06 (0.02)*
Avg. num. updates per question	-0.06 (0.02)*	-0.05 (0.02)*	-0.06 (0.02)*
Prop. fine-grained forecasts	0.07 (0.02)**	0.07 (0.02)**	0.08 (0.02)***
<i>Numerical fluency</i>			
Threshold of estimative precision	-0.01 (0.03)		
Composite Berlin Numeracy	-0.06 (0.03)		-0.06 (0.03)*
<i>Cognitive ability</i>			
Composite Raven's	-0.08 (0.03)**	-0.13 (0.02)***	-0.09 (0.03)***
Composite CRT	-0.05 (0.03)		-0.05 (0.03)
<i>Cognitive style</i>			
Composite Need for Cognition	0.02 (0.02)		
One question Fox-Hedgehog	0.01 (0.02)		
AOMT	-0.06 (0.02)*	-0.07 (0.02)**	-0.05 (0.02)*
<i>Demographics</i>			
Age	0.03 (0.02)		
Male	-0.08 (0.07)		
Education = bachelor's degree	0.04 (0.06)		
Education = master's degree	-0.08 (0.05)		
Education = doctorate	-0.05 (0.06)		
Multiple R <sup>2</sup>	0.628	0.616	0.617
Adjusted R <sup>2</sup>	0.618	0.612	0.613
RMSE	0.61	0.619	0.618

AIC	1437.52	1436.68	1436.72
BIC	1539.27	1482.93	1487.6

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

Table 9

[GJP data]: Predictors of bootstrapped beta (i.e., expertise;  $\hat{\beta}_i'$ ).

Ind. Diff. Measure	Variable Selection Approach		
	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	-0.13 (0.13)	-0.01 (0.04)	0.00 (0.04)
<i>Forecast accuracy</i>			
Average Brier score	0.27 (0.05)***	0.23 (0.04)***	
<i>Experimental condition</i>			
Individual, w/training	-0.07 (0.07)		
Grouped, no training	-0.21 (0.09)*	-0.09 (0.06)	
Grouped, w/training	0.12 (0.07)		
Superforecaster	0.14 (0.13)		
<i>Motivation and engagement</i>			
Number of questions addressed	0.00 (0.04)		
Avg. num. updates per question	0.01 (0.04)		
Prop. fine-grained forecasts	0.02 (0.04)		
<i>Numerical fluency</i>			
Threshold of estimative precision	-0.03 (0.04)		
Composite Berlin Numeracy	-0.01 (0.04)		
<i>Cognitive ability</i>			
Composite Raven's	0.01 (0.04)		
Composite CRT	0.04 (0.05)		
<i>Cognitive style</i>			
Composite Need for Cognition	-0.03 (0.04)		

One question Fox-Hedgehog	-0.02 (0.04)		
AOMT	-0.01 (0.04)		
<i>Demographics</i>			
Age	0.06 (0.04)		
Male	0.12 (0.11)		
Education = bachelor's degree	0.01 (0.09)		
Education = master's degree	0.02 (0.08)		
Education = doctorate	0.06 (0.09)		
Multiple R <sup>2</sup>	0.082	0.065	0
Adjusted R <sup>2</sup>	0.057	0.062	0
RMSE	0.957	0.967	0.999
AIC	2118.04	2096.41	2142.76
BIC	2219.79	2114.91	2152.01

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

Table 10

*[GJP data]: Predictors of bootstrapped sigma (i.e., consistency;  $\hat{\sigma}_i$ ).*

Ind. Diff. Measure	Variable Selection Approach		
	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	-0.03 (0.07)	-0.12 (0.02)***	-0.12 (0.02)***
<i>Forecast accuracy</i>			
Average Brier score	0.58 (0.03)***	0.59 (0.02)***	0.59 (0.02)***
<i>Experimental condition</i>			
Individual, w/training	0.20 (0.04)***	0.20 (0.04)***	0.20 (0.04)***
Grouped, no training	0.01 (0.05)		
Grouped, w/training	0.00 (0.04)		
Superforecaster	-0.59 (0.08)***	-0.58 (0.06)***	-0.58 (0.06)***
<i>Motivation and engagement</i>			

Number of questions addressed	-0.08 (0.02)***	-0.08 (0.02)***	-0.08 (0.02)***
Avg. num. updates per question	-0.11 (0.02)***	-0.11 (0.02)***	-0.11 (0.02)***
Prop. fine-grained forecasts	-0.01 (0.02)		
<i>Numerical fluency</i>			
Threshold of estimative precision	-0.01 (0.02)		-0.01 (0.02)
Composite Berlin Numeracy	-0.01 (0.03)		
<i>Cognitive ability</i>			
Composite Raven's	-0.11 (0.02)***	-0.12 (0.02)***	-0.12 (0.02)***
Composite CRT	-0.10 (0.03)***	-0.11 (0.02)***	-0.11 (0.02)***
<i>Cognitive style</i>			
Composite Need for Cognition	-0.02 (0.02)		
One question Fox-Hedgehog	0.03 (0.02)		
AOMT	-0.03 (0.02)		
<i>Demographics</i>			
Age	0.01 (0.02)		
Male	-0.07 (0.07)		
Education = bachelor's degree	0.01 (0.05)		
Education = master's degree	-0.05 (0.05)		
Education = doctorate	-0.05 (0.05)		
Multiple R <sup>2</sup>	0.695	0.691	0.691
Adjusted R <sup>2</sup>	0.686	0.688	0.687
RMSE	0.552	0.556	0.556
AIC	1288.6	1272.37	1274.22
BIC	1390.36	1314	1320.47

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

Table 11

[GJP data]: Predictors of average Brier score (i.e., forecast accuracy).

Variable Selection Approach

Ind. Diff. Measure	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	-0.01 (0.07)	0.05 (0.02)*	0.00 (0.02)
<i>Credibility</i>			
Estimated bias ( $\hat{\alpha}'_i$ )	0.34 (0.04)***	0.34 (0.04)***	0.34 (0.04)***
Estimated expertise ( $\hat{\beta}'_i$ )	0.16 (0.02)***	0.16 (0.02)***	0.16 (0.02)***
Estimated consistency ( $\hat{\sigma}_i$ )	0.43 (0.04)***	0.43 (0.04)***	0.38 (0.04)***
<i>Experimental condition</i>			
Individual, w/training	0.03 (0.04)		
Grouped, no training	0.00 (0.05)		
Grouped, w/training	-0.19 (0.04)***	-0.18 (0.04)***	-0.10 (0.03)***
Superforecaster	0.21 (0.08)**	0.24 (0.06)***	
<i>Motivation and engagement</i>			
Number of questions addressed	0.07 (0.02)**	0.07 (0.02)**	
Avg. num. updates per question	0.05 (0.02)*	0.05 (0.02)*	
Prop. fine-grained forecasts	-0.08 (0.02)***	-0.08 (0.02)***	
<i>Numerical fluency</i>			
Threshold of estimative precision	-0.17 (0.02)***	-0.17 (0.02)***	-0.14 (0.02)***
Composite Berlin Numeracy	-0.02 (0.03)		
<i>Cognitive ability</i>			
Composite Raven's	0.07 (0.02)**	0.07 (0.02)**	
Composite CRT	0.00 (0.03)		
<i>Cognitive style</i>			
Composite Need for Cognition	0.00 (0.02)		
One question Fox-Hedgehog	0.01 (0.02)		
AOMT	0.00 (0.02)		
<i>Demographics</i>			
Age	-0.02 (0.02)		
Male	0.04 (0.07)		
Education = bachelor's degree	0.02 (0.05)		
Education = master's degree	0.02 (0.05)		

Education = doctorate	0.04 (0.05)		
Multiple R <sup>2</sup>	0.698	0.697	0.672
Adjusted R <sup>2</sup>	0.689	0.693	0.67
RMSE	0.549	0.55	0.572
AIC	1282.97	1262.03	1312.64
BIC	1393.98	1317.54	1345.02

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

**Discussion.** Broadly speaking, the results of Analysis 2a.ii demonstrate that the linear credibility framework is capable of identifying valid and informative indicators of “skill” or “proficiency” in subjective probability judgment. Though far from perfect (indeed, several aspects of these results bear close examination), the general implication of these findings is that simple models of credibility can provide decision makers with a great deal of information about (a) who is likely to be a “better” vs. “worse” judge of uncertain events; and (b) why these differences are likely to exist — at least when fit to a sufficiently rich data-set, such as that provided by the GJP.

Beginning with simple, pairwise correlations, the validity of bootstrapped credibility estimates can be seen in (a) the strong positive relationship between forecast accuracy (average Brier score) and estimated *bias* ( $\hat{\alpha}'_i$ ); (b) the strong positive relationship between forecast accuracy and estimated *consistency* ( $\hat{\sigma}'_i$ ); and (c) the small-to-moderate positive relationship between accuracy and estimated *expertise* ( $\hat{\beta}'_i$ ). In and of themselves, these relationships indicate that credible forecasters are also likely to be accurate forecasters, lending support to the notion that credibility estimates are

meaningfully related to errors and biases in judgment. Practically speaking, however, the insight provided by bootstrapped credibility estimates is not limited to their value as predictors of forecast accuracy. Indeed, in examining the wider set of correlational relationships in Table 7, it is apparent that all three measures occupy sensible and interpretable places within the larger nomological network related to “skill” or “proficiency” in subjective probability judgment.

In the cases of both *bias* and *consistency*, for example, credibility estimates demonstrate consistent, convergent relationships with nearly all measures that the GJP has identified as important to successful forecasting (i.e., nearly all non-demographic items in Table 7), and little to no covariation with incidental variables such as age, sex, and education.<sup>19</sup> Though the same cannot be said for estimates of *expertise*, the relationships in Table 7 nevertheless provide some indication that better (lower) *expertise* scores are systematically related to environmental enrichment (teaming) and cognitive engagement (threshold of estimative precision). Though far from ideal for a variable that is meant to capture “good” judgment, post-hoc analyses indicate that the variance of non-standardized *expertise* scores was small in the GJP sample, suggesting a relatively homogenous population ( $Var(\hat{\beta}'_i) = 0.06$ , whereas  $Var(\hat{\alpha}'_i) = 0.37$  and  $Var(\hat{\sigma}_i) = 0.15$ ).

---

<sup>19</sup> Out of context, one might predict that successful forecasting strongly covaries with education. In practice, however, this is unlikely to be true for two reasons. First, while it is reasonable to expect that both forecasting accuracy and educational attainment are driven by something like general intelligence, the differential impact of this intelligence-like variable is likely to be relatively small among a population that excludes participants who have not yet attained a bachelor’s degree. Second, because GJP questions were limited to the geopolitical domain, it is unlikely that any domain-general variable such as education or intelligence would be sufficient to drive Brier scores. Instead, as was reported by the GJP, successful forecasting was predicted by an interactive constellation of factors such as fluid intelligence, intrinsic motivation, domain knowledge, and cognitive style (Mellers et al., 2015b).

Thus, is impressive that the linear credibility framework was able to identify covariates of *expertise* at all.

As can be seen in Tables 8-10, the results of exploratory regressions concerning the predictors of credibility tell a similar story. In Table 8, for example, bootstrapped estimates of *bias* ( $\hat{\alpha}'_i$ ) are (a) strongly related to both forecast accuracy and superforecaster status; and (b) are consistently predicted by prominent GJP measures such as Raven's progressive matrices (Bors & Stokes, 1998), actively open-minded thinking (Baron, *in press*; Baron et al., 2015; Haran et al., 2013; Baron, 2008), and various measures of motivation and engagement (e.g., number of questions addressed, average number of updates per question). In Table 10, the results are much the same for estimates of *consistency* ( $\hat{\sigma}_i$ ), with some additional evidence to suggest that training may play a role in improving the *consistency* of individuals working alone. Finally, in Table 9, the results of the reduced model indicate that estimates of *expertise* ( $\hat{\beta}'_i$ ) are meaningfully related to forecast accuracy, even if the Ridge-1SE model rejected this predictor as statistically tenuous.<sup>20</sup> Taken together, these analyses demonstrate that bootstrapped estimates of *bias*, *expertise*, and *consistency* all exhibit convergent validity with other indicators of "skill" or "proficiency" in GJP data, and — with the single exception of a

---

<sup>20</sup> Based on a qualitative examination of the data, this is likely because bootstrapped estimates of *expertise* ( $\hat{\beta}'_i$ ) in the GJP data-set are noticeably homogenous. When regressed on a (relatively) more heterogeneous criterion, the resulting regression line is likely to provide a poor fit to the data, as the limited range of the predictor values will constrain their ability to explain differences among individuals. Under such conditions, it is conceivable that an intercept-only model would provide a more regularized fit to the criterion than a model that includes *expertise*, as appears to have been the case in Table 9.

small positive relationship between *bias* and proportion of fine-grained forecasts — did not display any unexpected, inexplicable, or incorrectly-signed relationships.

Perhaps most notably, however, the results of Analysis 2a.ii demonstrate that bootstrapped estimates of credibility were strongly predictive of forecast accuracy (average Brier scores). Indeed, as can be seen in Table 11, the results of the simultaneous “contest” between GJP measures and linear credibility estimates suggests that the most parsimonious model of forecast accuracy (i.e., the Ridge-1SE model) was one that included all three measures of credibility and little else. Indeed, while this model included two predictors that provide a nod to the importance of enriched environments (experimental condition = grouped, w/training) and a general facility with numerical probabilities (threshold of estimative precision), a comparison of the ridge-1SE model with the reduced and kitchen-sink models suggests that bootstrapped estimates of credibility are in direct competition for nearly all of the explanatory variance provided by GJP variables. Though far from a conclusive test, this result is consistent with a highly attractive narrative. Specifically, the close overlap in the variance explained by these two sets of predictors suggests that the *reason* enriched forecasting environments, numerical fluency, cognitive ability, cognitive style, and motivation are all related to predictive accuracy is the *same reason* that credibility is related to predictive accuracy — namely, that all are related to less severe errors and biases in judgment. Though a direct test of this hypothesis was impossible with the existing data, I explore this possibility further, below.

### **Analysis 2a.iii: How effective are credibility estimates at predicting forecast accuracy? (GJP enrichment vs. credibility)**

In Analysis 2a.ii (GJP validity), exploratory linear regressions indicated that much of the variance in forecast accuracy (i.e., average Brier scores) that can be explained by “environmental enrichment” in the GJP (i.e., by experimental condition) can also be explained by bootstrapped estimates of credibility. Though far from conclusive, these results are consistent with the idea that the “skills” imparted by teaming, training, and superforecaster status may have increased accuracy in the GJP because they reduced the likelihood of errors and biases in judgment (i.e., increased credibility). Practically speaking, however, the results of Analysis 2a.ii do not provide any insight into the relative explanatory power (and/or redundancy) of these two sets of variables — only that “environmental enrichment” variables and linear credibility estimates are competing for the same portion of explanatory variance. Thus, in Analysis 2a.iii, I conducted a series of predictive contests between these two sets of variables to (a) compare the incremental validity of each set of predictors in models of forecast accuracy; and (b) determine whether one set of variables is redundant with (or subsumed by) the other.

#### **Method.**

**Detailed procedure.** To compare the predictive validity of bootstrapped credibility estimates with the GJP’s “environmental enrichment” variables (i.e., teaming, training, and superforecaster status), I constructed four linear regression models, each of which used forecast accuracy (i.e., average Brier scores) as its criterion. In the *baseline*

model, I included all predictors that were not drawn from the two categories of interest.

In the *environmental only* model, I added the four effects-coded predictors for “environmental enrichment” (i.e., experimental condition) to the baseline model.<sup>21</sup> In the *credibility only* model, I added the three bootstrapped credibility estimates to the baseline. And in the *full* model, I included all available predictors to establish a high-water mark for the GJP data’s overall predictive validity. To compare the incremental value of the GJP’s environmental enrichment variables with that of bootstrapped credibility estimates, I then conducted a series of likelihood-ratio tests in which I (a) contrasted the two intermediary models (*environmental only* and *credibility only*) with the *baseline* model; and (b) contrasted the *full* model with the two intermediaries. In cases where models weren’t nested (e.g., *environmental only* vs. *credibility only*), I made qualitative comparisons by examining each model’s multiple R<sup>2</sup> and Bayesian information criterion, or *BIC*.

**Results.** The four models used in Analysis 2a.iii can be seen in Table 12. Likelihood-ratio tests comparing nested models within this set indicate each of the following:

---

<sup>21</sup> When constructing the *environmental only* model and the *credibility only* model, I “added” predictors to the baseline model in the sense that I added the focal set of predictors to the pool of variables being included in the model. In all cases, all predictors for each model were estimated simultaneously.

- (a) The inclusion of environmental enrichment variables improved the fit of the *environmental only* model over the *baseline* model by a significantly greater degree than would be expected by chance,  $F(735, 739) = 29.92, p < 0.001$ .
- (b) The inclusion of bootstrapped credibility estimates improved the fit of the *credibility only* model over the *baseline* model by a significantly greater degree than would be expected by chance,  $F(736, 739) = 306.21, p < 0.001$ .
- (c) The inclusion of environmental enrichment variables improved the fit of the *full* model over the *credibility only* model by a significantly greater degree than would be expected by chance,  $F(732, 736) = 6.22, p < 0.001$ .
- (d) The inclusion of bootstrapped credibility estimates improved the fit of the *full* model over the *environmental only* model by a significantly greater degree than would be expected by chance,  $F(732, 735) = 243.75, p < 0.001$ .

This pattern of results is corroborated by an examination of multiple  $R^2$  and *BIC*.

In the case of both intermediary models (i.e., the *environmental only* model and the *credibility only* model), the addition of the focal predictors provided a sizeable improvement over the *baseline*, and the addition of the complementary predictors (to arrive at the *full* model) offered a noticeable improvement over each of the

intermediaries. Qualitatively speaking, however, the impact of adding bootstrapped credibility estimates to the *environmental only* model was considerably larger than that of adding the predictors in the reverse order.

Table 12

*[GJP data]: A comparison of credibility measures vs. environmental enrichment variables as predictors of average Brier score (i.e., forecast accuracy).*

Model				
Ind. Diff. Measure	Baseline	Envir. Only	Cred. Only	Full
(Intercept)	-0.01 (0.11)	-0.08 (0.1)	-0.06 (0.07)	-0.01 (0.07)
<i>Credibility</i>				
Est. bias ( $\hat{\alpha}'_i$ )			0.36 (0.04)***	0.34 (0.04)***
Est. expert. ( $\hat{\beta}'_i$ )			0.16 (0.02)***	0.16 (0.02)***
Est. consist. ( $\hat{\sigma}_i$ )			0.42 (0.04)***	0.43 (0.04)***
<i>Exp. condition (Envir. enrich.)</i>				
Indiv., w/train.		0.33 (0.06)***		0.03 (0.04)
Group, no train.		-0.05 (0.07)		0.00 (0.05)
Group, w/train.		-0.36 (0.06)***		-0.19 (0.04)***
Superforecaster		-0.32 (0.11)**		0.21 (0.08)**
<i>Motiv. &amp; Engage.</i>				
Num. questions	0.07 (0.03)*	0.03 (0.03)	0.08 (0.02)***	0.07 (0.02)**
Avg. n. updates	-0.06 (0.03)	-0.03 (0.03)	0.05 (0.02)*	0.05 (0.02)*
Prop. fine-grain	-0.13 (0.03)***	-0.12 (0.03)***	-0.08 (0.02)***	-0.08 (0.02)***
<i>Num. fluency</i>				
Thresh. precision	-0.45 (0.03)***	-0.36 (0.03)***	-0.16 (0.02)***	-0.17 (0.02)***
Comp. Numer.	-0.11 (0.04)**	-0.09 (0.04)*	-0.02 (0.03)	-0.02 (0.03)
<i>Cognitive ability</i>				

Comp. Raven's	-0.02 (0.04)	-0.01 (0.03)	0.08 (0.02)**	0.07 (0.02)**
Comp. CRT	-0.10 (0.04)*	-0.11 (0.04)**	0.01 (0.03)	0.00 (0.03)
<i>Cognitive style</i>				
Comp. NF Cog.	0.00 (0.03)	-0.01 (0.03)	0.00 (0.02)	0.00 (0.02)
Fox-Hedgehog	0.03 (0.03)	0.04 (0.03)	0.01 (0.02)	0.01 (0.02)
AOMT	-0.07 (0.03)*	-0.07 (0.03)*	0.01 (0.02)	0.00 (0.02)
<i>Demographics</i>				
Age	-0.03 (0.03)	0.01 (0.03)	-0.02 (0.02)	-0.02 (0.02)
Male	-0.02 (0.1)	0.01 (0.09)	0.04 (0.07)	0.04 (0.07)
Ed. = bachelor's	0.14 (0.08)	0.08 (0.07)	0.02 (0.05)	0.02 (0.05)
Ed. = master's	-0.02 (0.07)	-0.04 (0.07)	0.02 (0.05)	0.02 (0.05)
Ed. = doctorate	0.02 (0.08)	0.01 (0.07)	0.02 (0.05)	0.04 (0.05)
Multiple R <sup>2</sup>	0.298	0.397	0.688	0.698
Adjusted R <sup>2</sup>	0.284	0.381	0.68	0.689
RMSE	0.837	0.776	0.558	0.549
AIC	1905.61	1799.73	1300.22	1282.97
BIC	1984.24	1896.87	1392.73	1393.98

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

**Discussion.** The results of Analysis 2a.iii indicate that (a) the GJP's environmental enrichment variables; and (b) bootstrapped credibility estimates both demonstrate a high degree of predictive validity with respect to average Brier scores (i.e., forecast accuracy) in the GJP. In both cases, the inclusion of these variables significantly increased the fit of linear regression models designed to predict forecasting accuracy — both (a) with respect to a baseline model that included all non-focal predictors; and (b) with respect to each other. Taken together, these findings suggest that both environmental enrichment variables and bootstrapped credibility estimates are useful predictors of

forecasting accuracy in the GJP, and that neither set of variables was redundant with (or subsumed) by the other.

Critically, however, the results of the likelihood-ratio tests reported in Table 12 indicate that the incremental validity of environmental enrichment variables was considerably smaller than that of bootstrapped credibility estimates when added to the complementary intermediary model. From a qualitative perspective, this result suggests that most (but not all) of the explanatory power provided by environmental enrichment variables can also be provided by bootstrapped credibility estimates. This inference is also supported by disparate changes in multiple  $R^2$  and  $BIC$  when moving from each of the intermediary models to the *full* model. When bootstrapped credibility estimates were added to the *environmental only* model, the multiple  $R^2$  of the *full* model increased by 0.30 and its  $BIC$  decreased by 502.89 — the latter of which is more than 50-times larger than the conventional threshold for a “very large” improvement (Raftery, 1995). In contrast, when environmental enrichment variables were added to the *credibility only* model, the multiple  $R^2$  of the *full* model only increased by 0.01 and its  $BIC$  increased (i.e., got worse) by 1.25 to account for the model’s reduced degree of parsimony. Thus, while likelihood-ratio tests indicate that environmental enrichment variables can account for unique variance in forecast accuracy in GJP data, a closer examination suggests that the practical value of this effect is small.

## General Discussion

Despite the simplicity of the linear credibility framework, the results of Study 2a demonstrate that linear regression can provide reliable and valid estimates of credibility when fit to empirical data. Indeed, despite its statistical simplicity, the linear credibility framework provided informative estimates of “skill” or “proficiency” in subjective probability judgment at nearly every juncture of Study 2a. In Analysis 2a.i (GJP reliability), repeated trials of the credibility estimation procedure demonstrated that it is possible to derive reliable credibility estimates (both bootstrapped and non-bootstrapped) using a calibration sample-size of 50 judgments and as few as 10 bootstrap trials — both of which are small enough to have made the linear credibility framework a viable tool for researchers in the GJP. In Analysis 2a.ii (GJP validity), exploratory regressions demonstrated that bootstrapped credibility estimates (a) can be used to predict forecasting accuracy; (b) covary with a wide variety of measures related to “skill” or “proficiency” in the GJP; and (c) generally account for a large proportion of explanatory variance in models of forecast accuracy. Finally, in Analysis 2a.iii (GJP enrichment vs. credibility), likelihood-ratio tests demonstrated that bootstrapped credibility estimates are more strongly predictive of forecasting accuracy than the GJP’s environmental enrichment variables — very nearly to the point of making them redundant. Thus, it is evident from the results of Study 2a that linear credibility estimates can provide a great deal of insight into who is likely to be a “better” vs. “worse” forecaster in the GJP.

Critically, however, there was one area of Study 2a where the linear credibility framework did not perform as expected. In Analysis 2a.ii (GJP validity), exploratory linear regressions did not reveal a strong network of convergent validity between

individual difference measures and bootstrapped estimates of *expertise* ( $\hat{\beta}'_i$ ). Fortunately, a qualitative examination of the data suggests that convergent validity may have failed to manifest in this case because GJP forecasters were relatively homogenous in terms of *expertise* (see also: Footnote 20). Despite the ease with which this anomaly can be explained, however, its presence raises an important question. To what extent are the results of Study 2a a function of the GJP's uncharacteristically rich data-set? Furthermore, given that one of the main selling-points of the GJP is that its forecasters were exceedingly accurate, how likely are the results of Study 2a to generalize to less extraordinary (and perhaps less-well studied) populations? To address these questions, Study 2b examined the applicability of the linear credibility framework to a more ordinary sample of forecasters.

### **Study 2b: Reliability and Validity of Credibility Estimates Derived from March Madness Data (MM Reliability/Validity)**

In Study 2a (GJP reliability/validity), I demonstrated that the linear credibility framework can provide reliable and valid predictors of “skill” or “proficiency” in subjective probability judgment under ideal conditions. To be useful to decision makers, however — and, indeed, to represent a meaningful contribution to decision science — the linear credibility framework must be informative across a variety of domains. In practice, therefore, the results of Study 2a represent a rather weak (and perhaps non-representative) test of the linear credibility framework as a tool for identifying “better” vs. “worse” judges of subjective probability.

To remedy this problem, Study 2b examined the reliability and validity of credibility estimates derived from the 2017 NCAA Division-I Men’s Basketball Championship, or what is commonly known as the 2017 “March Madness” tournament. Unlike forecasts in the GJP, predictions in Study 2b were provided by novice forecasters who (a) had only minimal (and/or preexisting) training in subjective probability judgment (for the full text of the instructions provided to participants, see: Appendix B); and (b) did not have the benefit of working in collaborative teams. In addition, forecasters in Study 2b had less than a month to learn from their mistakes *in vivo* and were privy to only the most basic feedback about the quality of their predictions (i.e., which team won in a one-off event). Finally, and perhaps most critically, forecasters in Study 2b were constrained to a maximum of 67 judgments, thereby limiting the power of any credibility assessment to a rather stringent, real-world scope.

Based on these differences, the purpose of Study 2b was to provide a more representative test of the linear credibility framework. Strictly speaking, however, Study 2b also contained several features that were out of the ordinary. First, Study 2b used a sample that was even more “convenient” than the typical, behavioral sciences baseline (i.e., affluent, Western college students). Specifically, by targeting recruitment efforts at student organizations such as sports teams and fraternities, it is likely that participants in Study 2b were individuals who were particularly interested in the March Madness tournament. On the one hand, this self-selection bias might increase the validity of Study 2b’s results because the participants who completed the study were (a) especially likely to be engaged with the task; and (b) the same set of individuals who might use credibility

information to improve their March Madness predictions in future. On the other hand, the intrinsic motivation of participants in Study 2b might also have reduced its validity, in that preexisting loyalties to teams, conferences, and/or geographic regions might have led to biased reasoning about outcomes (thereby undermining the validity of crowdsourced beliefs).

Perhaps more importantly, however, Study 2b may have been unrepresentative in that it asked participants to provide SPJs about extremely uncertain events. Indeed, according to the forecasting blog FiveThirtyEight, a historical analysis of the 11.6 million tournament brackets registered on espn.com in 2015 reveals that only 273 individuals maintained a perfect prediction record beyond the tournament's *first day* (Paine & Boice, 2017, March 14). Practically speaking, therefore, it is evident that outcomes in the March Madness tournament are extremely difficult to predict. Even when pooling information across individuals, therefore, it is likely that the predictive "signal" associated with *estimated optima* in Study 2b was relatively weak. Thus, while Study 2b is more representative than the Good Judgment Project in its scope and design, it may also be unrepresentative as a test of the linear credibility framework in that it sets an exceedingly high bar.

### **Analysis 2b.i: Under what conditions are credibility estimates reliable? (MM reliability)**

To apply the linear credibility framework to March Madness data, I once again had to select three analytic parameters: a calibration sample size ( $n_{cal}$ ), a minimum

prediction sample size ( $n_{\text{pred}}$ ), and a number of bootstrap trials ( $n_{\text{boot}}$ ). Because forecasters were limited to a maximum of 67 predictions in the March Madness tournament, the minimum prediction sample size ( $n_{\text{pred}}$ ) for Study 2b was set to an *a priori* value of 1 to maximize statistical power. To select appropriate values for the other two parameters, I once again conducted an experiment. Similar to Analysis 2a.i (GJP reliability), the purpose of this experiment was to examine the empirical parameter ranges under which reliable credibility estimates could be extracted from the March Madness data.

### **Method.**

**Detailed Procedure.** To identify an appropriate set of analytic parameters for Study 2b, Analysis 2b.i employed the same two-armed design as Analysis 2a.i (GJP reliability). In the first arm, I examined the sensitivity of non-bootstrapped credibility estimates to five levels of calibration sample size,  $n_{\text{cal}} = \{10, 20, \dots, 50\}$ , with reliability in each cell estimated across 30 bootstrap trials. In the second arm, I examined the sensitivity of bootstrapped credibility estimates to changes in calibration sample size and number of bootstrap trials according to a  $5 \times 12$  design:  $n_{\text{cal}} = \{10, 20, \dots, 50\} \times n_{\text{boot}} = \{10, 20, \dots, 100, 200, 250\}$ , with reliability in each cell estimated across three runs of the General Procedure. Similar to Analysis 2a.i (GJP reliability), the purpose of the first arm of this experiment was to identify the minimum calibration sample size at which the effects of recalibration were likely to be consistent, and the purpose of the second arm was to ensure that bootstrapped credibility estimates would be appropriate for later analyses of validity.

**Results.** The results on Analysis 2b.i can be seen in Figures 10-13, below. In Figure 10, the reliability of non-bootstrapped estimates of credibility are once again graphed in a single plot, where the y-axis represents reliability (i.e., intraclass correlation, or ICC); the x-axis represents calibration sample size ( $n_{cal}$ ); and separate curves represent different components of credibility. In Figures 11-13, the reliability of bootstrapped estimates of *bias*, *expertise*, and *consistency* are graphed in separate plots. In each of these plots, the y-axis represents reliability (ICC); the x-axis represents the number of bootstrap trials ( $n_{boot}$ ); and separate curves represent different calibration sample sizes ( $n_{cal}$ ).

Figure 10

[MM data]: Reliability of non-bootstrapped estimates of credibility, varying by calibration sample size ( $n_{cal}$ ).

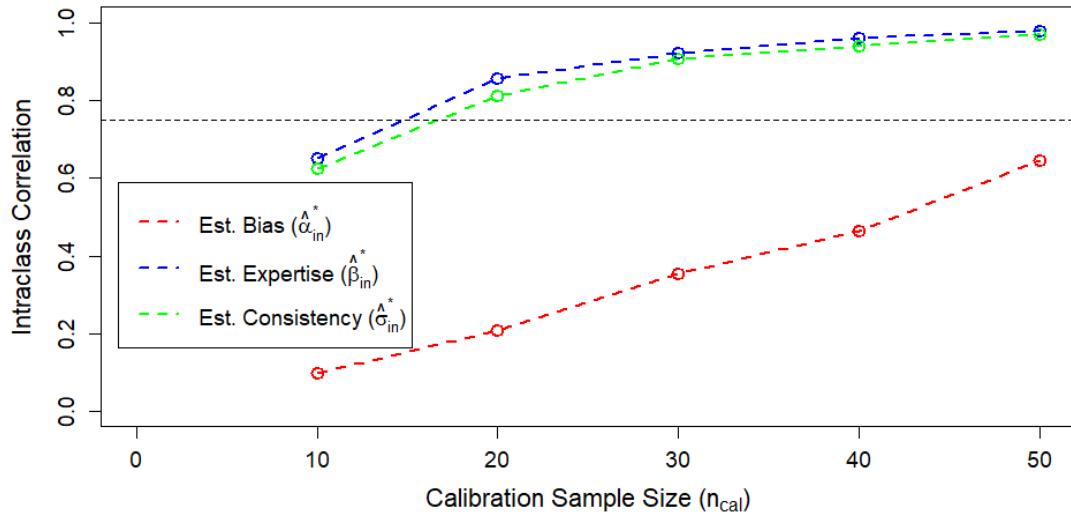


Figure 11

[MM data]: Reliability of bootstrapped estimates of bias ( $\hat{\alpha}_i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ).

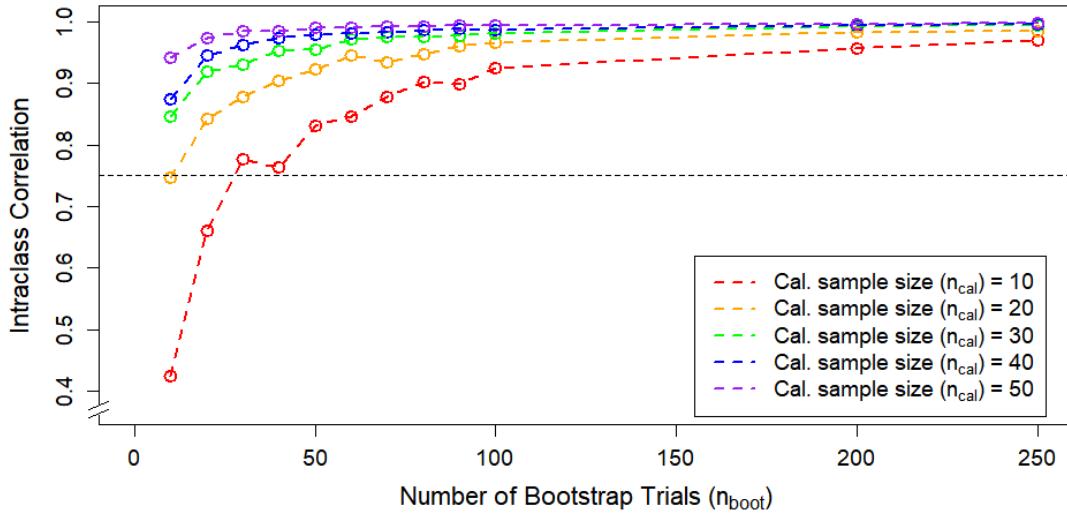


Figure 12

[MM data]: Reliability of bootstrapped estimates of expertise ( $\hat{\beta}_i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ).

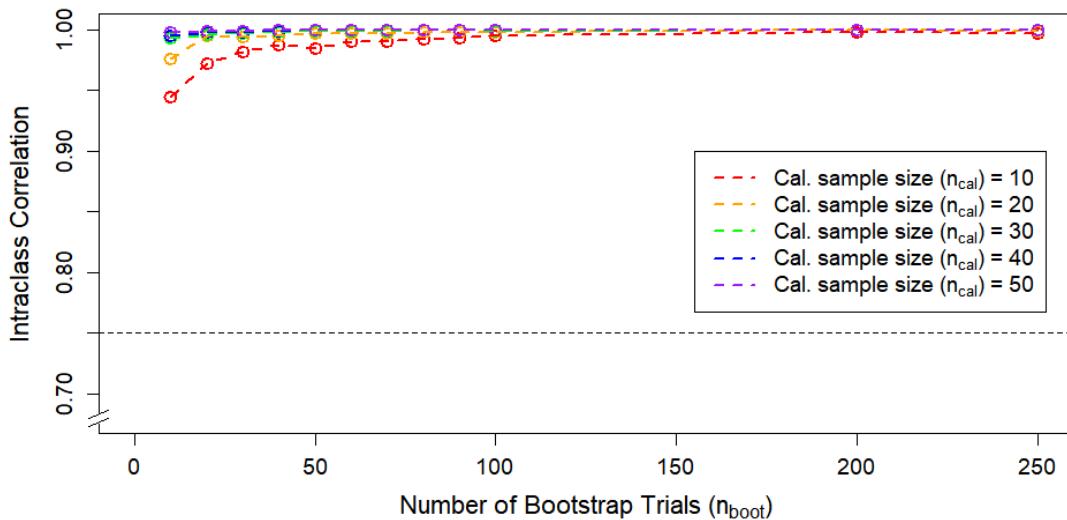
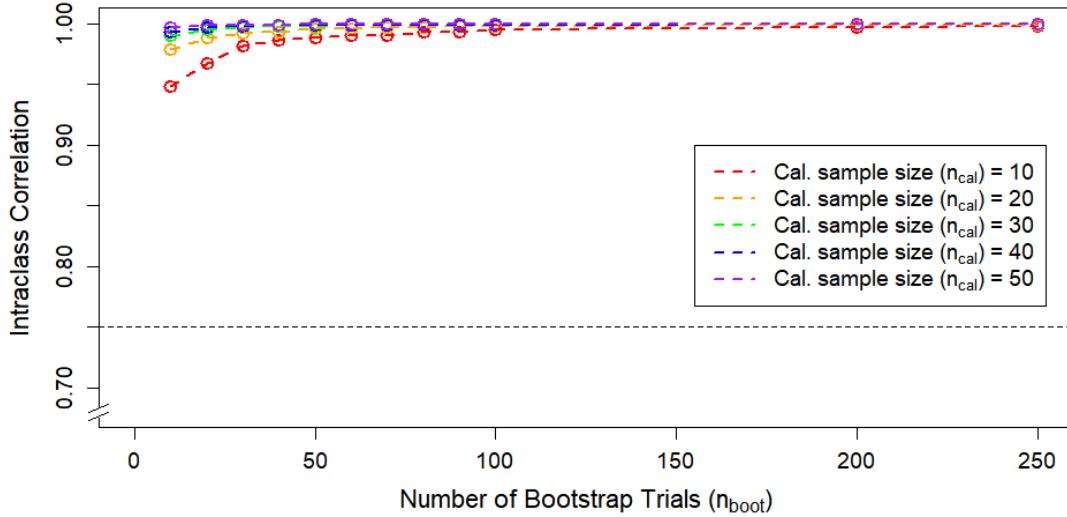


Figure 13

[MM data]: Reliability of bootstrapped estimates of consistency ( $\hat{\sigma}_i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ).



**Discussion.** Despite the high bar set by the March Madness study, the results of Analysis 2b.i indicate that the linear credibility framework was able to extract reliable credibility estimates from the available data. Indeed, the results of the first arm of Analysis 2b.i suggest that non-bootstrapped estimates of *expertise* and *consistency* exceeded the conventional threshold for “excellent” reliability (ICC  $\geq 0.75$ ; Ciccetti, 1994) at a smaller calibration sample-size than in the GJP ( $n_{cal} = 20$  vs. 50). Though the same cannot be said for non-bootstrapped estimates of *bias*, even these reached a conventionally “good” level of reliability ( $0.60 \leq \text{ICC} \leq 0.74$ ; Ciccetti, 1994), suggesting that errors and biases were no less prevalent in the March Madness data than elsewhere — even if crowd aggregates were generally less accurate.

In addition, Analysis 2b.i demonstrated that highly reliable credibility estimates could be extracted from the March Madness data by employing a nominal degree of

bootstrapping. Indeed, in all observed cases (Figures 12 and 13), intraclass correlations never fell below a value of 0.93 — a level that is very high for empirical data. As with non-bootstrapped estimates, bootstrapped estimates of *bias* (Figure 11) were considerably less reliable than the other two components of credibility. Even here, however, nearly all observations fell above the 0.75 threshold for “excellent” intraclass correlation, and in cases where they didn’t, calibration sample size was never more than  $n_{cal} = 20$ .

Based on these findings, the analytic parameters selected for Study 2b were: a calibration sample size of  $n_{cal} = 50$ , to maximize the likelihood of consistent recalibration while still leaving a non-trivial number of predictions out-of-sample; a minimum prediction sample-size of  $n_{pred} = 1$  to maximize statistical power; and — in the absence of strong concerns about the reliability of bootstrapped credibility estimates —  $n_{boot} = 100$  bootstrap trials to ensure a sufficient number of trials over which to examine the effects of recalibration.

### **Analysis 2b.ii: What are the predictors of credibility and what does credibility predict? (MM validity)**

In Analysis 2a.ii (GJP validity), exploratory linear regressions revealed a rich network of convergent validity between bootstrapped credibility estimates and predictors of accuracy in the GJP. Though far from providing a face-valid readout of each forecaster’s *bias*, *expertise*, and *consistency*, these results suggest that linear credibility estimates can help decision makers learn a great deal about the “quality” of an

individual's SPJs — especially if their goal is to predict forecast accuracy (i.e., average Brier scores).

Practically speaking, however, convergent validity with individual difference measures in the GJP represents a low bar for the linear credibility framework. Prior to the current research, Mellers et al. (2015a; 2015b; 2014) had already demonstrated that GJP forecasters (a) could produce remarkably accurate predictions about the outcomes of geopolitical events; (b) varied widely in their forecast accuracy; (c) generally tended to have greater accuracy when trained in probabilistic reasoning and/or working in groups; and (d) tended to become more accurate over time. Based on these results, Mellers et al. concluded that geopolitical forecasting is a domain where “skill” or “proficiency” can be actively cultivated, and that doing so can have a direct impact on forecast accuracy. Thus, Mellers et al. had already provided a strong basis for assuming the forecasters in the GJP varied in terms of credibility, and that these differences were at least partially responsible for differences in Brier scores. In Analysis 2a.ii (GJP validity), therefore, the only open question was whether linear credibility estimates could *capture* these differences.

In the March Madness data, by contrast, the conceptual coupling between accurate forecasting and credible forecasting was considerably weaker. Strictly speaking, the assumptions of the linear credibility framework were still the same: *if* there is predictive signal in the March Madness data and *if* one can use Baron et al.'s (2014) method to amplify it and *if* the resulting *estimated optima* tend to be more accurate than the beliefs held by most individuals and *if* an individual's SPJs tend to depart from *estimated optima* in systematic ways and *if* those departures can be meaningfully

captured by linear regression, then the linear credibility framework *should* yield estimates that are strongly predictive of forecast accuracy. However, because the March Madness data were more likely to violate these assumptions than forecasts from the GJP (e.g., participants likely had less domain knowledge; March Madness predictions likely yielded weaker predictive signal; team loyalties likely made some participants' judgement strategies less reliable), it was an empirical question whether the rich predictive validities observed in Analysis 2a.ii (GJP validity) would replicate.

To determine if they do, Analysis 2b.ii examined the validity of bootstrapped credibility estimates derived from the 2017 March Madness data. In an ideal scenario, the results of this analysis would reveal a rich network of convergent validity between credible forecasting and “good” forecasting, as defined by forecast accuracy (i.e., average Brier scores). Even if they don’t, however, the observation of any non-trivial predictive validity in Analysis 2b.ii would suggest that decision makers might still benefit from examining credibility. Regardless of whether linear regression is a “good” model of credibility, that is, the existence of predictive validity in Analysis 2b.ii would suggest that it is not a bankrupt one. Thus, as long as bootstrapped credibility estimates are not orthogonal to forecast accuracy, the March Madness data must contain some degree of information about the “quality” or relative validity of an individual’s judgments. If this turns out to be true, then even weak results in Analysis 2b.ii would suggest that decision

makers might benefit from examining credibility, as there are very few costs associated with probing this information and nothing to gain by leaving it on the table.<sup>22</sup>

## **Method.**

**Detailed Procedure.** Using the analytic parameters selected in Analysis 2b.i (MM reliability) (i.e.,  $n_{cal} = 50$ ;  $n_{pred} = 1$ ;  $n_{boot} = 100$ ), I used the General Procedure to arrive at bootstrapped credibility estimates for each forecaster in the March Madness study (i.e.,  $\hat{\alpha}_i$ ,  $\hat{\beta}_i$ , and  $\hat{\sigma}_i$ ). After excluding participants who provided too few forecasts (< 51), this data-set included a working sample of 118 participants. Among these participants, the mean age was 24.4 ( $Mdn. = 23.0$ ;  $SD = 5.7$ ), and 43% of participants self-identified as having completed “some undergraduate” education (0% reported no college education; 23% reported an associate’s or bachelor’s degree; 15% reported “some post-baccalaureate” education; and 20% reported an advanced degree). In addition, 48% of participants self-identified as female (50% male; 1% other/neither) and 50% self-identified as “white or Caucasian” (19% black or African American; 13% East Asian; 8% Hispanic or Latinx; 4% South Asian; 5% Other; 1% Middle Eastern; 0% Native American; 0% Pacific Islander; 0% Other Indigenous People).

---

<sup>22</sup> In practice, of course, the low-cost of credibility information does not mean that there is no risk in using it. As such, any decision maker interested in using credibility information to recalibrate SPJs (and by extension, influence decisions) should carefully consider the reliability and validity of credibility estimates before putting them into action. Fortunately, the research presented in this dissertation provides a preliminary set of analytic tools for doing just that.

After estimating credibility, I transformed estimates of *bias* and *expertise* to absolute differences from the normative values implied by identity (i.e., I calculated  $\hat{\alpha}'_i$  and  $\hat{\beta}'_i$  by taking the absolute difference of the untransformed values from 0 and 1, respectively). Then, using parallel procedures to those used in Analysis 2a.ii (GJP validity) (i.e., all continuous variables standardized; all categorical variables effects-coded), I calculated pairwise correlations and conducted exploratory linear regressions to examine the relationships among bootstrapped credibility estimates ( $\hat{\alpha}'_i$ ,  $\hat{\beta}'_i$ , and  $\hat{\sigma}_i$ ) and the individual difference measures captured in the March Madness study.

In the interest of replicating Analysis 2a.ii (GJP validity) as closely as possible, the March Madness study administered a similar set of individual difference measures as those captured by the GJP. Once again, the principal variable of interest was an individual's forecast accuracy (*average Brier score*), but exploratory analyses were conducted across a broad range of variables. Specifically, in the tables below, *Berlin Numeracy* reflects the number of questions a participant answered correctly on the Berlin Numeracy Test (Cokely et al., 2012). *CRT* reflects a forecaster's score on the extended Cognitive Reflection Task (Baron et al., 2015; see also: Frederick, 2005). *Need for Cognition* reflects a participant's score on the Need for Cognition questionnaire (Cacioppo & Petty, 1982). *One question Fox-Hedgehog* reflects a participant's response to a single-item scale measuring the extent to which an individual's approach to problem solving tends to rely on knowing "one big thing" rather than "many little things" (for additional details on this measure, see: Tetlock, 2005). And *AOMT* reflects a participant's score on Baron's 11-item Actively Open-Minded Thinking scale (*in press*).

In addition to these measures, the March Madness study also included measures of *Working memory*, which reflects the number of items a participant correctly recalled after memorizing six verbal cues under high cognitive load (here, timed arithmetic problems; for an overview of working memory, see Baddeley & Hitch, 1974); and a set of self-report *domain knowledge* questions that asked participants to rate their agreement with statements claiming “extensive expertise” in the areas of “college basketball,” “basketball, in general,” “probabilistic prediction,” and “another type of prediction or prediction, in general” (7-point Likert, anchored at “strongly disagree” and “strongly agree”). As in the GJP, participants also provided demographic information about their age, race, gender, and level of education, though the March Madness study offered a wider range of response options in each of the latter three categories.

Because this set of variables once again presented a non-trivial likelihood of redundancy, multi-collinearity, and/or suppression effects, exploratory regressions in Analysis 2b.ii were conducted under three approaches to variable selection: the *kitchen sink* approach, the *reduced* approach, and the *ridge-ISE* approach (for additional details, see the Detailed Procedure section of Analysis 2a.ii (GJP validity)). By comparing these three approaches, I was once again able to observe instances where variables were competing for overlapping segments of explanatory variance.

**Results.** The results on Analysis 2b.ii can be seen in Tables 13-17, below. Simple correlations between bootstrapped credibility estimates ( $\hat{\alpha}'_i$ ,  $\hat{\beta}'_i$ , and  $\hat{\sigma}_i$ ) and individual difference measures can be seen in Table 13; exploratory regression analyses examining

the predictors of credibility can be seen Tables 14-16; and an exploratory regression analysis concerning the predictors of forecast accuracy (i.e., average Brier scores) can be seen in Table 17.

Table 13

*[MM data]: Simple correlations between credibility estimates and individual difference measures.*

Ind. Diff. Measure	Credibility Estimate		
	$\hat{\alpha}'_i$ (Bias)	$\hat{\beta}'_i$ (Expertise)	$\hat{\sigma}_i$ (Consist.)
<i>Forecast accuracy</i>			
Average Brier score	0.13	0.27**	0.76***
<i>Numerical fluency</i>			
Berlin Numeracy score	-0.17	-0.13	-0.53***
<i>Cognitive ability</i>			
CRT	-0.15	-0.03	-0.29**
Working memory	0.07	-0.09	-0.30***
<i>Cognitive style</i>			
Need for Cognition	-0.14	-0.10	-0.17
One question Fox-Hedgehog	-0.11	0.01	0.00
AOMT	-0.20*	-0.08	-0.26**
<i>Domain knowledge (self-report)</i>			
College basketball	-0.06	-0.03	-0.30***
Basketball (general)	-0.02	-0.10	-0.28**
Probabilistic prediction	-0.07	-0.12	-0.24*
Prediction (general)	-0.01	-0.13	-0.15
<i>Demographics</i>			
Age	0.11	0.00	0.08
Gender = male	-0.01	-0.13	-0.35***

Gender = female	0.01	0.13	0.35***
Race = black/African American	0.10	-0.04	0.28**
Race = East Asian	0.01	0.01	-0.08
Race = Hispanic/Latinx	0.17	-0.01	0.09
Race = Middle Eastern	0.01	0.02	0.03
Race = South Asian	0.02	-0.02	0.06
Race = white/Caucasian	-0.05	-0.02	-0.28**
Education = some undergrad.	0.08	-0.15	0.01
Education = bachelor's degree	-0.09	0.12	-0.07
Education = master's degree	0.05	0.02	0.06
Education = doctorate	0.04	0.05	-0.07

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

Table 14

*[MM data]: Predictors of bootstrapped alpha (i.e., bias;  $\hat{\alpha}'_i$ ).*

Ind. Diff. Measure	Variable Selection Approach		
	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	-0.04 (0.51)	0.00 (0.09)	0.00 (0.09)
<i>Forecast accuracy</i>			
Average Brier score	0.12 (0.12)		
<i>Numerical fluency</i>			
Berlin Numeracy	-0.16 (0.15)		
<i>Cognitive ability</i>			
CRT	0.07 (0.15)		
Working memory	0.24 (0.11)*	0.07 (0.09)	
<i>Cognitive style</i>			
Need for Cognition	-0.11 (0.14)		
One question Fox-Hedgehog	-0.05 (0.11)		
AOMT	-0.11 (0.16)		

*Domain knowledge (self-report)*

College basketball	-0.06 (0.17)
Basketball (general)	0.04 (0.18)
Probabilistic prediction	-0.10 (0.22)
Prediction (general)	0.11 (0.21)

*Demographics*

Age	0.24 (0.15)
Gender = male	-0.13 (0.48)
Gender = female	-0.35 (0.46)
Race = Black/African American	0.15 (0.32)
Race = East Asian	0.14 (0.32)
Race = Hispanic/Latinx	0.63 (0.37)
Race = Middle Eastern	-1.07 (0.94)
Race = South Asian	0.11 (0.47)
Race = White/Caucasian	0.14 (0.24)
Education = some undergrad.	0.34 (0.28)
Education = bachelor's degree	-0.01 (0.28)
Education = some post-bac.	0.02 (0.34)
Education = master's degree	-0.04 (0.33)
Education = doctorate	-0.36 (0.46)

Multiple R <sup>2</sup>	0.18	0.005	0
Adjusted R <sup>2</sup>	-0.043	-0.004	0
RMSE	0.902	0.993	0.996
AIC	364.44	339.33	337.87
BIC	439.25	347.64	343.41

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

Table 15

[MM data]: Predictors of bootstrapped beta (i.e., expertise;  $\hat{\beta}'_i$ ).

Variable Selection Approach

Ind. Diff. Measure	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	0.34 (0.51)	0.00 (0.09)	0.00 (0.09)
<i>Forecast accuracy</i>			
Average Brier score	0.3 (0.12)*	0.26 (0.09)**	
<i>Numerical fluency</i>			
Berlin Numeracy	-0.05 (0.15)		
<i>Cognitive ability</i>			
CRT	0.16 (0.15)		
Working memory	-0.05 (0.11)		
<i>Cognitive style</i>			
Need for Cognition	-0.08 (0.14)		
One question Fox-Hedgehog	-0.03 (0.11)		
AOMT	-0.02 (0.16)		
<i>Domain knowledge (self-report)</i>			
College basketball	0.19 (0.17)		
Basketball (general)	-0.05 (0.18)		
Probabilistic prediction	-0.02 (0.22)		
Prediction (general)	-0.08 (0.21)		
<i>Demographics</i>			
Age	-0.10 (0.15)		
Gender = male	-0.32 (0.49)		
Gender = female	-0.14 (0.47)		
Race = Black/African American	-0.13 (0.32)		
Race = East Asian	0.07 (0.32)		
Race = Hispanic/Latinx	-0.09 (0.38)		
Race = Middle Eastern	0.3 (0.95)		
Race = South Asian	0.11 (0.48)		
Race = White/Caucasian	0.02 (0.24)		
Education = some undergrad.	-0.34 (0.29)		
Education = bachelor's degree	0.19 (0.28)		
Education = some post-bac.	0.02 (0.35)		

Education = master's degree	0.08 (0.34)		
Education = doctorate	-0.13 (0.46)		
Multiple R <sup>2</sup>	0.166	0.07	0
Adjusted R <sup>2</sup>	-0.061	0.062	0
RMSE	0.91	0.96	0.996
AIC	366.5	331.27	337.87
BIC	441.31	339.59	343.41

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

Table 16

[MM data]: Predictors of bootstrapped sigma (i.e., consistency;  $\hat{\sigma}_i$ ).

Ind. Diff. Measure	Variable Selection Approach		
	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	0.26 (0.28)	0.04 (0.06)	0.00 (0.06)
<i>Forecast accuracy</i>			
Average Brier score	0.65 (0.07)***	0.69 (0.06)***	0.65 (0.06)***
<i>Numerical fluency</i>			
Berlin Numeracy	-0.30 (0.08)***	-0.32 (0.07)***	-0.29 (0.06)***
<i>Cognitive ability</i>			
CRT	0.23 (0.08)**	0.12 (0.07)	
Working memory	-0.12 (0.06)*	-0.13 (0.06)*	
<i>Cognitive style</i>			
Need for Cognition	-0.10 (0.08)		
One question Fox-Hedgehog	0.03 (0.06)		
AOMT	0.01 (0.09)		
<i>Domain knowledge (self-report)</i>			
College basketball	0.08 (0.1)		
Basketball (general)	-0.08 (0.1)		
Probabilistic prediction	-0.17 (0.12)		

Prediction (general)	0.24 (0.11)*	0.11 (0.06)
<i>Demographics</i>		
Age	0.16 (0.08)*	0.04 (0.06)
Gender = male	-0.17 (0.27)	
Gender = female	-0.01 (0.26)	
Race = Black/African American	0.11 (0.18)	
Race = East Asian	-0.13 (0.18)	
Race = Hispanic/Latinx	-0.01 (0.21)	
Race = Middle Eastern	0.06 (0.52)	
Race = South Asian	0.30 (0.26)	
Race = White/Caucasian	-0.15 (0.13)	
Education = some undergrad.	0.01 (0.16)	
Education = bachelor's degree	-0.20 (0.16)	
Education = some post-bac.	0.03 (0.19)	
Education = master's degree	-0.42 (0.18)*	-0.30 (0.16)
Education = doctorate	-0.38 (0.25)	
Multiple R <sup>2</sup>	0.749	0.693
Adjusted R <sup>2</sup>	0.681	0.673
RMSE	0.499	0.552
AIC	224.8	212.56
BIC	299.61	237.49
		0.648
		0.642
		0.59
		218.51
		229.59

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

Table 17

*[MM data]: Predictors of average Brier score (i.e., forecast accuracy).*

Ind. Diff. Measure	Variable Selection Approach		
	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	-0.37 (0.3)	-0.04 (0.06)	0.00 (0.06)
<i>Credibility</i>			
Estimated bias ( $\hat{\alpha}'_i$ )	-0.03 (0.06)		

Estimated expertise ( $\hat{\beta}'_i$ )	0.17 (0.06)**	0.15 (0.06)*	
Estimated consistency ( $\hat{\sigma}_i$ )	0.78 (0.08)***	0.72 (0.06)***	0.76 (0.06)***
<i>Numerical fluency</i>			
Berlin Numeracy	0.14 (0.09)		
<i>Cognitive ability</i>			
CRT	-0.25 (0.09)**	-0.04 (0.06)	
Working memory	0.08 (0.07)		
<i>Cognitive style</i>			
Need for Cognition	0.11 (0.08)		
One question Fox-Hedgehog	-0.02 (0.06)		
AOMT	-0.05 (0.09)		
<i>Domain knowledge (self-report)</i>			
College basketball	-0.18 (0.1)		
Basketball (general)	0.02 (0.1)		
Probabilistic prediction	0.22 (0.13)		
Prediction (general)	-0.25 (0.12)*	-0.13 (0.06)*	
<i>Demographics</i>			
Age	-0.10 (0.09)		
Gender = male	0.31 (0.28)		
Gender = female	0.27 (0.27)		
Race = Black/African American	-0.03 (0.19)		
Race = East Asian	0.04 (0.19)		
Race = Hispanic/Latinx	-0.05 (0.22)		
Race = Middle Eastern	-0.19 (0.56)		
Race = South Asian	-0.24 (0.28)		
Race = White/Caucasian	-0.03 (0.14)		
Education = some undergrad.	-0.01 (0.17)		
Education = bachelor's degree	0.09 (0.17)		
Education = some post-bac.	-0.04 (0.2)		
Education = master's degree	0.49 (0.2)*	0.32 (0.16)*	
Education = doctorate	0.46 (0.27)		

Multiple R <sup>2</sup>	0.721	0.636	0.577
Adjusted R <sup>2</sup>	0.637	0.62	0.573
RMSE	0.526	0.6	0.648
AIC	241.2	228.45	238.4
BIC	321.55	247.85	246.71

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: missingness in individual difference data resolved through mean imputation. All non-categorical measures standardized; all categorical variables effects-coded.

**Discussion.** From a purely descriptive perspective, the results of Analysis 2b.ii tell a mixed story about the validity of bootstrapped credibility estimates derived from the March Madness data. As can be seen in Table 13, simple correlations reveal (a) a strong positive relationship between bootstrapped estimates of *consistency* ( $\hat{\sigma}_i$ ) and forecast accuracy (average Brier score); and (b) a rich network of convergent validity with other relevant variables (the only effects with unexpected signs and/or magnitudes are: gender; and race = white/Caucasian). In this same table, however, bootstrapped estimates of *expertise* ( $\hat{\beta}'_i$ ) only demonstrate a small-to-moderate relationship with forecast accuracy and bootstrapped estimates of *bias* ( $\hat{\alpha}'_i$ ) don't show any significant relationship with Brier score, whatsoever. Somewhat surprisingly, bootstrapped estimates of *bias* and *expertise* also don't seem to share any meaningful relationships with the other variables measured in the March Madness Study, beyond generally tending to have the expected signs.

As shown in Tables 14-16, exploratory linear regressions tell a similar story. In Table 16, bootstrapped estimates of *consistency* demonstrate a strong, predictive relationship with forecast accuracy and show a small (yet promising) degree of convergent validity with numeracy and working memory (the former of which can be

seen in the *reduced* model but does not make the cut for the statistically stringent *ridge-ISE* model). By contrast, bootstrapped estimates of *expertise* demonstrate only a moderate degree of convergent validity with forecast accuracy (Table 15), and bootstrapped estimates of *bias* are predicted by nothing except working memory (Table 14). Notably, however, this singular effect is (a) fairly weak; (b) only observed in the atheoretical *kitchen sink* model; and (c) not in the expected direction. Thus, the results of Analysis 2b.ii indicate that *consistency* and *expertise* are the only components of credibility that could have provided any insight into “skill” or “proficiency” in the March Madness study, and that the latter’s contribution was quite small — a conclusion that is corroborated by Table 17.

Based on these results, it is tempting to conclude that the linear credibility framework was not well suited to capturing an individual’s “true” degree of credibility in Study 2b. When considered in the context of the March Madness data, however, there are at least two reasons why this conclusion may be overstated. First, bootstrapped estimates of *consistency* ( $\hat{\sigma}_i$ ) were able to explain a large proportion of the variance in forecast accuracy. Indeed, even though bootstrapped measures of *bias* ( $\hat{\alpha}'_i$ ) and *expertise* ( $\hat{\beta}'_i$ ) didn’t add much to the story, the results presented in Table 17 indicate that *consistency* alone was able to account for 57% of the observed variance in average Brier scores. In practice, therefore, this result suggests that the construct-validity of *bias*, *expertise*, and *consistency* may be rather weak (i.e., the linear credibility framework’s three-component model of credibility may not have been empirically justified). From the perspective of the “big picture,” however, the overall quality of the information provided by the linear

credibility framework was strong. Thus, it is an overstatement to conclude that the linear credibility framework “was not informative” in Study 2b.

Perhaps more importantly, however, the second reason to be circumspect about the results of Analysis 2b.ii is that there are two plausible explanations for *bias* and *expertise*’s weak predictive validity. The first is that the linear credibility framework did not provide an accurate model of the latent construct of credibility, suggesting that the model, the construct, or both were poorly specified. The second is that the March Madness data did not contain much information about “skillful” forecasting to begin with, suggesting that there was little signal for the linear credibility framework to amplify. Because Analysis 2b.ii cannot disambiguate between these two explanations, it is premature to conclude that the linear credibility framework was “ill suited” to its task in Study 2b.

In the interest of disentangling these two explanations, however, I conducted several post-hoc analyses. In general, these analyses revealed a relatively weak degree of predictive signal in the March Madness data — suggesting, perhaps, that the weak results of Analysis 2b.ii can be attributed to the data, rather than the linear credibility framework. Consider, for example, the *estimated optima* used in Study 2b (MM reliability/validity). Even though these aggregates were used as a stand-in for “optimal” judgments, average Brier scores indicate that they were less accurate than 16% of individual forecasters. In addition, these so-called “optima” were only slightly better than the average forecaster at predicting win/loss outcomes. If all probability estimates in the March Madness study were collapsed to binary (i.e., yes/no) predictions, that is, then the

predictions implied by *estimated optima* would have been correct 75% of the time, whereas the average forecaster was correct 66% of the time ( $Mdn. = 66\%$ ;  $SD = 6\%$ ). Though a Wilcoxon signed-rank test indicates that this difference is significant ( $V = 68$ ,  $p < 0.001$ ), the practical value of this improvement is debatable, given that 6% (7/118) of Study 2b's untrained, amateur forecasters provided more accurate predictions.

Though far from conclusive, these post-hoc analyses suggest that the linear credibility framework may have struggled to identify predictive signal in the March Madness data because that signal was weak to begin with. Critically, however, the results of Analysis 2b.ii indicate that — with the predictive signal that was available — bootstrapped estimates of *consistency* were able to account for 57% of the observed variance in forecast accuracy. Thus, even if the linear credibility framework was a poor statistical model in Study 2b, it was still an *informative* one, and could still be used to identify important predictors of who was likely to be a “better” vs. “worse” judge of subjective probability.

## General Discussion

From a statistical perspective, the results of Study 2b suggest that the linear credibility framework is an imperfect tool for examining “skill” or “proficiency” in subjective probability judgment. In Analysis 2b.i (MM reliability), both bootstrapped and non-bootstrapped credibility estimates demonstrated acceptable degrees of reliability, but in Analysis 2b.ii (MM validity), bootstrapped estimates failed to demonstrate consistent validity as predictors of forecast accuracy. Upon examining these result more closely, it

is worth noting that bootstrapped estimates of *expertise* ( $\hat{\beta}'_i$ ) and *consistency* ( $\hat{\sigma}_i$ ) showed non-trivial degrees of predictive validity, and were together able to explain somewhere between 57% and 62% of the observed variance in Brier scores (see also: Table 17; the discussion section of Analysis 2b.ii (MM validity)). Nevertheless, the overall implication of Study 2b was that the linear credibility framework was not a well-specified model of “skill” or “proficiency” in March Madness predictions.

Practically speaking, however, there is reason to believe that Study 2b may not have been a fair test of the linear credibility framework. Specifically, outcomes in the March Madness tournament were so uncertain that forecasters in Study 2b may not have differed meaningfully in terms of their “true,” or latent credibility. To put this in perspective, consider the odds of constructing a “perfect bracket” in the March Madness tournament (i.e., correctly predicting the outcomes of all 67 games, *ex ante*). From a combinatorial perspective, the probability of constructing a perfect bracket by chance is less than 1 in 9.2 quintillion, and the forecasting blog FiveThirtyEight suggests that even sophisticated methods of forecast aggregation can only shorten these odds to about 1 in one-or-two-billion (Paine & Boice, 2017, March 14). In Study 2b, participants had better odds than these, as the study’s design asked them to make predictions about only those games that were played (rather than attempting to divine the entire path of the tournament, *ex ante*). However, a simple extrapolation from the March Madness data suggests that the chances of predicting all 67 games correctly were still about 1 in 138

million — an outcome that would have required a forecaster’s yes/no predictions to be 5.67 standard deviations above the mean.<sup>23</sup>

Based on these calculations, it is unlikely that participants’ performance in Study 2b represented a (globally) wide range of “skill” or “proficiency” in subjective probability judgment. As a result, small differences in credibility may not have been useful for distinguishing “better” vs. “worse” forecasters in this data-set. Thus, the mixed results observed in Study 2b (and Analysis 2b.ii (MM validity), in particular) may be partially attributable to the inherent uncertainty of March Madness outcomes, rather than any failing of the linear credibility framework. With this possibility in mind, the results of Study 2b provide room for optimism. Despite the fact that (a) outcomes in the March Madness tournament were exceedingly difficult to predict; and (b) the linear credibility framework was only able to provide rudimentary information about “true” or latent credibility in these data, examining credibility still provided statistical traction on the question of who is likely to be a “better” vs. “worse” judge of subjective probability. Thus, even at the boundaries of the linear credibility framework’s validity, the results of Study 2b suggest that the predictive utility of credibility information was surprisingly robust — a promising result for decision makers.

---

<sup>23</sup> This extrapolation was conducted by modeling forecasters’ rates of correct binary prediction as a Normal distribution with a mean of 0.66 and a standard deviation of 0.06— the empirical mean and standard deviation observed in the March Madness data. Strictly speaking, of course, this is an inappropriate statistical model, as plausible values for correct prediction rates are bounded at 0 and 1. As a first order approximation, however, this model represents only a small abuse of statistical realism, as all implausible values are at least 5.67 standard deviations from the empirical mean. Thus, the inappropriately-long tails of the Normal distribution used in this analysis represent a negligible departure from a more appropriately specified probability density function.

## **Study 2c: Reliability and Validity of Credibility Estimates Derived from Philadelphia Air Temperature Data (PHL Reliability/Validity)**

Up to this point, I have examined two applications of the linear credibility framework to empirical data. In the first application (Study 2a; GJP reliability/validity), data from the Good Judgment Project (GJP) revealed strong evidence for the reliability and validity of linear credibility estimates. However, because GJP forecasters were exceptional in a variety of ways, a broader consideration of these findings suggests that the general performance of the linear credibility framework may often be somewhat weaker. In the second application (Study 2b; MM reliability/validity), predictions from the March Madness study revealed weak (or, at least inconsistent) evidence for the reliability and validity of linear credibility estimates. However, because outcomes in the March Madness tournament were exceedingly uncertain (and thus, differences in “true” credibility may have been small), a broader consideration of these findings suggests that the general performance of the linear credibility framework may often be somewhat stronger.

For reasons that should be apparent to the reader, neither of these studies is entirely satisfactory. In Study 2a, the richness of the GJP data-set provides a plausible explanation for the richness of the predictive information uncovered by the linear credibility framework. And in Study 2b, a lack of heterogeneity in “true” credibility provides a plausible explanation for why linear credibility estimates were inconsistent

predictors of forecast accuracy. In general, therefore, the results of both studies present a non-trivial chance that the observed relationships between linear credibility estimates and forecast accuracy were not mediated by the underlying construct of credibility (i.e., a forecaster's relative tendencies towards error and bias in judgment). Thus, in both Study 2a (GJP reliability/validity) and Study 2b (MM reliability/validity), it is conceivable that the strength of the relationship between “good” judgment and credible judgment is not diagnostic of the latent covariation between linear credibility estimates and an individual's tendencies towards error and bias in judgment. Or, in simpler terms, neither Study 2a nor Study 2b provides an ironclad test of whether linear credibility estimates were valid indicators of “true” credibility.

To address this issue, Study 2c examined the performance of the linear credibility framework when fit to predictions about Philadelphia air temperature — a domain where between-subjects differences in accuracy and estimated credibility were unlikely to be driven by anything other than an individual's tendencies towards error and bias in judgment (i.e., an individual's “true” degree of credibility). As described in the General Method section, Study 2c was designed to be administered to a large number of participants in a short period of time. As such, Study 2c was less comprehensive than Studies 2a and 2b in that it only examined the convergent validity of linear credibility estimates with a handful of measures related to “good” judgment. However, Study 2c represented an improvement over Studies 2a and 2b in several ways:

- (1) The “true” credibility of participants in Study 2c was likely to vary due to differences in (e.g.) domain knowledge; proficiency in probabilistic reasoning; and differential susceptibility to errors and biases in judgment (e.g., representativeness; anchoring and adjustment; baserate neglect; for an overview, see: Kahneman, Slovic, & Tversky, 1982).
- (2) Air temperatures are familiar enough that ordinary people can make reasonable predictions, but outcomes were neither so certain nor so uncertain that individuals did not vary in terms of accuracy.
- (3) And finally, participants in Study 2c were not expected to be extraordinary in any way. Thus, if the linear credibility framework can provide meaningful insight into “skill” or “proficiency” in subjective probability judgment in this study, then there is little reason to suspect that these results won’t generalize.

#### **Analysis 2c.i: Under what conditions are credibility estimates reliable? (PHL reliability)**

As with previous applications of the linear credibility framework, Study 2c required that I select three analytic parameters: a calibration sample size ( $n_{cal}$ ), a minimum prediction sample size ( $n_{pred}$ ), and a number of bootstrap trials ( $n_{boot}$ ). Because most participants in Study 2c provided a total of 120 SPJs, there was little concern that participants would be excluded for providing too-few judgments. Thus, the minimum

prediction sample size for Study 2c was set at an *a priori* value of  $n_{\text{pred}} = 30$  to ensure that each participant would have a sufficient number of predictions held out-of-sample. To select appropriate values for the other two parameters, I conducted a parallel experiment to those conducted in Analysis 2a.i (GJP reliability) and 2b.i (MM reliability). Thus, in Analysis 2c.i, I manipulated calibration sample size ( $n_{\text{cal}}$ ) and number of bootstrap trials ( $n_{\text{boot}}$ ) to determine the empirical parameter ranges under which reliable credibility estimates could be extracted from the Philadelphia air temperature data.

### **Method.**

**Detailed Procedure.** As with previous reliability experiments, Analysis 2c.i, employed a two-armed experimental design. In the first arm, I examined the sensitivity of non-bootstrapped credibility estimates to five levels of calibration sample size,  $n_{\text{cal}} = \{10, 20, \dots, 50\}$ , with reliability in each cell estimated across 30 bootstrap trials. In the second arm, I examined the sensitivity of bootstrapped credibility estimates to changes in calibration sample size and number of bootstrap trials according to a  $5 \times 12$  design:  $n_{\text{cal}} = \{10, 20, \dots, 50\} \times n_{\text{boot}} = \{10, 20, \dots, 100, 200, 250\}$ , with reliability in each cell estimated across three runs of the General Procedure. Similar to Analyses 2a.i (GJP reliability) and 2b.i (MM reliability), the purpose of the first arm was to identify the minimum calibration sample size at which the effects of recalibration were likely to be consistent across trials, and the purpose of the second arm was to ensure that bootstrapped credibility estimates would be appropriate for later examinations of validity.

**Results.** The results of Analysis 2c.i can be seen in Figures 14-17. In Figure 14, the reliability of non-bootstrapped estimates of credibility are graphed in a single plot, where the y-axis represents reliability (i.e., intraclass correlation, or ICC); the x-axis represents calibration sample size ( $n_{cal}$ ); and separate curves represent different components of credibility. In Figures 15-17, the reliability of bootstrapped estimates of *bias*, *expertise*, and *consistency* are graphed in separate plots. In each of these plots, the y-axis represents reliability (ICC); the x-axis represents the number of bootstrap trials ( $n_{boot}$ ); and separate curves represent different calibration sample sizes ( $n_{cal}$ ).

Figure 14

[PHL data]: Reliability of non-bootstrapped estimates of credibility, varying by calibration sample size ( $n_{cal}$ ).

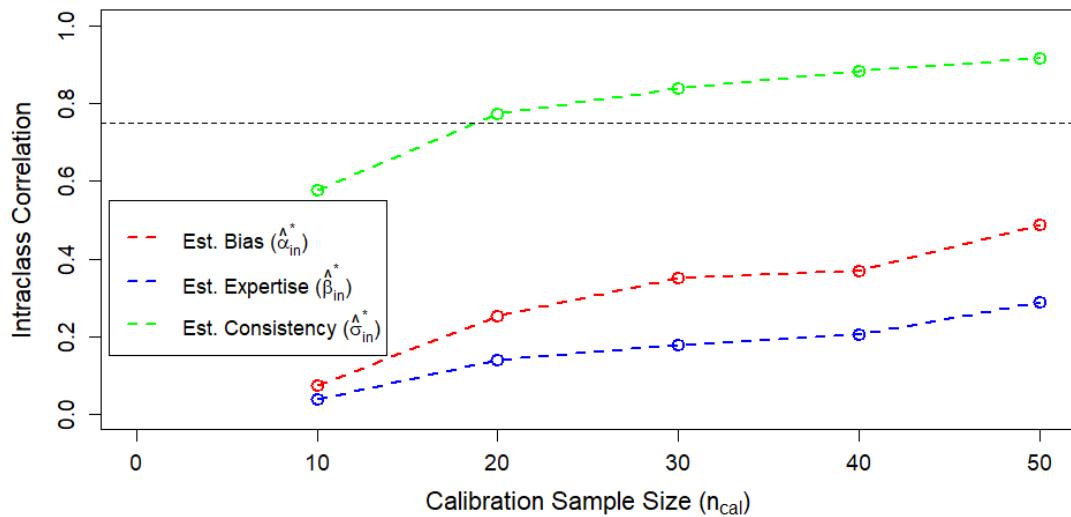


Figure 15

[PHL data]: Reliability of bootstrapped estimates of bias ( $\hat{\alpha}_i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ).

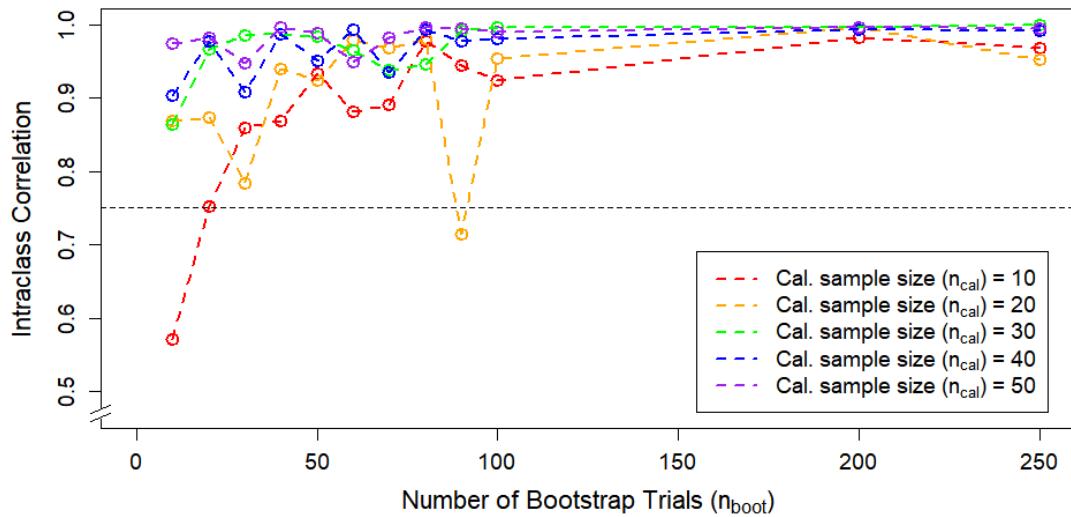


Figure 16

[PHL data]: Reliability of bootstrapped estimates of expertise ( $\hat{\beta}_i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ).

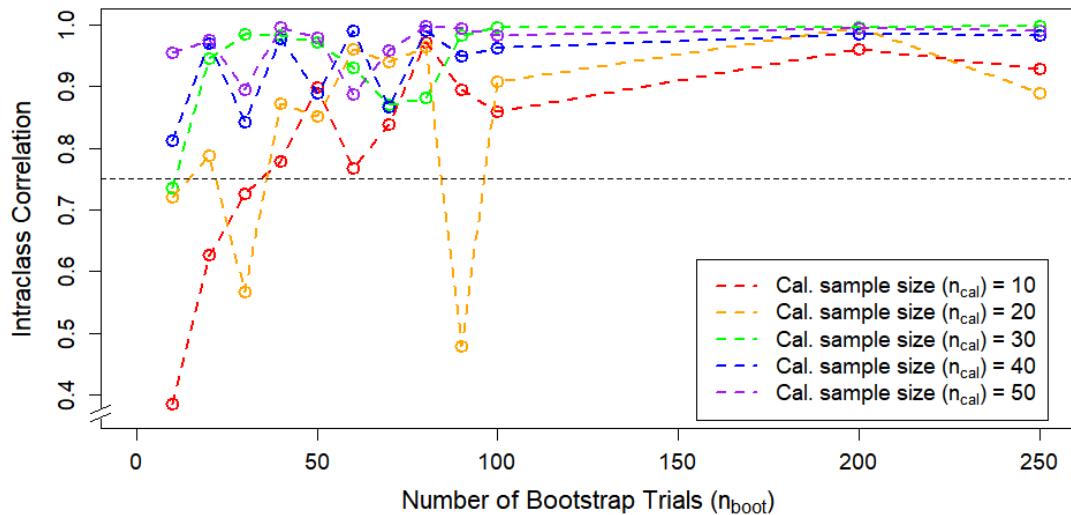
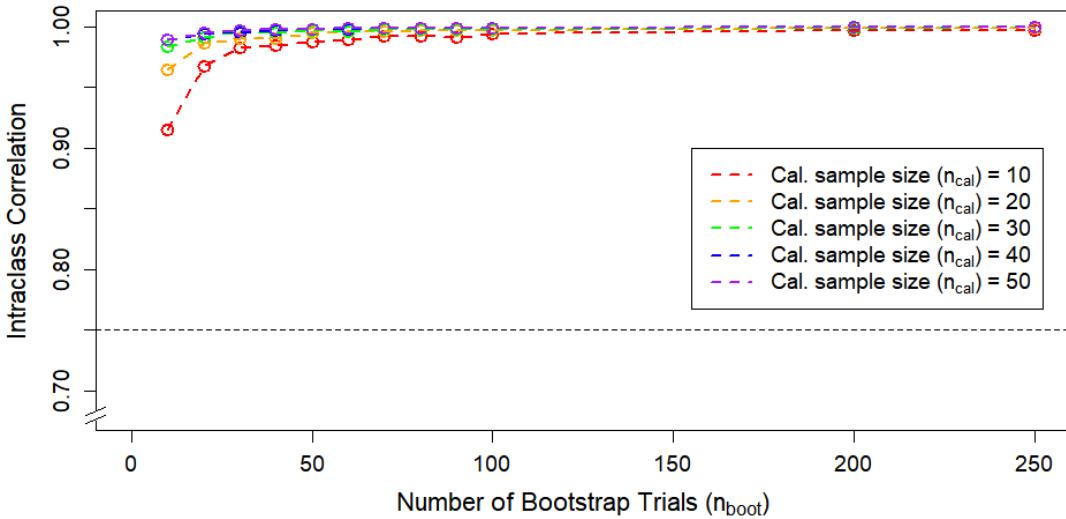


Figure 17

[PHL data]: Reliability of bootstrapped estimates of consistency ( $\hat{\sigma}_i$ ), varying by number of bootstrap trials ( $n_{boot}$ ) and calibration sample size ( $n_{cal}$ ).



**Discussion.** In contrast with previous reliability experiments, the results of Analysis 2c.i indicate that the reliabilities of non-bootstrapped estimates of credibility were somewhat inconsistent when fit to the Philadelphia air temperature data. As can be seen in Figure 14, non-bootstrapped estimates of *consistency* ( $\hat{\sigma}_{in}^*$ ) surpassed the conventional threshold for “excellent” intraclass correlation (ICC  $\geq 0.75$ ; Cicchetti, 1994) at a calibration sample size of  $n_{cal} = 20$  and remained high throughout the experiment. However, non-bootstrapped estimates of *bias* ( $\hat{a}_{in}^*$ ) and *expertise* ( $\hat{\beta}_{in}^*$ ) started extremely low in Analysis 2c.i, and never surpassed the thresholds for “poor” (ICC  $< 0.40$ ) and “fair” ( $0.40 \leq \text{ICC} \leq 0.59$ ), respectively (Cicchetti, 1994). Fortunately for Study 2c, these results did not compromise the remaining analyses in the Philadelphia air temperature study, as the reliability of bootstrapped estimates for all three components of credibility stabilized at above-excellent levels for experimental cells with  $n_{boot} \geq 100$  bootstrap trials (Figures 15-17).

Despite these somewhat puzzling results, Analysis 2c.i led me to two decisions. The first was to forge-ahead with my planned analyses of the Philadelphia air temperature data, even though the low reliabilities of non-bootstrapped estimates of *bias* and *expertise* would set an upper-bound on the validity of these estimates and create a high bar for credibility-based recalibration. For the purposes of Studies 2c (PHL reliability/validity) and 3c (PHL recalibration), therefore, I selected the following set of analytic parameters: a calibration sample size of  $n_{cal} = 50$  to maximize the consistency of the recalibration transformation from trial-to-trial (while still remaining realistic for real-world data-sets); a minimum prediction sample size of  $n_{pred} = 30$  to ensure that each forecaster would have an adequate number of judgments with which to examine the effects of recalibration out-of-sample; and  $n_{boot} = 100$  bootstrap trials to ensure that the validities of bootstrapped credibility estimates would not be unduly constrained by being unreliable.

Perhaps more importantly, however, the results of Analysis 2c.i led me to a second decision. Because Study 2c (PHL reliability/validity) was designed to examine the linear credibility framework under representative (and fairly favorable) conditions, I was surprised to see the reliabilities of non-bootstrapped estimates of *bias* and *expertise* depart so dramatically from those reported in Analyses 2a.i (GJP reliability) and 2b.i (MM reliability) — to say nothing of their contrast with non-bootstrapped estimates of *consistency* in the same study. Based on these results, I decided to conduct a variety of follow-up analyses in which I examined linear credibility estimates derived from subsets of the Philadelphia air temperature data (e.g., January predictions only vs. July

predictions only). As I report in Chapter 4, the results of these analyses revealed an important limitation of the linear credibility framework: namely, that “errors” and “biases” can only be detected if they are (reasonably) stable across judgments. If they are not — as was the case in the Philadelphia air temperature data, then the reliability and validity of the linear credibility estimates will be limited (for additional details about these analyses, see: Study 4).

### **Analysis 2c.ii: What are the predictors of credibility and what does credibility predict? (PHL validity)**

Despite the limitations of the Philadelphia air temperature data (suboptimal reliability; inconsistent errors and biases across question-types), the primary goal of Study 2c was to determine whether linear credibility estimates could be used to distinguish between “better” vs. “worse” judges of subjective probability. Consequently, in Analysis 2c.ii, I once-again examined the validity of linear credibility estimates as predictors of forecast accuracy (i.e., average Brier scores). As with other examinations of validity, the prevailing assumption of this analysis was not that linear credibility estimates would represent face-valid measures of a participant’s errors and biases, but rather that forecast accuracy and linear credibility estimates would both covary with the (latent) soundness of a participant’s judgments. Thus, even in the noisy world of judgment under uncertainty, I expected accurate judgment and “credible” judgment to move together.

## **Method.**

**Detailed Procedure.** Using the analytic parameters selected in Analysis 2c.i (PHL reliability) (i.e.,  $n_{\text{cal}} = 50$ ;  $n_{\text{pred}} = 30$ ;  $n_{\text{boot}} = 100$ ), I applied the General Procedure to arrive at bootstrapped credibility estimates for each forecaster in the Philadelphia air temperature study (i.e.,  $\hat{\alpha}_i$ ,  $\hat{\beta}_i$ , and  $\hat{\sigma}_i$ ). After excluding participants who provided too few forecasts ( $< 80$ ), this data-set included forecasts from 73 participants. Within this sample, the mean age was 48.8 ( $Mdn. = 49.0$ ;  $SD = 12.5$ ) and 36% self-identified as male.

After estimating credibility, I then transformed estimates of *bias* and *expertise* to measures of absolute difference from identity (i.e., I calculated  $\hat{\alpha}'_i$  and  $\hat{\beta}'_i$  by taking the absolute differences of the untransformed values from 0 and 1, respectively). Then, using similar procedures to Analysis 2a.ii (GJP validity) (i.e., all continuous variables standardized; sex dummy-coded), I used pairwise correlations and exploratory linear regressions to probe the relationships among bootstrapped credibility estimates ( $\hat{\alpha}'_i$ ,  $\hat{\beta}'_i$ , and  $\hat{\sigma}_i$ ) and the handful of individual difference measures captured in the Philadelphia air temperature study.

In the interest of simplicity, Study 2c (PHL reliability/validity) focused on the relationships between linear credibility estimates and forecast accuracy (i.e., *average Brier score*). However, because this study was conducted in collaboration with Jonathan Baron (the principal advisor on this thesis), participants also completed Baron's 11-item Actively Open-Minded Thinking scale (*in press*) (*AOMT Score* in Tables 18-22, below) in addition to providing basic information about their age and sex. Similar to previous studies, exploratory regressions were conducted under three approaches to variable

selection: the *kitchen sink* approach, the *reduced* approach, and the *ridge-ISE* approach (for additional details, see the Detailed Procedure section of Analysis 2a.ii (GJP validity)) — a methodology that allowed me to look for instances where variables were competing for overlapping segments of explanatory variance.

**Results.** The results on Analysis 2c.ii can be seen in Tables 18-22, below. Simple correlations between bootstrapped credibility estimates ( $\hat{\alpha}'_i$ ,  $\hat{\beta}'_i$ , and  $\hat{\sigma}_i$ ) and individual difference measures can be seen in Table 18; exploratory regression analyses examining the predictors of credibility can be seen Tables 19-21; and an exploratory regression analysis concerning the predictors of forecast accuracy (average Brier score) can be seen in Table 22.

Table 18

[PHL data]: Simple correlations between credibility estimates and individual difference measures.

Ind. Diff. Measure	Credibility Estimate		
	$\hat{\alpha}'_i$ (Bias)	$\hat{\beta}'_i$ (Expertise)	$\hat{\sigma}_i$ (Consist.)
Average Brier score	0.75***	0.41***	0.19
AOMT Score	0.27*	-0.11	-0.20
Age	-0.32**	0.04	0.09
Male	0.08	-0.19	-0.22

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: All non-categorical measures standardized; gender dummy-coded.

Table 19

[PHL data]: Predictors of bootstrapped alpha (i.e., bias;  $\hat{\alpha}'_i$ ).

	Variable Selection Approach		
Ind. Diff. Measure	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	0.03 (0.09)	0.00 (0.07)	0.00 (0.08)
Average Brier score	0.72 (0.08)***	0.72 (0.07)***	0.75 (0.08)***
AOMT Score	0.05 (0.08)		
Age	-0.24 (0.08)**	-0.25 (0.07)**	
Male	-0.1 (0.17)		
Multiple R <sup>2</sup>	0.623	0.62	0.558
Adjusted R <sup>2</sup>	0.601	0.609	0.552
RMSE	0.61	0.612	0.66
AIC	146.91	143.51	152.58
BIC	160.66	152.67	159.45

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: All non-categorical measures standardized; gender dummy-coded.

Table 20

[PHL data]: Predictors of bootstrapped beta (i.e., expertise;  $\hat{\beta}'_i$ ).

	Variable Selection Approach		
Ind. Diff. Measure	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	0.16 (0.13)	0.00 (0.11)	0.00 (0.11)
Average Brier score	0.49 (0.11)***	0.41 (0.11)***	0.41 (0.11)***
AOMT Score	-0.15 (0.12)		
Age	0.06 (0.11)		
Male	-0.44 (0.23)		
Multiple R <sup>2</sup>	0.259	0.168	0.168
Adjusted R <sup>2</sup>	0.215	0.156	0.156
RMSE	0.855	0.906	0.906
AIC	196.28	198.74	198.74
BIC	210.02	205.61	205.61

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: All non-categorical measures standardized; gender dummy-coded.

Table 21

[PHL data]: Predictors of bootstrapped sigma (i.e., consistency;  $\hat{\sigma}_i$ ).

Ind. Diff. Measure	Variable Selection Approach		
	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	0.15 (0.14)	0.00 (0.12)	0.00 (0.12)
Average Brier score	0.28 (0.12)*	0.19 (0.12)	
AOMT Score	-0.18 (0.13)		
Age	0.09 (0.12)		
Male	-0.43 (0.25)		
Multiple R <sup>2</sup>	0.144	0.036	0
Adjusted R <sup>2</sup>	0.094	0.022	0
RMSE	0.919	0.975	0.993
AIC	206.81	209.48	210.16
BIC	220.55	216.35	214.74

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: All non-categorical measures standardized; gender dummy-coded.

Table 22

[PHL data]: Predictors of average Brier score (i.e., forecast accuracy).

Ind. Diff. Measure	Variable Selection Approach		
	Kitchen-Sink	Reduced	Ridge-1SE
(Intercept)	-0.06 (0.1)	0.00 (0.08)	0.00 (0.08)
Estimated bias ( $\hat{\alpha}'_i$ )	0.74 (0.1)***	0.75 (0.08)***	0.75 (0.08)***
Estimated expertise ( $\hat{\beta}'_i$ )	0.10 (0.11)		
Estimated consistency ( $\hat{\sigma}_i$ )	-0.07 (0.1)		
AOMT Score	0.06 (0.09)		
Age	0.15 (0.09)		

Male	0.17 (0.18)		
Multiple R <sup>2</sup>	0.598	0.558	0.558
Adjusted R <sup>2</sup>	0.561	0.552	0.552
RMSE	0.63	0.66	0.66
AIC	155.7	152.58	152.58
BIC	174.03	159.45	159.45

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

Note: All non-categorical measures standardized; gender dummy-coded.

**Discussion.** The results of Analysis 2c.ii demonstrate that bootstrapped estimates of *bias* ( $\hat{\alpha}'_i$ ) and *expertise* ( $\hat{\beta}'_i$ ) demonstrated strong convergent validity with forecast accuracy (average Brier scores) in the Philadelphia air temperature study. As can be seen in Table 18, simple correlations suggest that estimated *bias* and *expertise* were both strongly related to better (lower) Brier scores, and that estimated *bias* was moderately related to participants' tendency towards actively open-minded thing (*AOMT*). The convergent relationships between *bias*, *expertise*, and forecast accuracy are also evident in Tables 19 and 20 (*bias* and accuracy in Table 19; and *expertise* and accuracy in Table 20), and the convergent relationship between *bias* and forecast accuracy in Table 22. Unfortunately, throughout these exploratory analyses, bootstrapped estimates of *consistency* ( $\hat{\sigma}_i$ ) were not strongly related to forecast accuracy (Table 21). However, the post-hoc reliability analyses conducted in Analysis 2c.i suggest that this may have been because participants' tendencies towards error and bias varied across question types. Thus, I return to this issue when I re-examine subsets of the Philadelphia air temperature data in Study 4.

## General Discussion

Despite Study 2c's intentions, the Philadelphia air temperature data were not as unimpeachable as originally intended. However, the results of Study 2c suggest that it is generally reasonable to construe linear credibility estimates as individual-difference-type measures of "skill" or "proficiency" in subjective probability judgment. In Analysis 2c.i (PHL reliability), results demonstrated that *bias*, *expertise*, and *consistency* can be reliably estimated with moderate calibration sample-sizes and/or a large number of bootstrap trials, and in Analysis 2c.ii (PHL validity), results demonstrated that estimates of *bias* and *expertise* can be used to predict forecast accuracy (average Brier scores). In practice, of course, the results of Study 2c also demonstrate that the reliability and validity of credibility estimates (and especially those that are derived from simple models such as the linear credibility framework) are constrained by the stability and generalizability of an individual's tendencies towards errors and biases in judgment.

Despite an experimental design that limited stability and generalizability, however, linear credibility estimates in Study 2c were still able to explain somewhere between 56% and 60% of the observed variance in forecast accuracy. Given that these estimates were derived from truly amateur forecasters with only lay knowledge of Philadelphia air temperatures — to say nothing of an oversimplified model of credibility — these results suggest that examining credibility is likely to provide decision makers with insight into "skill" or "proficiency" in subjective probability judgment across a wide variety of decision environments.

## Conclusions

Taken together, the results of Studies 2a-2c suggest that the linear credibility framework is an informative (albeit imperfect) tool for examining “skill” or “proficiency” in subjective probability judgment. In general, therefore, it is likely that decision makers will gain valuable insight by examining credibility, though they would be wise to treat such insight with a degree of skepticism. Given the inconsistent performance of linear credibility estimates across studies, the results of Study 2a-2c suggest that (linear) estimates of *bias*, *expertise*, and *consistency* may not be orthogonal components of credibility. Indeed, the rigidity and simplicity of the linear credibility framework make it an unlikely candidate for modeling credibility in a descriptively accurate way.

Nevertheless, Studies 2a-2c provided strong evidence to suggest that linear credibility estimates (in one combination or another) are often powerfully explanatory with respect to judgmental accuracy. As such, these studies provide a strong empirical basis for concluding that credibility can be probed with simple statistical tools and that doing so is often informative for decision makers. Furthermore, because it is reasonable to assume that the results of Studies 2a-2c represent an empirical lower-bound for the performance of empirical models of credibility, these results suggest that examinations of credibility may be informative, in general. Thus, when considered in a broader context, it is evident that credibility is a useful theoretical construct and that additional studies of credibility are warranted.

## References

- Armstrong, J. S. (2001). Combining forecasts. *Principles of forecasting: a handbook for researchers and practitioners*. J. S. Armstrong (Ed.). Norwell, MA: Kluwer Academic Publishers.
- Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2016). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3), 691-706.
- Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Vol. Ed.), *Psychology of learning and motivation, Vol. 8*, 47–89. New York: Academic Press.
- Baron, J. (2008). *Thinking and deciding*. New York: Cambridge University Press.
- Baron, J. (in press). Actively open-minded thinking in politics. *Cognition*.
- Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, 4(3), 265-284.
- Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, 58(3), 382-398.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3.
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1), 116-131.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284.
- Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, 7(1), 25-47.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25-42.

- Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, 62(2), 410-422.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1-22.
- Galton, F. (1907). Vox populi (The wisdom of crowds). *Nature*, 75(7), 450-451.
- Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, 8(3), 188-201.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgement under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015a). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, 21(1), 1-14.
- Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015b). Identifying and cultivating superforecasters as a method of improving probabilistic prediction. *Perspectives on Psychological Science*, 10(3), 267-281.
- Mellers, B., Ungar, L., Baron, J., Ramos, J., Gürçay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5), 1-10.
- Paine N., & Boice, J. (2017, March 14). The odds you'll fill out a perfect bracket [Blog post]. Retrieved from <https://fivethirtyeight.com/features/the-odds-youll-fill-out-a-perfect-bracket/>
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111-163.
- Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. New York: Random House, Inc.

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

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, 10(3), 243-268.

## CHAPTER 3

### THE LINEAR CREDIBILITY FRAMEWORK IS OFTEN USEFUL AND CREDIBILITY (IN GENERAL) IS WORTH EXAMINING

#### **Abstract:**

Errors in judgment can lead to errors in decision making. Thus, decision makers have an incentive to ensure that their subjective probability judgments (SPJs) are as well-calibrated as possible. However, decision makers do not always have the necessary information to assess calibration. To address this issue, models of credibility are intended to provide a substitute for calibration information. To determine whether empirical credibility models can deliver on this promise, this chapter examines whether the linear credibility framework can be used to reduce individuals' tendencies toward error and bias in judgment. Specifically, in Studies 3a-3c, I examine the effects of credibility-based recalibration on absolute judgment error (AJE), absolute linear error (ALE), and *reliability* (Murphy, 1973) to determine whether the linear credibility framework can improve accuracy. Across three data-sets, the results of these studies generally suggest that it can. Thus, I argue that even simple models of credibility can often be useful.

#### **Introduction**

When facing an uncertain yet consequential choice, a decision maker's principal concern is that errors in judgment will lead to errors in decision making. Though it is possible for such errors to exist without changing a decision maker's behavior (indeed, in some cases it is quite common; for an elaboration of this point, see: Arkes, Gigerenzer, & Hertwig, 2016), the presence of poorly-calibrated beliefs necessarily opens the door to edge-cases where the balance of (subjective) expected utility shifts. Thus, in any case where errors in belief are persistent or consequential, decision makers have a rational (i.e., utilitarian) incentive to ensure that their subjective probability judgments (SPJs) are as well-calibrated as possible (for an overview of probabilistic calibration, see: Lichtenstein, Fischhoff, & Phillips, 1982).

In practice, however, examining calibration is not always feasible. Because empirical baserates (and other approximations of “objective” probabilities) are often inestimable or unknown, decision makers generally do not have access to the necessary information to determine *if* and *how* their beliefs may err. In principle, empirical models of credibility are useful because they can provide decision makers with a proxy or substitute for this information. Though necessarily less informative than a direct, quantitative measure of an SPJ’s agreement with the “truth,” measures of credibility do this by estimating systematic differences between an individual’s judgments and an idealized set of *estimated optima*. If a credibility model can succeed in this task — and if the focal set of *estimated optima* is worth emulating — then some or all of the differences captured by this model should reflect systematic tendencies towards error and bias in judgment. Consequently, while empirical models of credibility are unlikely to be a panacea, they should allow decision makers to move judgments *closer* to the truth by identifying and ameliorating suboptimal judgment strategies that have historically prevented an individual’s SPJs from being as accurate and/or as well-calibrated as possible.

Following from this line of reasoning, one way to test the usefulness of an empirical model of credibility is to determine whether it can help a decision maker “undo” errors and biases in judgment (i.e., increase accuracy and/or calibration) by correcting for an individual’s personalized pattern of errors. In practice, of course, credibility estimates derived from simple models such as the linear credibility framework may not reflect the full scope of errors and biases in an individual’s judgments, and the

scope it does reflect may not be captured perfectly. Thus, the “correction” procedures encapsulated in an individual’s credibility function may not allow for perfect recalibration or even improve judgments all the time. However, if an empirical credibility model is to be useful *in general* — and certainly if a decision maker is to put stock in its estimates — then the typical effects of credibility-based recalibration must be positive and its effect-size practically significant.

To achieve this standard, a credibility model must meet three empirical criteria:

Criterion 1. The estimated relationship between an individual’s SPJs and a focal set of *estimated optima* must be stable and generalizable within a given decision environment.

Criterion 2. Reversing an individual’s estimated errors and biases in a given decision environment (i.e., employing credibility-based recalibration) must generally improve the accuracy and/or calibration of her SPJs.

Criterion 3. The typical effect-size associated with credibility-based recalibration must be both statistically and practically significant.

In Studies 3a-3c, I will examine whether the linear credibility framework meets these criteria when applied to empirical judgments. If it does (despite being a deliberately simple model), then decision makers may benefit from examining credibility more

generally, especially to the extent that their credibility models are fit with descriptively accurate curves rather than a single, atheoretical line.

### **Study 3a: Typical Effects of Recalibration with GJP Data (GJP Recalibration)**

In Study 2a (GJP reliability/validity), exploratory regression analyses revealed a strong degree of predictive validity between linear credibility estimates and forecast accuracy in the Good Judgment Project (GJP). From an empirical perspective, these relationships suggest that the linear credibility framework is sensitive to systematic trends in an individual's judgment strategy that bear on overall accuracy. As a practical matter, however, it is unclear whether these trends represent stable, generalizable tendencies toward error and bias in judgment, or whether they were an empirical catchall for circumstantial variations in forecast accuracy (e.g., differences in domain knowledge; random and/or uncharacteristic variations in accuracy; etc.). To address this issue, Study 3a will examine the degree to which the linear credibility framework meets the three criteria for a useful credibility model (see above) when applied to forecasts from the GJP.

As in Study 1, I will probe these criteria by examining the effects of credibility-based recalibration on out-of-sample SPJs. If correcting for the ostensible "errors" and "biases" in a forecaster's calibration sample systematically improves the accuracy and/or calibration of judgments in her prediction sample, then these estimates must reflect stable and generalizable relationships with *estimated optima* (criterion 1). When accuracy is measured in terms of absolute judgment error (i.e., AJE; the absolute difference between an SPJ and the corresponding *estimated optimum*), the existence of these relationships

would suggest that an individual's judgments are predictably different from *estimated optima*, though they cannot speak to the origin of these differences.<sup>24</sup> However, when accuracy is measured in terms of absolute linear error (i.e., ALE; the absolute difference between an SPJ and the corresponding empirical outcome), a systematic improvement due to recalibration would suggest that an individual's judgments are predictably less accurate than *estimated optima* — an outcome that solidifies their interpretation as errors and biases in judgment (criterion 2). Finally, if the typical effects of recalibration are substantial and systematic — either because they reduce departures from *estimated optima* (AJE) or empirical outcomes (ALE) — then it is likely that forecasters would have seen practical benefits from examining (and accounting for) their historical degree of credibility over the course of the GJP (criterion 3).

## Method

**Detailed procedure.** As described in the General Method section, Study 3a was conducted at the same time as Study 2a (GJP reliability/validity). Consequently, the dependent variables I examined in Study 3a were observed during the same run of the General Procedure that was used to calculate bootstrapped credibility estimates in Analysis 2a.ii (GJP validity). In this analysis, credibility functions were fit to a

---

<sup>24</sup> Consider, for example, a case where some individuals have “inside information” that is not available to other members of the crowd. In this case, a positive effect of recalibration on AJE would indicate a systematic relationship between an individual’s judgments and *estimated optima*, but this “improvement” in AJE might correspond to a decrease in predictive accuracy if (crowdsourced) *estimated optima* are systematically less accurate than SPJs derived from inside information.

calibration sample size of  $n_{\text{cal}} = 50$ ; forecasters were only included if they provided enough SPJs to accommodate a minimum prediction sample size of  $n_{\text{pred}} = 30$ ; and the effects of recalibration were observed over  $n_{\text{boot}} = 100$  bootstrap trials.

As a brief reminder, data for Study 3a were generated by recalibrating each participant's SPJs across a large number of bootstrap trials. For a given participant on a given trial, this was accomplished by (a) using a random sampling procedure to partition participant  $i$ 's SPJs into a calibration sample ( $n_{\text{cal}} = 50$ ) and a prediction sample (all remaining SPJs); (b) estimating a credibility function by regressing participant  $i$ 's calibration sample on the corresponding *estimated optima*; (c) using the resulting credibility function to recalibrate SPJs in participant  $i$ 's prediction sample; and (d) recording the effects of recalibration for the given trial (for a list of specific outcome measures, see below). For the purposes of Study 3a, this process was conducted a total of 100 times per forecaster (i.e., over 100 bootstrap trials), with a new calibration sample and a new credibility function being selected each time.

To examine the effects of credibility-based recalibration, Study 3a recorded ten outcome measures for each forecaster on each bootstrap trial. Eight of these outcome measures were variants of judgmental accuracy and can be divided into two general categories: (a) absolute judgment error (AJE), which measures to the absolute difference between an SPJ and the corresponding *estimated optimum*; and (b) absolute linear error (ALE), which measures the absolute difference between an SPJ and the corresponding

empirical outcome.<sup>25</sup> For each type of accuracy, I recorded four summary statistics per trial, yielding eight of the ten total outcome measures:

- The proportion of individual judgments for which recalibration improved (reduced) AJE / ALE;
- The mean pairwise difference in AJE / ALE due to recalibration;<sup>26</sup>
- The effect-size (Cohen's  $d$ ) associated with pairwise changes in AJE / ALE due to recalibration;
- And a binary indicator of whether recalibration improved (reduced) the sample's mean AJE / ALE.<sup>27</sup>

In addition to judgmental accuracy, I also recorded two outcome measures related to changes in *reliability* on each trial. As discussed in the General Methods section, *reliability* is one of the three components of Murphy's (1973) three-component decomposition of the Brier score (Brier, 1950) and is defined as the weighted sum of squared-differences between an individual's SPJs and observed baserates (here, the relative frequencies of event occurrence for SPJs separated into 101 percentage-point

---

<sup>25</sup> In all three studies (3a-3c), I also examined the empirical effects of recalibration on Brier scores. However, because the Brier score is a variant of the quadratic scoring rule, *changes* in Brier scores often produce highly skewed distributions. As such, it is not always informative to examine the “typical” effects of recalibration in terms of mean or median changes in Brier scores. Despite this limitation, however, the typical effects of credibility-based recalibration on Brier scores were not substantively different than those reported below.

<sup>26</sup> Note that these measures (i.e., the mean differences in AJE / ALE) are mathematically equivalent to the difference in mean AJE / ALE, due to recalibration. Thus, I will only discuss the former and not the latter when presenting results.

<sup>27</sup> In all cases, binary indicators were coded as “1” if the stated event occurred, and “0” if it did not.

“bins”). From a mathematical perspective, *reliability* is closely related to the forecasting measure *calibration* (see: Lichtenstein et al., 1982) and can be conceptualized as a measure of agreement between subjective probability judgments and empirical baserates (though it is defined in such a way that larger values indicate worse agreement). Though conceptually similar to *consistency* (i.e., both measures describe the extent to which SPJs depart from a set of benchmark judgments), the value of examining *reliability* in this context is that *reliability* is independent of subjective beliefs. By contrast, *consistency* describes the extent to which an individual’s SPJs tend to agree with *estimated optima* — a standard which here depends on intersubjective agreement. In principle, therefore, *reliability* provides a more “objective” measure of SPJ validity, though the strength of this claim is limited by the extent to which within-sample baserates are representative of “objective” probabilities. Thus, in each trial of Study 3a (GJP recalibration), I recorded two outcome measures related to *reliability*:

- A binary indicator of whether recalibration improved (reduced) *reliability*;
- And the pairwise change in sample *reliability* associated with recalibration.

By averaging outcome measures (within-measure) across each participant’s 100 bootstrap trials, Study 3a (GJP recalibration) produced 10 summary-statistics per participant:

- The typical<sup>28</sup> proportion of participant  $i$ 's judgments for which recalibration improved (reduced) AJE / ALE.
- The typical pairwise change in participant  $i$ 's AJE / ALE due to recalibration;
- The typical effect-size (Cohen's  $d$ ) of recalibration on participant  $i$ 's AJE / ALE;
- The proportion of prediction samples in which recalibration improved (reduced) participant  $i$ 's mean AJE / ALE.
- The proportion of prediction samples in which recalibration improved (reduced) *reliability*;
- And typical pairwise change in sample *reliability* due to recalibration.

As in Study 1, each of these summary-statistics represented a different aspect of the *typical* or expected effect of recalibration for a given participant. Thus, these measures were used as the primary dependent variables (DVs) in the tests that follow.

## Results

**Analysis 3a.i: typical effects of recalibration on AJE (GJP recal., AJE).** In Analysis 3a.i, I examined the typical effects of recalibration on absolute judgment error

---

<sup>28</sup> As in Chapter 1, I will use the word “typical” in this chapter to indicate bootstrapped averages, calculated over each participant’s 100 recalibration trials. This is done to prevent confusion in instances where “typical” values are bootstrapped averages of sample-level means.

(AJE). As discussed above, the effects of recalibration on AJE can be used to draw inferences about the stability and generalizability of the observed relationship between an individual's judgments and *estimated optima*. If credibility-based recalibration can be used to improve (reduce) AJE in out-of-sample judgments, then this would suggest that the ostensible “errors” and “biases” observed in an individual’s calibration sample were similar to those that remained unobserved in her prediction sample. Though a reliable improvement in AJE is not sufficient to conclude that the differences between an individual’s judgments and *estimated optima* were detrimental (see: Footnote 24), it is enough to conclude that some portion of this relationship was stable and generalizable. Thus, the observation of a reliable, positive effect of recalibration on AJE would suggest that the linear credibility framework meets the first criterion of a useful model of credibility.

To determine whether this was the case, Analysis 3a.i examined the typical effects of recalibration on AJE, across forecasters. The results of this analysis can be seen in the figures and tables below. Figures 18-21 show the empirical distributions of the four AJE-related DVs across forecasters, and Figure 22 shows a visual comparison of mean AJEs before and after recalibration. Table 23 provides descriptive statistics for each of the distributions represented in Figures 18-21, and Table 24 shows the results of Wilcoxon signed-rank tests examining the null hypothesis that the median of each distribution does not differ from chance (represented by red dotted lines in Figures 18-21).

Figure 18

[GJP data]: Typical proportion of judgments for which recalibration improved (reduced) AJE.

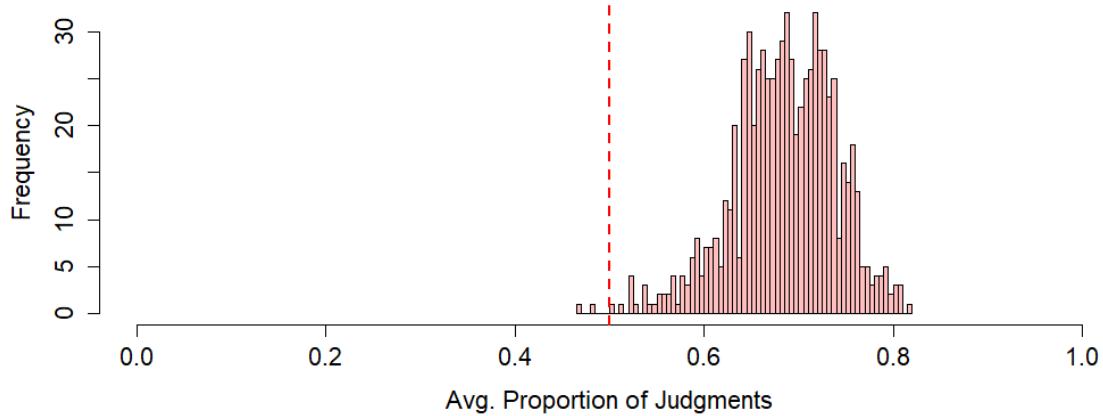
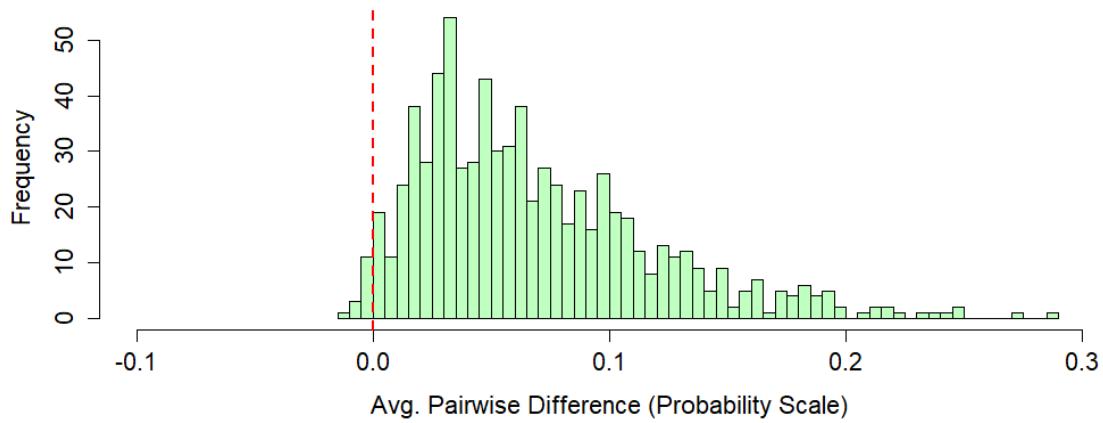


Figure 19

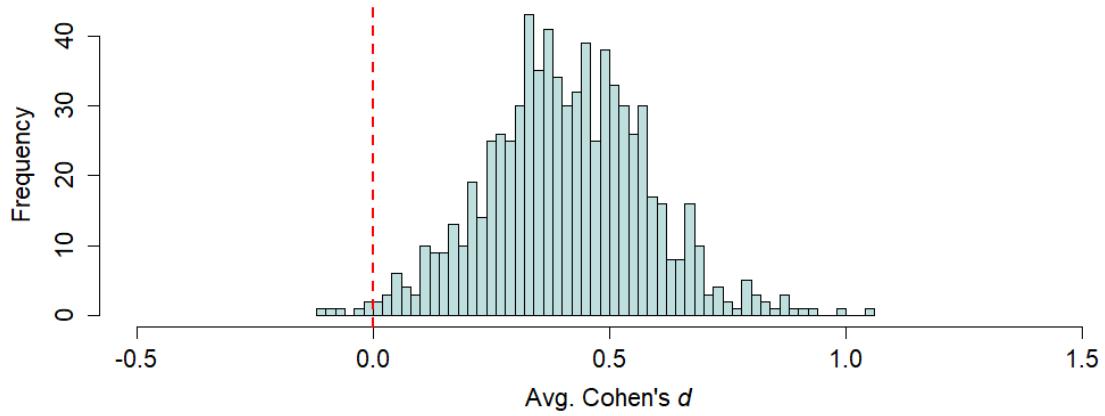
[GJP data]: Typical pairwise change in AJE (pre – post), due to recalibration.



Note: positive values indicate an improvement (reduction) in AJE.

Figure 20

[GJP data]: Typical effect-size (Cohen's  $d$ ) of recalibration on AJE.



*Note:* positive values indicate an improvement (reduction) in AJE.

Figure 21

[GJP data]: Proportion of samples in which recalibration improved (reduced) mean AJE.

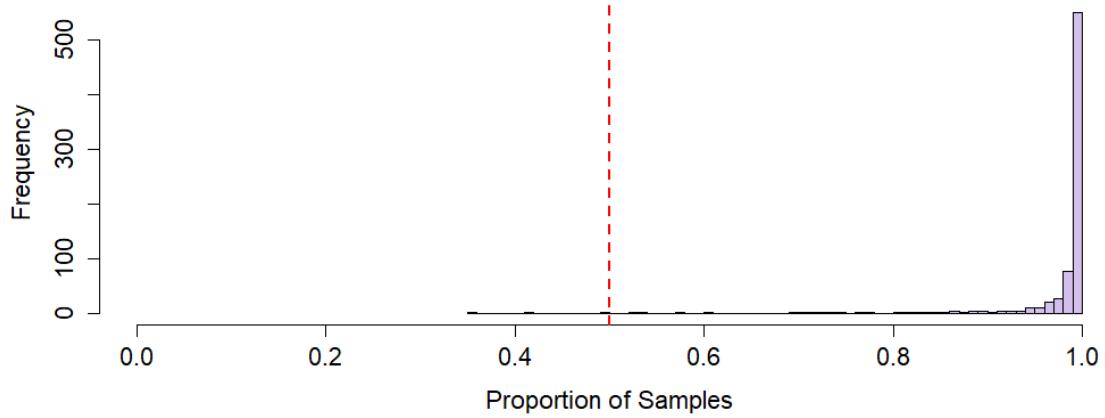
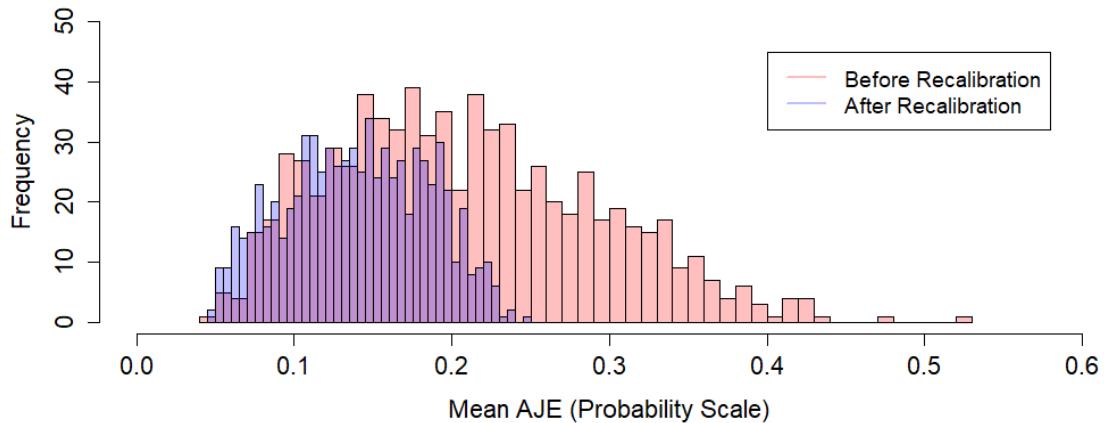


Figure 22

[GJP data]: Mean AJE, before and after recalibration.



*Note:* smaller values indicate more accurate judgements (smaller errors), on average.

Table 23

*[GJP data]: Typical effects of credibility-based recalibration on AJE, summarized across forecasters.*

Outcome Measure	Between-Subjects Summary Statistics				
	Mean	Mdn.	SD	Min.	Max.
Typical proportion of judgments for which recalibration improved AJE	69%	69%	5%	47%	82%
Typical pairwise change in AJE (pre - post), due to recalibration	6.84 $\times 10^{-2}$	5.84 $\times 10^{-2}$	5.00 $\times 10^{-2}$	-1.32 $\times 10^{-2}$	28.98 $\times 10^{-2}$
Typical effect (Cohen's $d$ ) of recalibration on AJE	0.41	0.41	0.17	-0.11	1.05
Proportion of samples in which recalibration improved mean AJE	98%	100%	7%	35%	100%

Table 24

*[GJP data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced AJE) beyond chance.*

Outcome Measure	$H_0$	$Prop. Mass > H_0$	Stat. ( $V$ )	p-value
Typical proportion of judgments for which recalibration improved AJE	Mdn. = 0.5	100%	$2.85 \times 10^5$	<.001***
Typical pairwise change in AJE (pre - post), due to recalibration	Mdn. = 0	98%	$2.84 \times 10^5$	<.001***
Typical effect (Cohen's $d$ ) of recalibration on AJE	Mdn. = 0	99%	$2.85 \times 10^5$	<.001***
Proportion of samples in which recalibration improved mean AJE	Mdn. = 0.5	99%	$2.83 \times 10^5$	<.001***

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

**Discussion.** The results of Analysis 3a.i overwhelmingly suggest that credibility-based recalibration typically led to improvements in AJE. Depending on how AJE was summarized, the study-wide expected effects of recalibration (i.e., the typical effects of recalibration for the average GJP forecaster) were to reduce AJE in 69% of individual judgments, each by an average of 6.84 points on the probability scale. Overall, this corresponded to an average, pairwise Cohen's  $d$  of 0.41, and a smaller mean AJE in 98% of bootstrap samples, on average. For all four AJE-based outcome measures, this corresponded to a distribution of typical effect sizes that was significantly greater than chance, suggesting that credibility-based recalibration would have been expected to improve AJE in the Good Judgment Project, in general. Indeed, even when examined in terms of the worst-performing measure (typical pairwise change in AJE), fewer than 2.5% of forecasters could have expected a negative effect of recalibration. Consequently, the results of Analysis 3a.i strongly suggest that the linear credibility framework fulfilled

the first criterion of a useful credibility model (stable, generalizable relationships) when fit to forecasts from the GJP.

**Analysis 3a.ii: typical effects of recalibration on ALE (GJP recal., ALE).** In Analysis 3a.ii, I examined the typical effects of recalibration on absolute linear error (ALE). Much like AJE, ALE can be used to draw inferences about the stability and generalizability of the observed relationship between individual judgments and *estimated optima*. However, because recalibration will only improve (reduce) ALE if *estimated optima* are more accurate than individual judgments, observing a reliable, positive effect of recalibration would suggest that the linear credibility framework is capturing genuine errors and biases in judgment. Practically speaking, therefore, if the results of Analysis 3a.ii reveal a reliable improvement (reduction) in ALE, then this would suggest that the linear credibility framework meets the first two criteria of a useful model of credibility (stable/generalizable relationships; capable of improving judgments). Furthermore, if the practical effect-size of this improvement is large, a case can be made that the same result also fulfills criterion 3 (practical value).

To test these possibilities, Analysis 3a.ii examined the typical effects of credibility-based recalibration on ALE, across forecasters. The results of this analysis can be seen in the figures and tables below. Figures 23-26 show the empirical distributions of the four ALE-related DVs across forecasters, and Figure 27 shows a visual comparison of mean ALEs before and after recalibration. Table 25 provides descriptive statistics for each of the distributions represented in Figures 23-26, and Table 26 shows the results of

Wilcoxon signed-rank tests examining the null hypothesis that the median of each distribution does not differ from chance (represented by red dotted lines in Figures 23-26).

Figure 23

[GJP data]: Typical proportion of judgments for which recalibration improved (reduced) ALE.

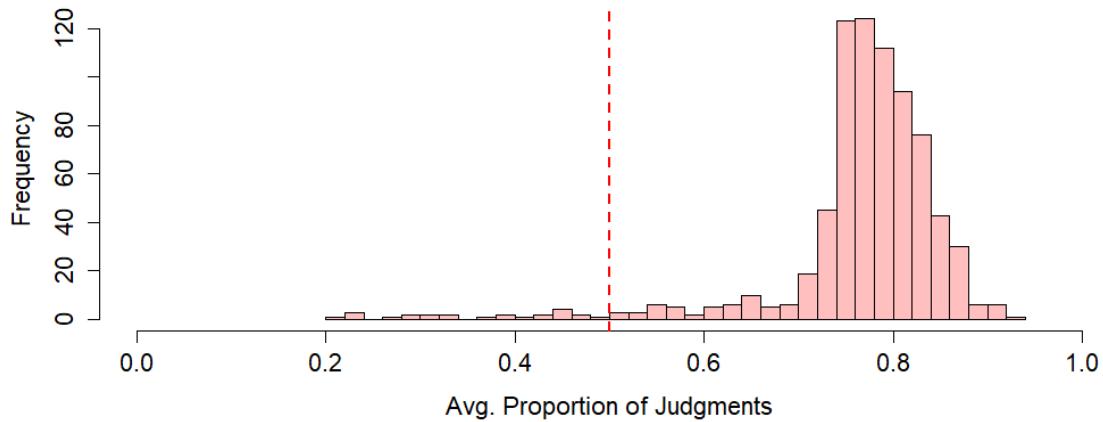
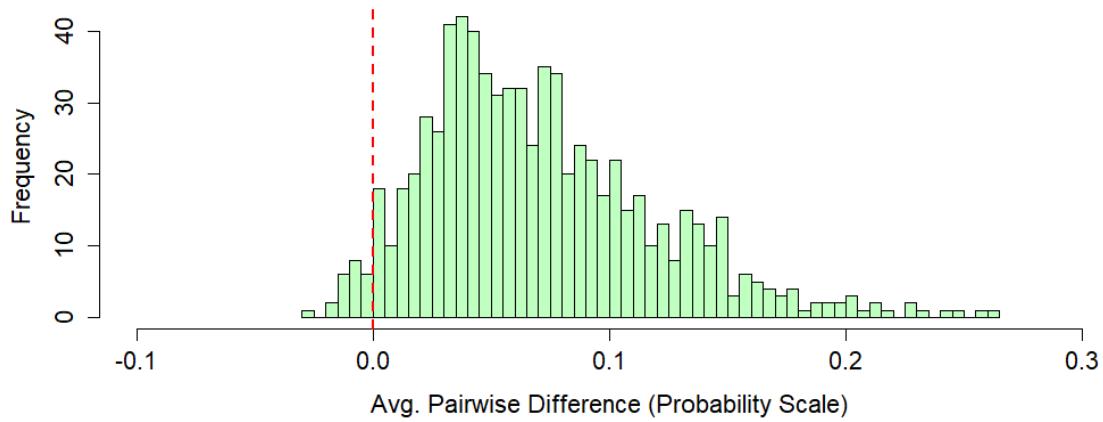


Figure 24

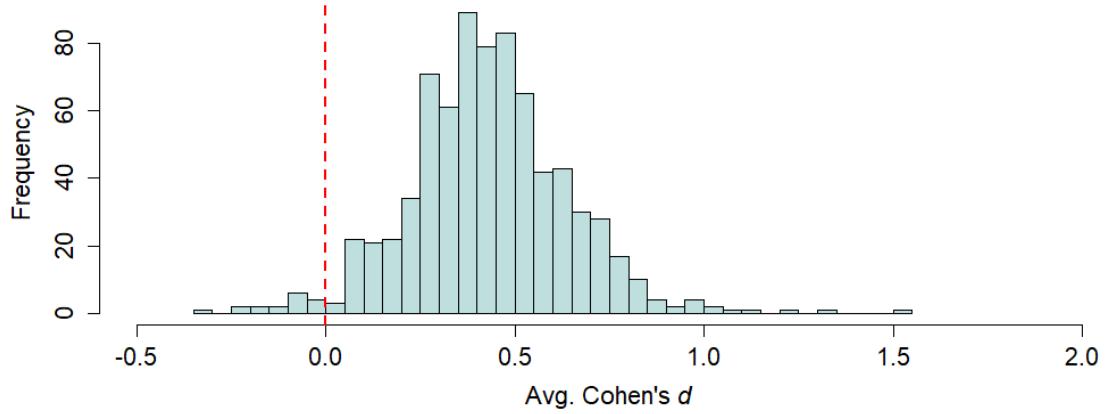
[GJP data]: Typical pairwise change in ALE (pre – post), due to recalibration.



Note: positive values indicate an improvement (reduction) in ALE.

Figure 25

[GJP data]: Typical effect-size (Cohen's  $d$ ) of recalibration on ALE.



Note: positive values indicate an improvement (reduction) in ALE.

Figure 26

[GJP data]: Proportion of samples in which recalibration improved (reduced) mean ALE.

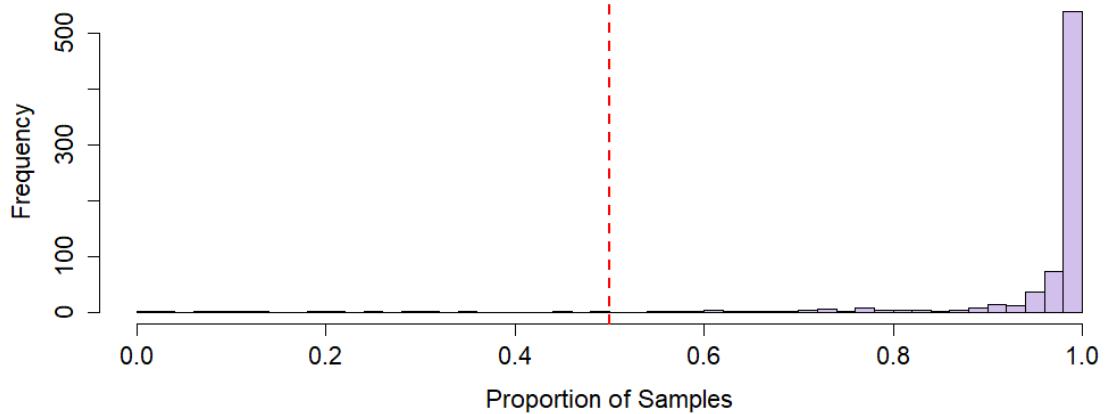
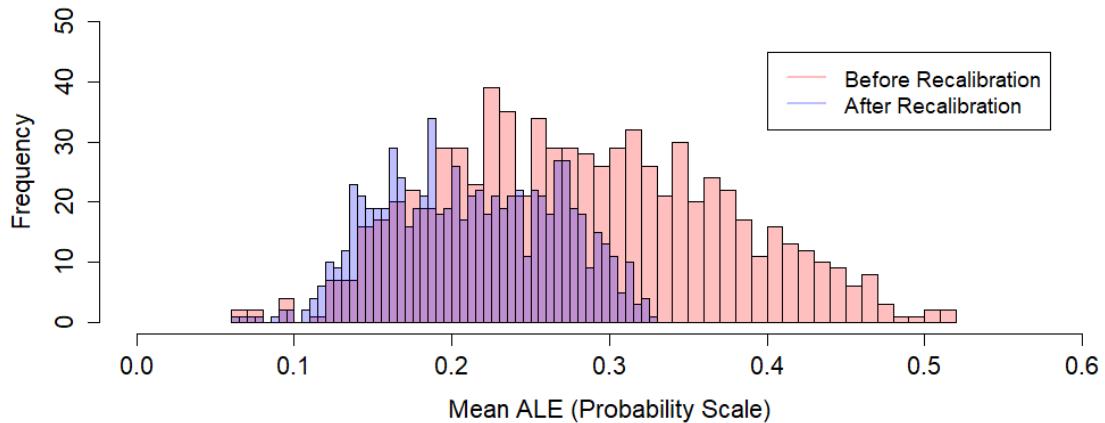


Figure 27

[GJP data]: Mean ALE, before and after recalibration.



*Note:* smaller values indicate more accurate judgements (smaller errors), on average.

Table 25

*[GJP data]: Typical effects of credibility-based recalibration on ALE, summarized across forecasters.*

Outcome Measure	Between-Subjects Summary Statistics				
	Mean	Mdn.	SD	Min.	Max.
Typical proportion of judgments for which recalibration improved ALE	76%	78%	10%	21%	93%
Typical pairwise change in ALE (pre - post), due to recalibration	$7.07 \times 10^{-2}$	$6.23 \times 10^{-2}$	$4.87 \times 10^{-2}$	-2.84	$26.36 \times 10^{-2}$
Typical effect (Cohen's $d$ ) of recalibration on ALE	0.43	0.42	0.21	-0.30	1.51
Proportion of samples in which recalibration improved mean ALE	95%	100%	14%	0%	100%

Table 26

*[GJP data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced ALE) beyond chance.*

Outcome Measure	$H_0$	$Prop. Mass > H_0$	Stat. ( $V$ )	p-value
Typical proportion of judgments for which recalibration improved ALE	Mdn. = 0.5	97%	$2.82 \times 10^5$	<.001***
Typical pairwise change in ALE (pre - post), due to recalibration	Mdn. = 0	97%	$2.84 \times 10^5$	<.001***
Typical effect (Cohen's $d$ ) of recalibration on ALE	Mdn. = 0	98%	$2.84 \times 10^5$	<.001***
Proportion of samples in which recalibration improved mean ALE	Mdn. = 0.5	98%	$2.82 \times 10^5$	<.001***

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

**Discussion.** The results of Analysis 3a.ii demonstrate that credibility-based recalibration typically led to large improvements in ALE. As with AJE, this conclusion was supported by all four ALE-based outcome measures, with the average GJP forecaster being able to expect 76% of her individual judgments to improve, each by an average of 7.07 points on the probability scale. Overall, this corresponded to an average, pairwise Cohen's  $d$  of 0.43, and a better (smaller) mean ALE in 95% of a forecaster's samples, on average. Once again, these averages corresponded to distributions of typical effect sizes that were significantly better than chance, suggesting that credibility-based recalibration would likely have improved ALE in the Good Judgment Project, in general. Critically, all four distributions of DVs suggested that the expected effects of recalibration were very rarely negative (< 3.5% of forecasters), and that the practical effects of recalibration were substantial. Thus, the results of Analysis 3a.ii suggest that the linear credibility

framework fulfilled all three criteria for a useful credibility model when applied to judgments from the GJP.

**Analysis 3a.iii: typical effects of recalibration on *reliability* (GJP recal., *reliability*)**. In Analysis 3a.iii, I examined the typical effects of recalibration on the *reliability* of GJP forecasts. As discussed above, *reliability* is closely related to the forecasting term *calibration* (see: Lichtenstein et al., 1982), though defined in such a way that better *reliability* is indicated by smaller numbers (Murphy, 1973). As with ALE, a reliable improvement (decrease) in *reliability* due to recalibration would suggest that the linear credibility framework had captured genuine errors and biases in judgment. Critically, however, improvements in ALE might result in better scores simply because judgments became more extreme (and thus were closer to empirical outcomes of 0 or 1). Examining *reliability* avoids this problem by comparing SPJs to within-sample baserates that need not be extreme. Consequently, if credibility-based recalibration reliably improves *reliability*, it is because the calibration of forecasts has genuinely improved, relative to empirical baserates.

To determine whether the linear credibility framework could systematically improve forecast calibration, Analysis 3a.iii examined the typical effects of credibility-based recalibration on *reliability*, across forecasters. The results of this analysis can be seen in the figures and tables below. Figures 28 and 29 show the empirical distributions of the two *reliability*-based DVs across forecasters. Table 27 provides descriptive statistics for each of the distributions represented in Figures 28 and 29, and Table 28

shows the results of Wilcoxon signed-rank tests examining the null hypothesis that the median of each distribution does not differ from chance (represented by red dotted lines in Figures 28 and 29).

Figure 28

[GJP data]: Proportion of samples in which recalibration improved (reduced) reliability.

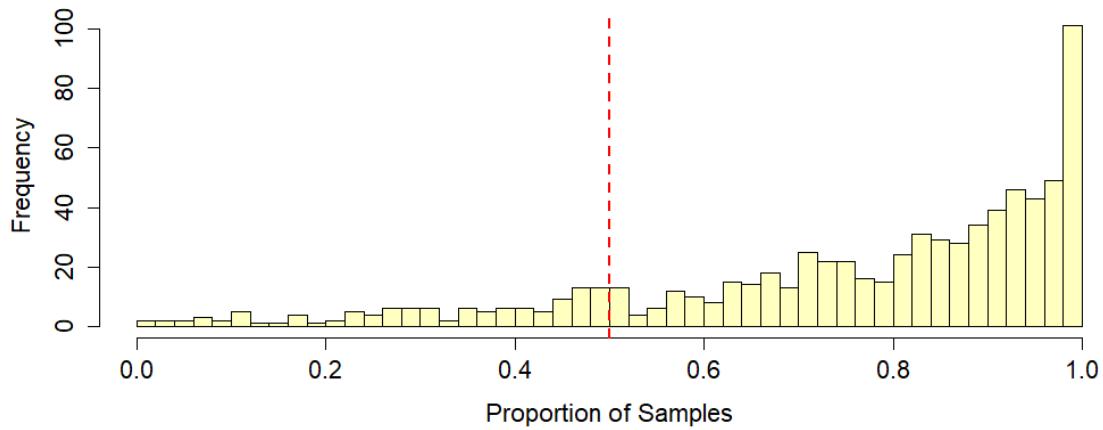
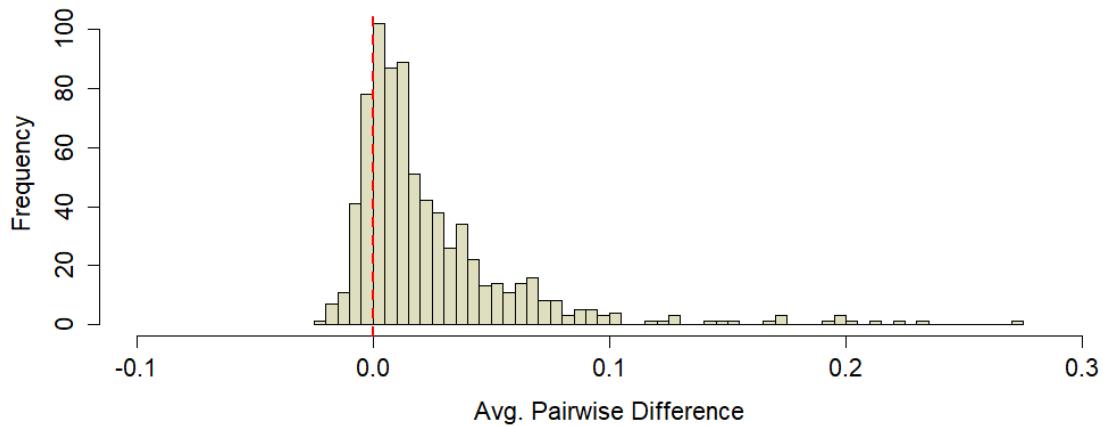


Figure 29

[GJP data]: Typical pairwise change in reliability (pre – post), due to recalibration.



Note: positive values indicate an improvement (reduction) in reliability.

Table 27

[GJP data]: Typical effects of credibility-based recalibration on reliability, summarized across forecasters.

Outcome Measure	Between-Subjects Summary Statistics				
	Mean	Mdn.	SD	Min.	Max.
Proportion of samples in which recalibration improved <i>reliability</i>	77%	84%	23%	1%	100%
Typical pairwise change in <i>reliability</i> (pre - post), due to recalibration	0.02	0.01	0.04	-0.02	0.27

Table 28

[GJP data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced reliability) beyond chance.

Outcome Measure	$H_0$	Prop. Mass $> H_0$	Stat. ( $V$ )	p-value
Proportion of samples in which recalibration improved <i>reliability</i>	Mdn. = 0.5	84%	$2.60 \times 10^5$	<.001***
Typical pairwise change in <i>reliability</i> (pre - post), due to recalibration	Mdn. = 0	82%	$2.61 \times 10^5$	<.001***

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

**Discussion.** Much like the other findings in Study 3a (GJP recalibration), the results of Analysis 3a.iii demonstrate that credibility-based recalibration typically had a beneficial effect on judgments in the GJP. Indeed, recalibration could be expected to improve (reduce) *reliability* in 77% of samples for the average GJP forecaster and in 84% of samples for the median GJP forecaster. Though less interpretable than measures of

absolute error, this was accompanied by an average pairwise improvement (reduction) of 0.02 in *reliability* across samples. In the case of both *reliability*-based outcome measures, the typical effects of recalibration were better than chance for more than 80% of forecasters, both of which corresponded to statistically significant effects. In general, therefore, the results of Analysis 3a.iii provide evidence to suggest that credibility-based recalibration did not improve judgmental accuracy simply because it made SPJs more extreme, but instead because it genuinely improved forecast calibration.

## General Discussion

In nearly all cases, the results of Study 3a demonstrate that credibility-based recalibration can identify and ameliorate errors and biases in GJP data. As such, these results provide a clear example of a domain in which empirical models of credibility can be useful. As in Study 2a (GJP reliability/validity), it is conceivable that these results can be attributed to the unusual skill of GJP forecasters or the remarkable richness of the GJP data-set. However, a qualitative assessment suggests that this is not the case. Consider, for example, the sheer breadth of forecasters included in this sample. While it is true that *some* GJP forecasters demonstrated exceptional skill; provided a large number of forecasts; and were uncommonly motivated (etc.), this was certainly not true of *all* GJP forecasters. Despite this, the results of Study 3a demonstrate that the linear credibility framework was able to produce reliable improvements in AJE and ALE for *nearly every forecaster* ( $\geq 97\%$ , in all cases). Though the results were not quite as strong for measures of *reliability* (84% and 82% of forecasters saw positive typical effect-sizes, respectively),

these results nevertheless indicate that the usefulness of the linear credibility framework was not limited to the most exceptional participants in the GJP. Thus, the results of Study 3a provide a strong preliminary basis for arguing that the linear credibility framework — and perhaps other models of credibility — may be useful to decision makers, in general.

### **Study 3b: Empirical Effects of Recalibration with March Madness Data (MM recalibration)**

In Study 2b (MM reliability/validity), the linear credibility framework uncovered surprisingly reliable trends in participants' judgments, but exploratory analyses indicated these relationships were not consistently related to forecast accuracy. These findings present two possible explanations. The first is that bootstrapped estimates of *bias* ( $\hat{\alpha}'_i$ ) and *expertise* ( $\hat{\beta}'_i$ ) — i.e., the two measures that were not strongly related to forecast accuracy — reflected systematic departures from *estimated optima* but had little bearing on the participant's overall performance. However, for this explanation to be true, one would need to posit a rather contrived scenario in which the costs and benefits to recalibration were symmetric, such as a case in which *estimated optima* hovered around 0.50 and the outcome of each game was a toss-up. From an empirical perspective, the results of the March Madness study demonstrate that this was not true. While it is true that empirical outcomes were split 49:51 over the course of the tournament (i.e., 49% of outcomes resolved to 0 and 51% resolved to 1), this near-symmetric split in outcomes could only have created symmetric costs and benefits to recalibration if *estimated optima* were systematically less extreme (i.e., closer to 0.50) than most participants' SPJs. As reported

in Study 2b (MM reliability/validity), *estimated optima* in the March Madness tournament were more accurate (i.e., had a smaller average Brier score) than approximately 84% of participants. Thus, it is highly unlikely that bootstrapped estimates of *bias* ( $\hat{\alpha}'_i$ ) and *expertise* ( $\hat{\beta}'_i$ ) were unrelated to forecast accuracy in this study for purely artifactual reasons.

In Study 2b (MM reliability/validity), therefore, the most likely explanation for the inconsistent validity of bootstrapped credibility estimates is that *estimated optima* in the March Madness study were not particularly accurate. Thus, the degree to which an individual's SPJs resembled *estimated optima* in Study 2b was predictive of accuracy only to the extent that their judgment strategy was relatively stable — an aspect of “skill” or “proficiency” that is captured by *consistency* ( $\hat{\sigma}_i$ ). In Study 3b, this would suggest that credibility-based recalibration may succeed in shrinking an individual's SPJs towards *estimated optima*, but that it is unlikely to improve their accuracy. Or, in operational terms, a lack of predictive signal among *estimated optima* would suggest that recalibration is likely to produce a reliable improvement in AJE, but perhaps not an improvement in ALE.

If this turns out to be the case, then the linear credibility framework will fail to meet the criteria for a useful credibility model in this Study 3b. However, this will not be because linear regression is a poor descriptive model (a possibility that would be indicated by little improvement in *both* AJE and ALE). Instead, an empirical split between ALE and AJE in Study 3b would suggest that the use of linear regression is sound (at least to a degree), but that my faith in the focal set of *estimated optima* was not

— largely because prediction in this study was such a difficult task. To distinguish between these possibilities, Study 3b will examine the fit of the linear credibility framework to data from the March Madness study by examining its performance against the three criteria of a useful model. If the linear credibility framework does not fulfil any of the three criteria in Study 3b, then linear regression may not be a useful tool for examining credibility in this domain (and thus, may not be useful, in general); if it fulfills only the first criterion (stable/generalizable relationship with *estimated optima*), then linear regression may be a generally useful tool, but one that can backfire when fit to a poor set of *estimated optima*; and finally, if it fulfils two or more criteria, then it may be a more useful tool than the results of Study 2b (MM reliability/validity) suggest.

## Method

**Detailed procedure.** Study 3b followed the same procedure as Study 3a (GJP recalibration) — excepting, of course, that it focused on March Madness data rather than forecasts from the GJP. As described in the Detailed Procedure section of Study 2b (MM reliability/validity), credibility functions in this analysis were fit to a calibration sample size of  $n_{\text{cal}} = 50$ ; the minimum prediction sample size was set to  $n_{\text{pred}} = 1$  to maximize statistical power; and the effects of recalibration were observed over  $n_{\text{boot}} = 100$  bootstrap trials.

## Results

**Analysis 3b.i: typical effects of recalibration on AJE (MM recal., AJE).** In Analysis 3b.i, I examined the typical effects of recalibration on absolute judgment error (AJE). As before, the effects of recalibration on AJE were used to draw inferences about the stability and generalizability of the observed relationship between an individual's judgments and *estimated optima* (i.e., the first criterion of a useful credibility model). If credibility-based recalibration can reliably improve AJE in out-of-sample judgments, then this would suggest that the errors and biases observed in the calibration sample were similar to those that remained unobserved in the prediction sample — i.e., that these relationships were stable and generalizable. To determine whether this was true in the March Madness data, Analysis 3b.i examined the typical effects of recalibration on AJE, across forecasters.

The results of Analysis 3b.i can be seen in the figures and tables below. Figures 30-33 show the empirical distributions of the four AJE-related DVs across forecasters, and Figure 34 shows a visual comparison of mean AJEs before and after recalibration. Table 29 provides descriptive statistics for each of the distributions represented in Figures 30-33, and Table 30 shows the results of Wilcoxon signed-rank tests examining the null hypothesis that the median of each distribution does not differ from chance (represented by red dotted lines in Figures 30-33).

Figure 30

[MM data]: Typical proportion of judgments for which recalibration improved (reduced) AJE.

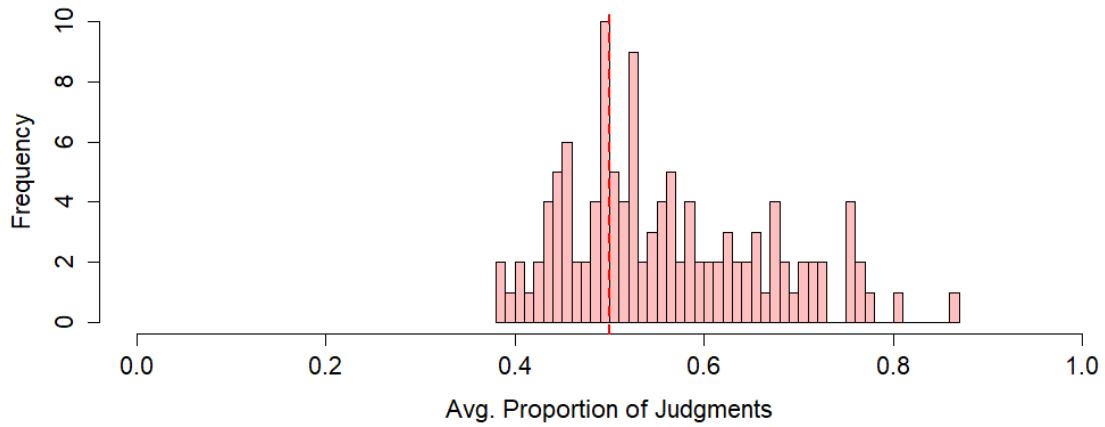
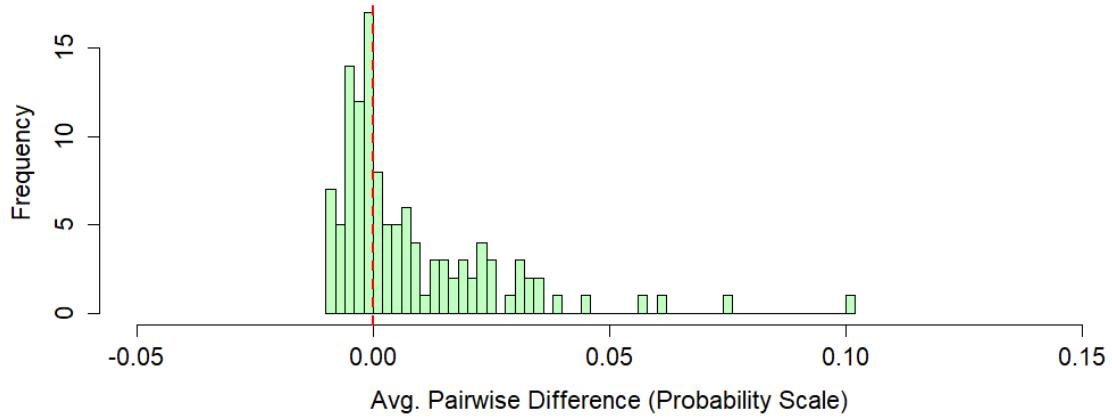


Figure 31

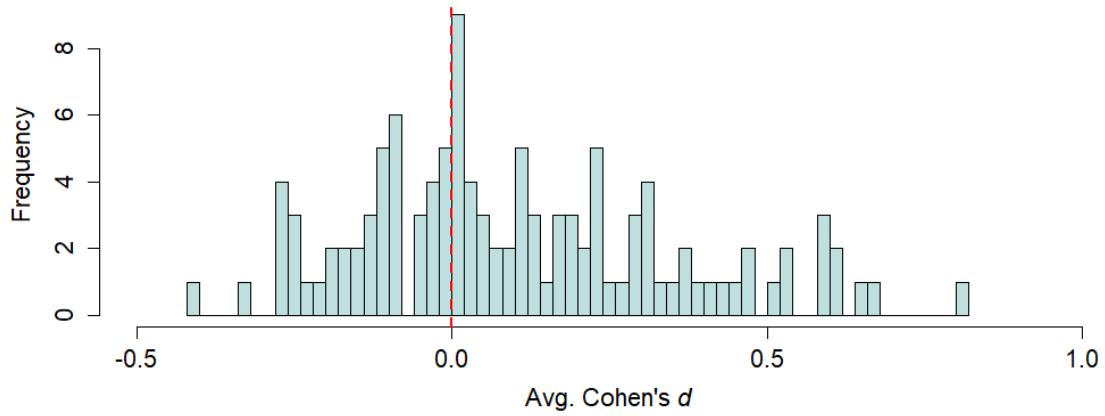
*[MM data]: Typical pairwise change in AJE (pre – post), due to recalibration.*



*Note:* positive values indicate an improvement (reduction) in AJE.

Figure 32

*[MM data]: Typical effect-size (Cohen's d) of recalibration on AJE.*



Note: positive values indicate an improvement (reduction) in AJE.

Figure 33

[MM data]: Proportion of samples in which recalibration improved (reduced) mean AJE.

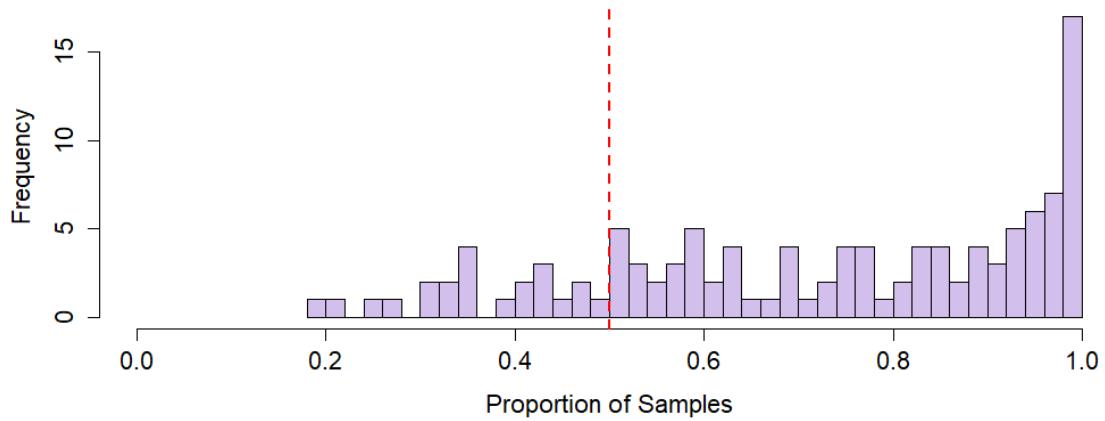
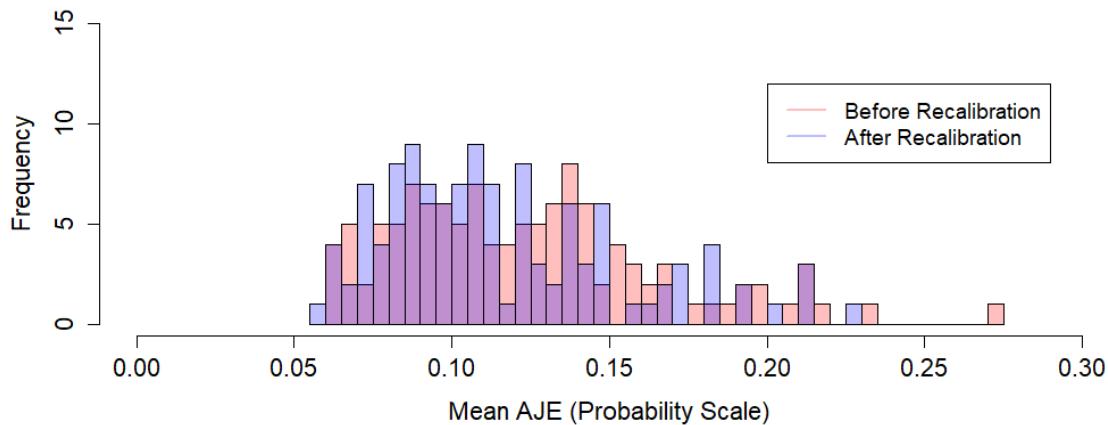


Figure 34

[MM data]: Mean AJE, before and after recalibration.



*Note:* smaller values indicate more accurate judgements (smaller errors), on average.

Table 29

*[MM data]: Typical effects of credibility-based recalibration on AJE, summarized across forecasters.*

Outcome Measure	Between-Subjects Summary Statistics				
	Mean	Mdn.	SD	Min.	Max.
Typical proportion of judgments for which recalibration improved AJE	56%	53%	10%	38%	86%
Typical pairwise change in AJE (pre - post), due to recalibration	0.79 × 10 <sup>-2</sup>	0.05 × 10 <sup>-2</sup>	1.79 × 10 <sup>-2</sup>	-1.00 × 10 <sup>-2</sup>	10.18 × 10 <sup>-2</sup>
Typical effect (Cohen's <i>d</i> ) of recalibration on AJE	0.10	0.05	0.25	-0.42	0.80
Proportion of samples in which recalibration improved mean AJE	73%	77%	23%	18%	100%

Table 30

*[MM data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced AJE) beyond chance.*

Outcome Measure	$H_0$	$Prop. Mass > H_0$	Stat. ( $V$ )	p-value
Typical proportion of judgments for which recalibration improved AJE	Mdn. = 0.5	65%	5391	<.001***
Typical pairwise change in AJE (pre - post), due to recalibration	Mdn. = 0	53%	4732	0.01**
Typical effect (Cohen's $d$ ) of recalibration on AJE	Mdn. = 0	63%	4652	<.001***
Proportion of samples in which recalibration improved mean AJE	Mdn. = 0.5	81%	6236.50	<.001***

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

**Discussion.** The results of Analysis 3b.i suggest that credibility-based recalibration typically led to small, but consistent improvements in AJE. Depending on how AJE was summarized, the study-wide expected effects of recalibration (i.e., the typical effects of recalibration for the average March Madness forecaster) were to reduce AJE in 56% of individual judgments, each by an average of 0.79 points on the probability scale. Overall, this corresponded to an average, pairwise Cohen's  $d$  of 0.10, and a smaller mean AJE in 73% of bootstrap samples, on average. Though considerably less impressive than the effects observed in the Good Judgment Project, the medians of the distributions for all four outcome measures were significantly greater than chance, suggesting that credibility-based recalibration would have generally improved AJE in the March Madness study. Thus, while the results of Analysis 3b.i are small, they nevertheless indicate that the linear credibility framework was able to identify stable, generalizable relationships in the March Madness data — a rather impressive feat, in and of itself.

**Analysis 3b.ii: typical effects of recalibration on ALE (MM recal., ALE).** In Analysis 3b.ii, I examined the typical effects of recalibration on absolute linear error (ALE) in the March Madness data. As with the previous study, a reliable improvement in ALE due to recalibration would suggest that the trends observed in Analysis 3b.i (MM recal., AJE) represent genuine errors and biases in judgment and that correcting them would systematically lead to increased accuracy. As discussed at the outset of Study 3b (MM recalibration), however, there is reason to suspect that *estimated optima* in the March Madness study were not particularly accurate as “model” judgments. As such, shrinking an individual’s SPJs towards *estimated optima* in Analysis 3b.ii may not result in a reliable increase in ALE. If this turns out to be the case, then the linear credibility framework would not meet the criteria for a useful model of credibility when applied to the March Madness data. Because Analysis 3b.i (MM recal., AJE) revealed a consistent effect of recalibration on AJE, however, this would suggest that the principal reason the linear credibility framework was not useful was because *estimated optima* were a poor stand-in for “objective” judgments in the March Madness study, not because linear regression was too simple of a model to capture credibility.

To better tease-apart these possibilities, Analysis 3b.ii examined the typical effects of credibility-based recalibration on ALE, across forecasters. The results of this analysis can be seen in the figures and tables below. Figures 35-38 show the empirical distributions of the four ALE-related DVs across forecasters, and Figure 39 shows a visual comparison of mean ALEs before and after recalibration. Table 31 provides

descriptive statistics for each of the distributions represented in Figures 35-38, and Table 32 shows the results of Wilcoxon signed-rank tests examining the null hypothesis that the median of each distribution does not differ from chance (represented by red dotted lines in Figures 35-38).

Figure 35

*[MM data]: Typical proportion of judgments for which recalibration improved (reduced) ALE.*

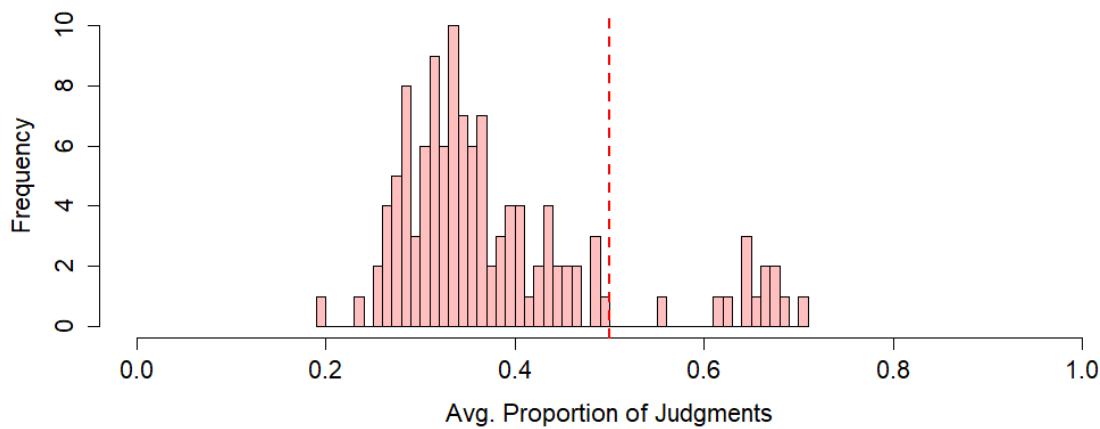
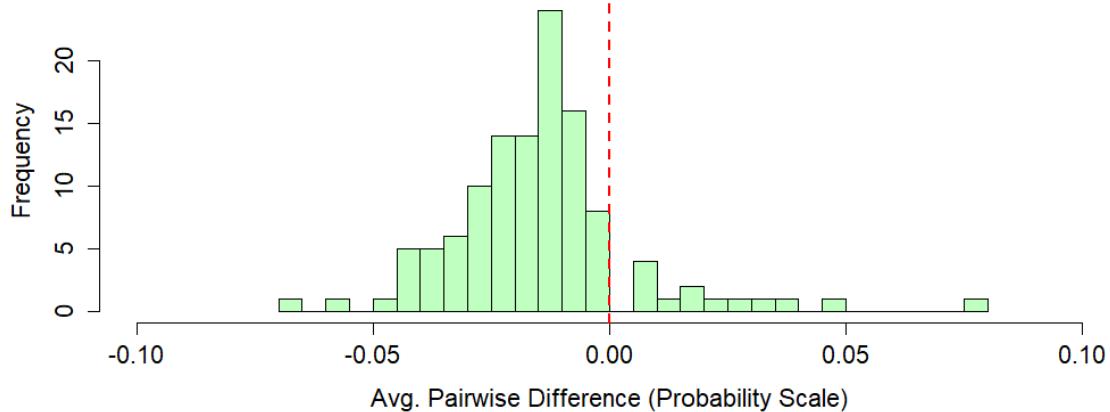


Figure 36

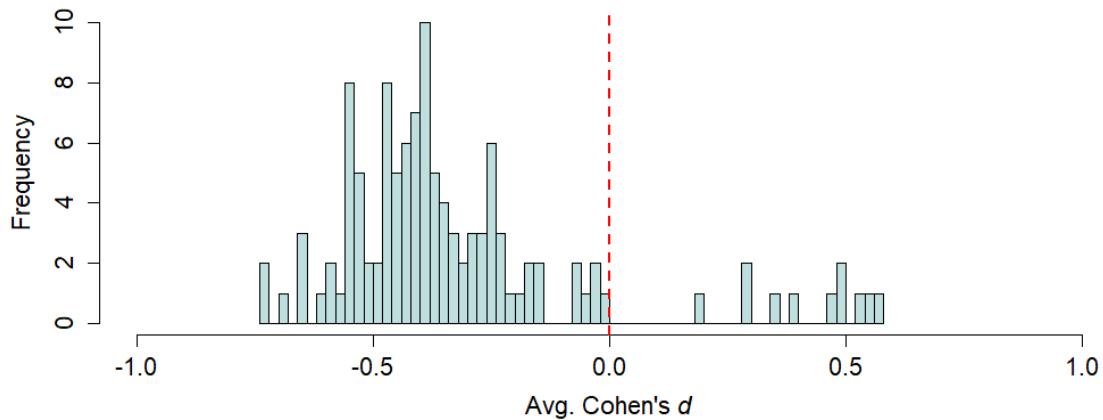
*[MM data]: Typical pairwise change in ALE (pre – post), due to recalibration.*



Note: positive values indicate an improvement (reduction) in ALE.

Figure 37

[MM data]: Typical effect-size (Cohen's  $d$ ) of recalibration on ALE.



Note: positive values indicate an improvement (reduction) in ALE.

Figure 38

[MM data]: Proportion of samples in which recalibration improved (reduced) mean ALE.

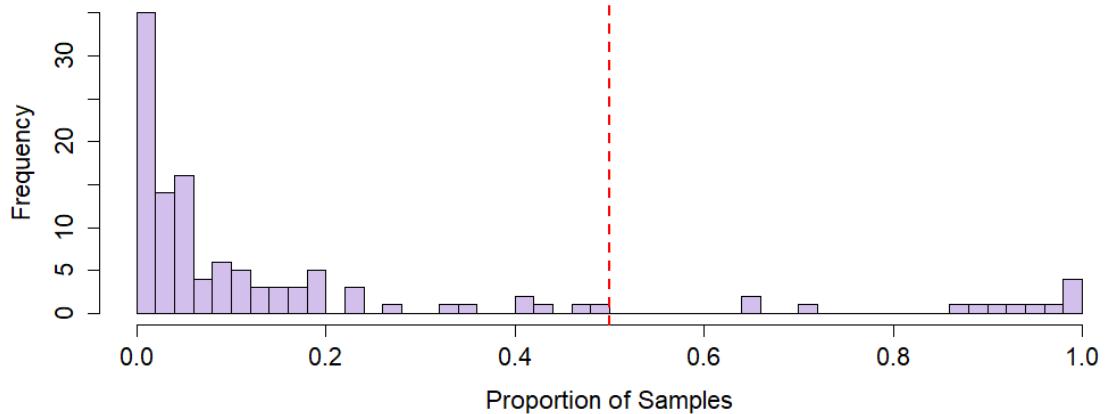
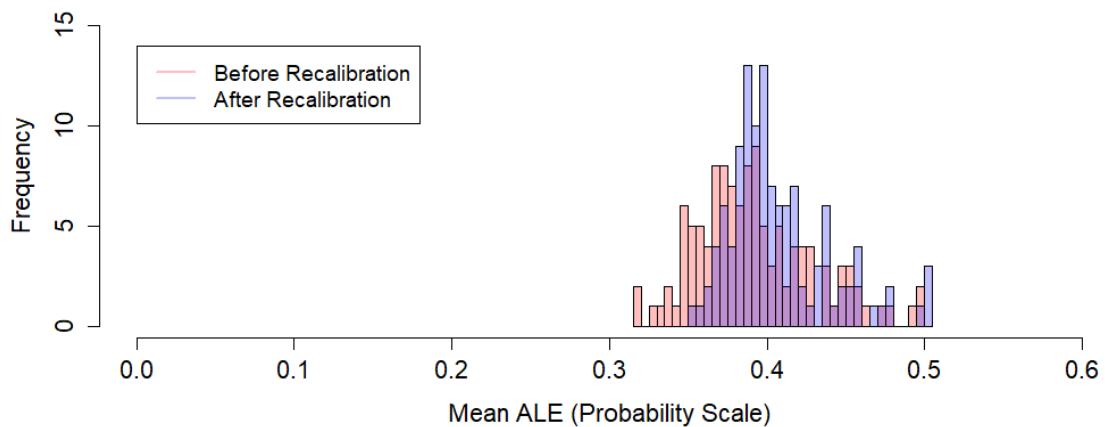


Figure 39

*[MM data]: Mean ALE, before and after recalibration.*



Note: smaller values indicate more accurate judgements (smaller errors), on average.

Table 31

*[MM data]: Typical effects of credibility-based recalibration on ALE, summarized across forecasters.*

Outcome Measure	Between-Subjects Summary Statistics				
	Mean	Mdn.	SD	Min.	Max.

Typical proportion of judgments for which recalibration improved ALE	38%	35%	11%	19%	71%
Typical pairwise change in ALE (pre - post), due to recalibration	-1.48 $\times 10^{-2}$	-1.46 $\times 10^{-2}$	1.93 $\times 10^{-2}$	-6.66 $\times 10^{-2}$	7.51 $\times 10^{-2}$
Typical effect (Cohen's $d$ ) of recalibration on ALE	-0.32	-0.39	0.28	-0.73	0.56
Proportion of samples in which recalibration improved mean ALE	18%	6%	27%	0%	100%

Table 32

[MM data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced ALE) beyond chance.

Outcome Measure	$H_0$	Prop. Mass $> H_0$	Stat. ( $V$ )	p-value
Typical proportion of judgments for which recalibration improved ALE	Mdn. = 0.5	11%	726.5	<.001***
Typical pairwise change in ALE (pre - post), due to recalibration	Mdn. = 0	11%	829	<.001***
Typical effect (Cohen's $d$ ) of recalibration on ALE	Mdn. = 0	10%	704	<.001***
Proportion of samples in which recalibration improved mean ALE	Mdn. = 0.5	11%	737.5	<.001***

Significance levels: \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$

Note: p-values in red indicate that effects were not in the expected direction

**Discussion.** Somewhat unexpectedly — though generally in-line with the concerns outlined at the beginning of this study — the results of Analysis 3b.ii indicate that credibility-based recalibration typically had a negative effect on (i.e., increased)

ALE. This conclusion was supported by all four ALE-based outcome measures, with the average GJP forecaster being able to expect only 38% of her individual judgments to improve in terms of ALE, and the average size of her errors to increase by 1.48 points on the probability scale. Overall, this corresponded to an average, pairwise Cohen's  $d$  of -0.32, and a worse (larger) mean ALE in 82% of a forecaster's samples, on average. Across the board, distributions of typical effect sizes were significantly worse than chance, suggesting that credibility-based recalibration would have decreased accuracy in the March Madness study, in general. As such, the results of Analysis 3b.ii suggest that the linear credibility framework would not have been a useful model for improving March Madness predictions, primarily because *estimated optima* were not particularly accurate (relative to individuals).

Despite this disappointing performance, however, the results of Analysis 3b.ii suggest that the linear credibility framework was statistically powerful enough to pick-up on small trends in individual judgments. Consequently, it is still conceivable that the linear credibility framework might be of some use, if a better standard of judgment can be identified. Thus, in Chapter 4, I will return to the March Madness data to determine whether it is possible to salvage the benefits of credibility-based recalibration by using a more accurate set of *estimated optima*.

**Analysis 3b.iii: typical effects of recalibration on *reliability* (MM recal., *reliability*).** In Analysis 3b.iii, I examined the typical effects of recalibration on *reliability*, a measure that is closely related to *calibration* in the forecasting literature

(see: Lichtenstein et al., 1982). As with ALE, a systematic improvement in *reliability* due to recalibration would suggest that shrinking SPJs towards *estimated optima* reduced errors and biases in judgment. Critically, while the results of Analysis 3b.ii (MM recal., ALE) have already shown that credibility-based recalibration was generally detrimental to forecast accuracy in the March Madness data, these results cannot speak to whether that accuracy was lost because judgments were made “truly” worse, or simply less extreme. Examining *reliability* can help disambiguate this effect because it measures the (dis)agreement between SPJs and within-sample baserates, rather than observed outcomes. As such, a consistent improvement in *reliability* paired with a consistent worsening of ALE would suggest that part of the reason *estimated optima* were less accurate than 16% of individual forecasters in the March Madness study was that those forecasters benefitted from short-term overconfidence. As such, observing the predictions of the same forecasters over a larger number of events might have seen the top performers regress to the mean, causing *estimated optima* to become (relatively) more useful.

To determine whether this was the case, Analysis 3a.iii examined the typical effects of credibility-based recalibration on *reliability*, across forecasters. The results of this analysis can be seen in the figures and tables below. Figures 40 and 41 show the empirical distributions of the two *reliability*-related DVs across forecasters. Table 33 provides descriptive statistics for each of the distributions represented in Figures 40 and 41, and Table 34 shows the results of Wilcoxon signed-rank tests examining the null

hypothesis that the median of each distribution does not differ from chance (represented by red dotted lines in Figures 40 and 41).

Figure 40

*[MM data]: Proportion of samples in which recalibration improved (reduced) reliability.*

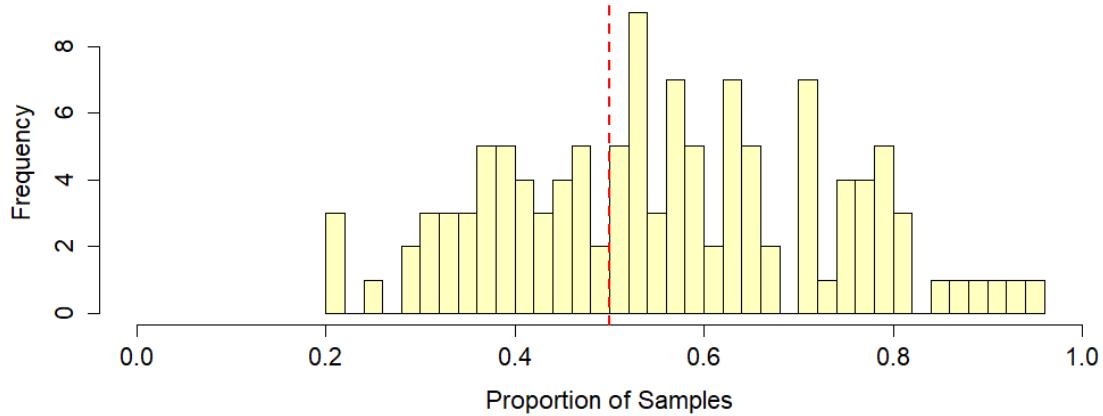
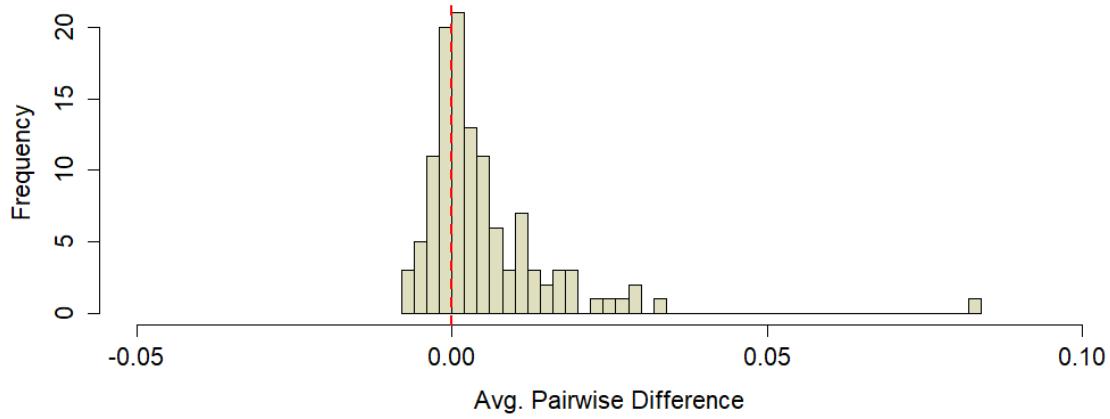


Figure 41

*[MM data]: Typical pairwise change in reliability (pre – post), due to recalibration.*



Note: positive values indicate an improvement (reduction) in reliability.

Table 33

[MM data]: Typical effects of credibility-based recalibration on reliability, summarized across forecasters.

Outcome Measure	Between-Subjects Summary Statistics				
	Mean	Mdn.	SD	Min.	Max.
Proportion of samples in which recalibration improved <i>reliability</i>	56%	56%	17%	20%	95%
Typical pairwise change in <i>reliability</i> (pre - post), due to recalibration	0.49 $\times 10^{-2}$	0.19 $\times 10^{-2}$	1.08 $\times 10^{-2}$	-0.73 $\times 10^{-2}$	8.25 $\times 10^{-2}$

Table 34

[MM data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced reliability) beyond chance.

Outcome Measure	$H_0$	Prop. Mass $> H_0$	Stat. ( $V$ )	p-value
Proportion of samples in which recalibration improved <i>reliability</i>	Mdn. = 0.5	64%	4754	<.001***
Typical pairwise change in <i>reliability</i> (pre - post), due to recalibration	Mdn. = 0	67%	5396	<.001***

Significance levels: \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$

**Discussion.** Though the effects are quite small, the results of Analysis 3b.iii indicate that credibility-based recalibration had a systematically beneficial effect on *reliability* in the March Madness data. For the average participant, recalibration was expected to improve (reduce) *reliability* in 56% of samples, and the average pairwise change in *reliability* was positive. In both cases, the typical effects of recalibration were positive for about 65% of forecasters, corresponding to statistically significant effects. In

general, therefore, the results of Analysis 3a.iii indicate that credibility-based recalibration genuinely improved calibration (albeit to a small degree). When considered in conjunction with the results of Analysis 3b.ii (MM recal., ALE), this suggests that some participants in the March Madness study may have outperformed *estimated optima* due to short-term overconfidence — an effect that helps to explain why *estimated optima* did not *appear* to be optimal in the March Madness study, as measured by forecast accuracy.

## General Discussion

As in Study 2b (MM reliability/validity), a provisional glance at the results of Study 3b suggests that the linear credibility framework may not have been worth examining in the March Madness study. Indeed, as defined by my own criteria for a useful model of credibility, the results of Study 3b indicate that the linear credibility framework was *not useful* for improving March Madness SPJs. In practice, however, this does not mean that the results of Study 3b are not informative. Instead, when considered in a broader context, the pattern of findings uncovered in Study 3b suggests that the linear credibility framework was not useful for a very predictable reason — namely that I failed to identify a strong set of *estimated optima*. Unsurprisingly, therefore, building a model of credibility around a weak set of “model” judgments did not reveal much about “good” forecasting in the March Madness study and served as a poor basis for improving SPJs.

Despite this shortcoming, however, it is worth noting that the linear credibility framework still performed admirably in Study 3b. Indeed, even when applied to a

situation that systematically undermined its chances for success, Analysis 3b.i (MM recal., AJE) demonstrates that the linear credibility framework was able to identify stable, generalizable trends in participants' judgments. In addition, Analysis 3b.iii (MM recal., *reliability*) indicates that — above and beyond the fact that March Madness games are notoriously hard to predict — one of the reasons that *estimated optima* were so poor in this study (or, rather *seemed* so poor) is that some forecasters benefited from short-term overconfidence. As such, the results of Study 3b suggest that there is still hope for the linear credibility framework, despite its limitations and real-world boundary conditions. To explore these boundary conditions further, I will return to the March Madness data in Chapter 4 to see if the usefulness of the linear credibility framework can be improved by fitting it to a more accurate set of *estimated optima*.

### **Study 3c: Empirical Effects of Recalibration with Philadelphia Air Temperature**

#### **Data (PHL recalibration)**

Much as in Chapter 2, the first two studies presented in Chapter 3 occupy opposite ends of the “usefulness” spectrum. In Study 3a (GJP recalibration), I tested the usefulness of the linear credibility framework when fit to the remarkably rich data-set provided by the Good Judgment Project; in Study 3b (MM recalibration), I examined the surprisingly spartan data gathered during the 2017 March Madness tournament. Though both studies provide a great deal of information about the capabilities of the linear credibility framework, neither can be called representative, *per se*. Thus, in Study 3c, I examined the usefulness of the linear credibility framework when applied to a thoroughly

ordinary data-set: judgments from the Philadelphia air temperature study. As discussed in Study 2c (PHL reliability/validity), exploratory regression analyses revealed a promising but somewhat haphazard degree of predictive validity between linear credibility estimates and forecast accuracy in these data — as one might expect from SPJs provided by truly amateur judges.

As in previous studies, I tested these data against the three criteria of a useful credibility model by examining the effects of credibility-based recalibration on out-of-sample judgments. If “undoing” the trends in judgment observed in a participant’s calibration sample systematically improves judgments in her prediction sample, then these trends must reflect a set of stable and generalizable relationships with *estimated optima* (criterion 1). Furthermore, by manipulating how “improvements” are defined, the effects of credibility-based recalibration can provide insight into the type of relationship that existed between an individual’s judgments and *estimated optima* in the first place. In Study 3c, this flexible testing structure was important because it allowed me to examine a set of outcome measures that were not available in previous studies. Specifically, because detailed historical records of Philadelphia air temperatures are widely available, building a statistical model that predicts one day’s air temperature from the previous day’s temperature is a relatively simple matter. Consequently, in Study 3c, I was able to test the effects of recalibration on *three* types of judgmental accuracy: AJE vs. crowd aggregates (previously just AJE; criterion 1: stable/generalizable trends); AJE vs. baserates (criteria 1 & 2: stable, generalizable, and detrimental trends); and the existing standard of ALE (criteria 1 & 2: stable, generalizable, and detrimental trends).

## Method

**Detailed procedure.** To examine the usefulness of the linear credibility framework when applied to the Philadelphia air temperature data, Study 3c followed a procedure that was largely parallel to Studies 3a (GJP recalibration) and 3b (MM recalibration). As mentioned above, however, the one major difference was that the existence of detailed historical records made it possible to estimate baserates for various temperatures observed in Philadelphia during January and July. Thus, in Study 3c, I examined four additional outcome measures for each participant, corresponding to “AJE vs. baserates” variants of the four AJE-based outcome measures that were used in previous studies (here, referred to as “AJE vs. crowd aggregates”).

To estimate baserates in Study 3c, I constructed two linear regression models: one for January and one for July. Using daily air temperature data recorded at the Philadelphia International Airport from 2008-2017, these models were constructed by regressing the high temperature from each day  $t$  on the high temperature from each day  $t + 1$  in January and July (excluding  $t = 31$  for both months, where the temperature on day  $t + 1$  would fall in the next month). After estimating these models, it was possible to estimate the probability of observing a temperature  $q$  the day after observing a temperature  $p$  by evaluating the likelihood of drawing a temperature at least as large (or at least as small) as  $q$  from a Normal distribution with a mean equal to the model’s fitted value for  $q(p)$  and a standard deviation equal to the model’s standard error of the estimate.

Beyond examining AJE vs. baserates, Study 3c was conducted in a manner similar to previous studies in Chapter 3. Thus, the dependent variables in Study 3c were observed during the same run of the general procedure that was used to calculate bootstrapped credibility estimates in Analysis 2c.ii (PHL validity). As described in Study 2c (PHL reliability/validity), credibility functions in this analysis were fit to a calibration sample size of  $n_{cal} = 50$ ; participants were only included if they had provided enough SPJs to accommodate a minimum prediction sample size of  $n_{pred} = 30$ ; and the effects of recalibration were observed over  $n_{boot} = 100$  bootstrap trials.

## Results

**Analysis 3c.i: typical effects of recalibration on AJE vs. crowd aggregates (PHL recal., AJE vs. crowd).** In Analysis 3c.i, I examined the typical effects of recalibration on absolute judgment error relative to *estimated optima* (i.e., AJE vs. crowd aggregates, or what was previously referred to as AJE). As before, the effects of recalibration on AJE vs. crowd aggregates were used to draw inferences about the stability and generalizability of trends captured by linear credibility estimates (i.e., the first criterion of a useful credibility model).

The results of this analysis can be seen in the figures and tables below. Figures 42-45 show the empirical distributions of the four AJE-vs.-crowd-aggregates-related DVs across forecasters, and Figure 46 shows a visual comparison of mean AJEs vs. crowd aggregates before and after recalibration. Table 35 provides descriptive statistics for each of the distributions represented in Figures 42-45, and Table 36 shows the results of

Wilcoxon signed-rank tests examining the null hypothesis that the median of each distribution does not differ from chance (represented by red dotted lines in Figures 42-45).

Figure 42

[PHL data]: Typical proportion of judgments for which recalibration improved (reduced) AJE vs. crowd aggregates.

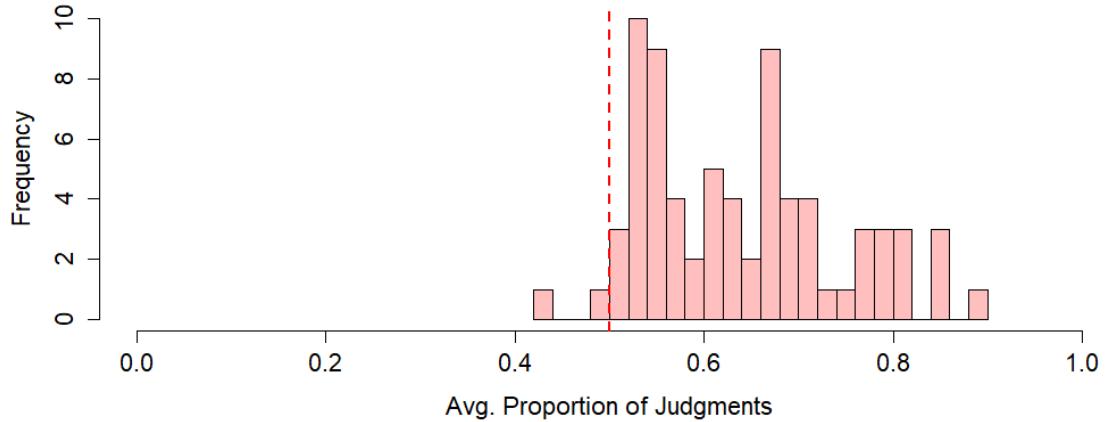
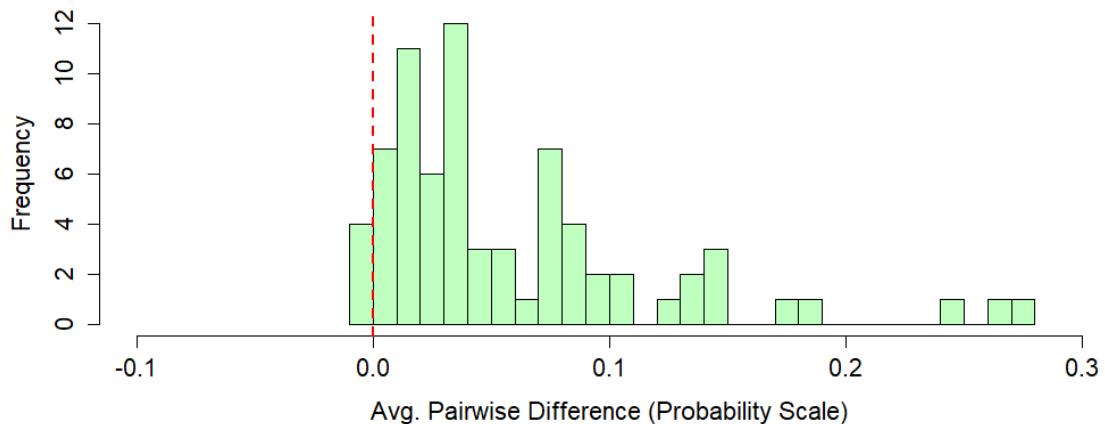


Figure 43

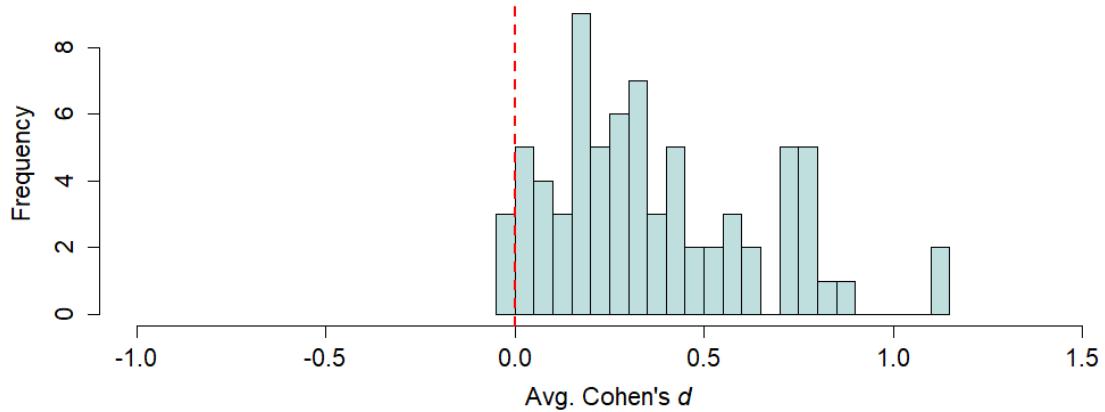
[PHL data]: Typical pairwise change in AJE vs. crowd aggregates (pre – post), due to recalibration.



Note: positive values indicate an improvement (reduction) in AJE.

Figure 44

[PHL data]: Typical effect-size (Cohen's  $d$ ) of recalibration on AJE vs. crowd aggregates.



Note: positive values indicate an improvement (reduction) in AJE.

Figure 45

[PHL data]: Proportion of samples in which recalibration improved (reduced) mean AJE vs. crowd aggregates.

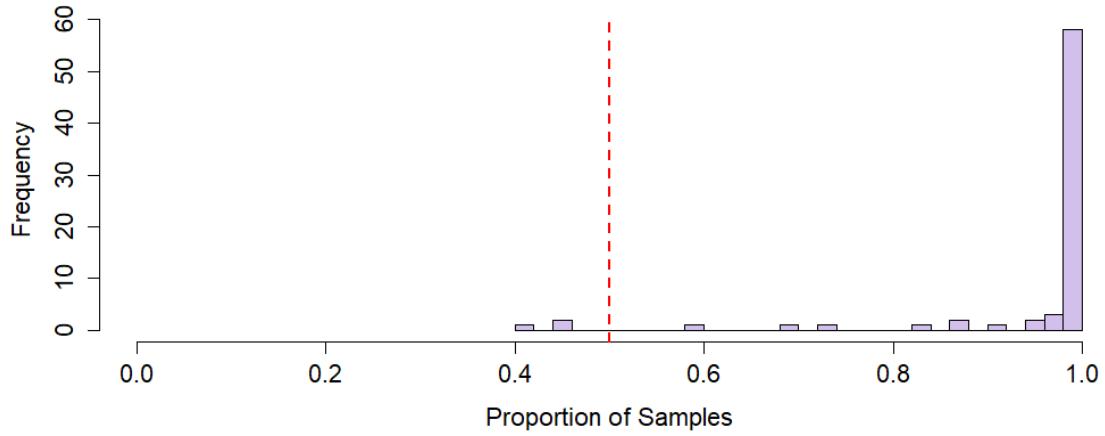
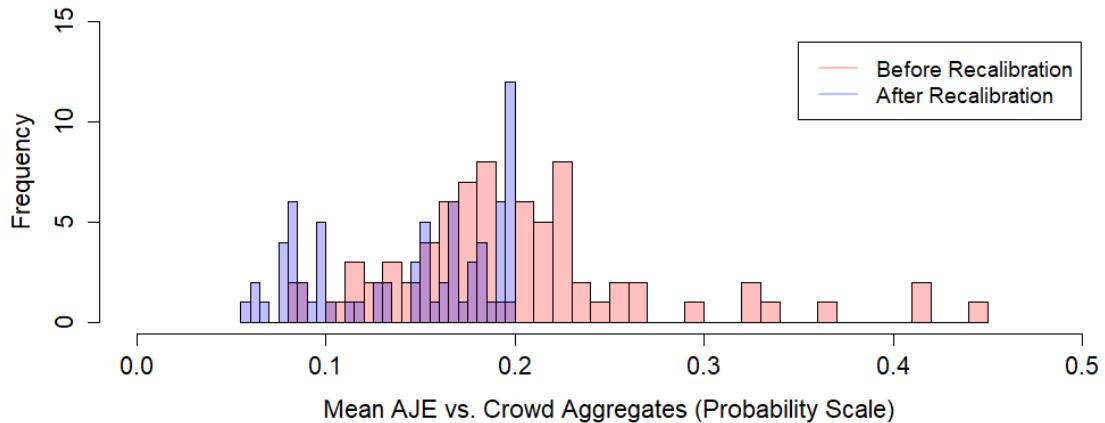


Figure 46

[PHL data]: Mean AJE vs. crowd aggregates, before and after recalibration.



*Note:* smaller values indicate more accurate judgements (smaller errors), on average.

Table 35

*[PHL data]: Typical effects of credibility-based recalibration on AJE vs. crowd aggregates, summarized across forecasters.*

Outcome Measure	Between-Subjects Summary Statistics				
	Mean	Mdn.	SD	Min.	Max.
Typical proportion of judgments for which recalibration improved AJE vs. crowd agg.	64%	62%	11%	43%	90%
Typical pairwise change in AJE vs. crowd agg. (pre - post), due to recalibration	5.99 $\times 10^{-2}$	3.40 $\times 10^{-2}$	6.16 $\times 10^{-2}$	-0.27 $\times 10^{-2}$	27.06 $\times 10^{-2}$
Typical effect (Cohen's $d$ ) of recalibration on AJE vs. crowd agg.	0.37	0.31	0.28	-0.02	1.13
Prop. of samples in which recalibration improved mean AJE vs. crowd agg.	95%	100%	13%	41%	100%

Table 36

[PHL data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced AJE vs. crowd aggregates) beyond chance.

Outcome Measure	$H_0$	Prop. Mass $> H_0$	Stat. ( $V$ )	p-value
Typical proportion of judgments for which recalibration improved AJE vs. crowd agg.	Mdn. = 0.5	97%	2670	<.001***
Typical pairwise change in AJE vs. crowd agg. (pre - post), due to recalibration	Mdn. = 0	95%	2682	<.001***
Typical effect (Cohen's $d$ ) of recalibration on AJE vs. crowd agg.	Mdn. = 0	96%	2692	<.001***
Prop. of samples in which recalibration improved mean AJE vs. crowd agg.	Mdn. = 0.5	96%	2694	<.001***

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

**Discussion.** Though less impressive than the results of Analysis 3a.i (GJP recal., AJE), the results of Analysis 3c.i demonstrate that credibility-based recalibration once again led to a universal improvement in AJE vs. crowd aggregates (which, for the remainder of this analysis, I will simply refer to as AJE). Depending on how AJE was summarized, the expected effect of recalibration for the average study participant was to reduce AJE in 64% of individual judgments, each by an average of 5.99 points on the probability scale. Overall, this corresponded to an average, pairwise Cohen's  $d$  of 0.37, and a better (smaller) mean AJE in 95% of bootstrap samples, on average. For all four AJE-based outcomes, this corresponded to distributions of typical effect sizes that were significantly greater than chance, suggesting that credibility-based recalibration could have been expected to improve AJE in the Philadelphia air temperature study, in general.

Consequently, the results of Analysis 3a.i suggest that the linear credibility framework fulfilled the first criterion of a useful credibility model (stable, generalizable relationships) when fit to predictions from the Philadelphia air temperature study.

**Analysis 3c.ii: typical effects of recalibration on AJE vs. baserates (PHL recal., AJE vs. baserates).** In Analysis 3c.ii, I examined the typical effects of recalibration on absolute judgment error with respect to historical baserates (AJE vs. baserates). As discussed above, this category of outcome measure was unique to Study 3c, in that the Philadelphia air temperature data were the only ones that allowed for the estimation of empirical baserates. Though still just another set of *estimated optima*, these baserates represent a high-water mark the usefulness of the linear credibility framework, as they are the only criteria in Chapter 3 that can be unambiguously defended as a normative standard of “good” judgment. Thus, if credibility-based recalibration can be shown to systematically improve (reduce) AJE vs. baserates, then Analysis 3c.ii will definitively show that even simple models of credibility can be used to identify and ameliorate errors and biases in judgment (criteria 1 & 2: identify stable, generalizable, and detrimental trends in judgment).

To determine whether the Philadelphia air temperature could support this conclusion, Analysis 3c.ii examined the effects of credibility-based recalibration (with credibility functions fit to crowd aggregates as *estimated optima*) on AJE vs. baserates. The results of this analysis can be seen in the figures and tables below. Figures 47-50 show the empirical distributions of the four AJE vs. baserates DVs across forecasters, and

Figure 51 shows a visual comparison of mean AJEs vs. baserates before and after recalibration. Table 37 provides descriptive statistics for each of the distributions represented in Figures 47-50, and Table 38 shows the results of Wilcoxon signed-rank tests examining the null hypothesis that the median of each distribution does not differ from chance (represented by red dotted lines in 47-50).

Figure 47

*[PHL data]: Typical proportion of judgments for which recalibration improved (reduced) AJE vs. baserates.*

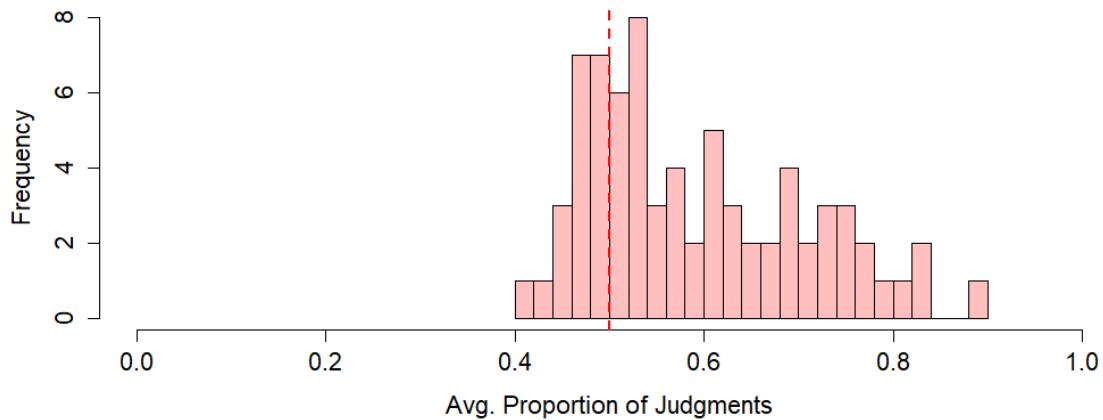
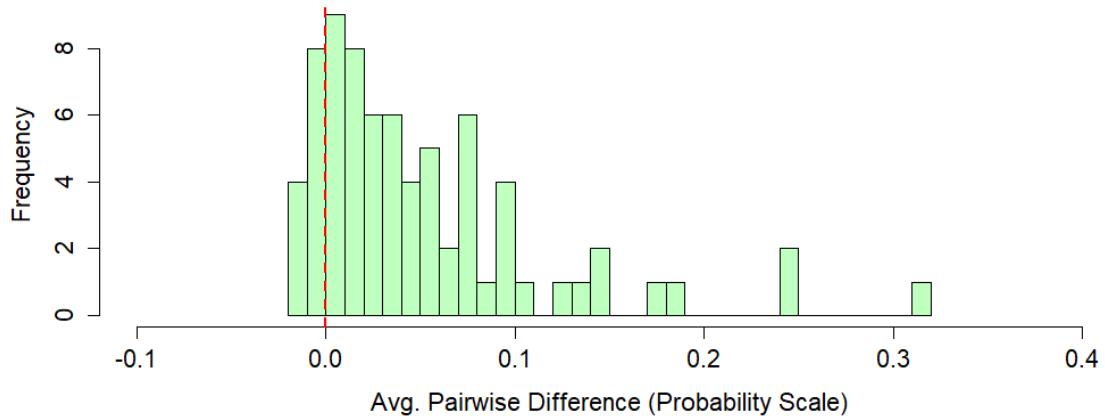


Figure 48

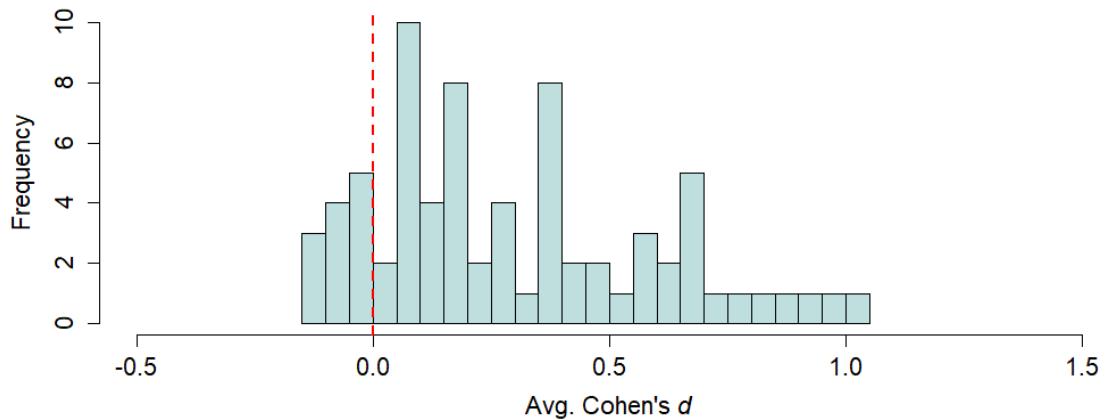
*[PHL data]: Typical pairwise change in AJE vs. baserates (pre – post), due to recalibration.*



Note: positive values indicate an improvement (reduction) in AJE.

Figure 49

[PHL data]: Typical effect-size (Cohen's  $d$ ) of recalibration on AJE vs. baserates.



Note: positive values indicate an improvement (reduction) in AJE.

Figure 50

[PHL data]: Proportion of samples in which recalibration improved (reduced) mean AJE vs. baserates.

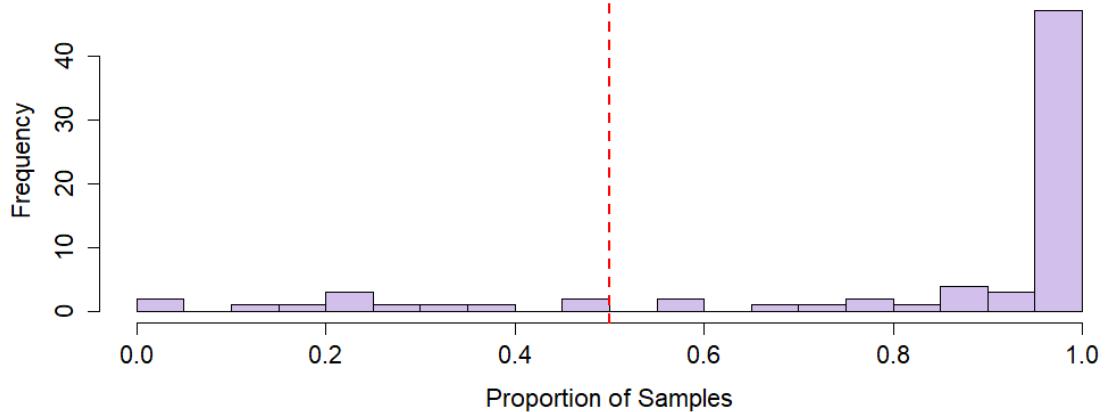
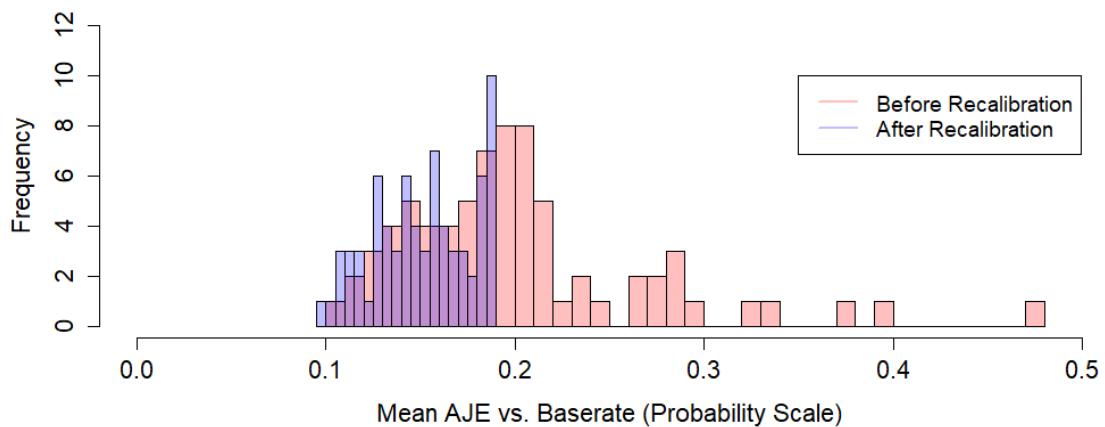


Figure 51

*[PHL data]: Mean AJE vs. baserates, before and after recalibration.*



Note: smaller values indicate more accurate judgements (smaller errors), on average.

Table 37

*[PHL data]: Typical effects of credibility-based recalibration on AJE vs. baserates, summarized across forecasters.*

Outcome Measure	Between-Subjects Summary Statistics				
	Mean	Mdn.	SD	Min.	Max.

Typical proportion of judgments for which recalibration improved AJE vs. baserates	60%	56%	11%	42%	88%
Typical pairwise change in AJE vs. baserates (pre - post), due to recalibration	$5.13 \times 10^{-2}$	$3.25 \times 10^{-2}$	$6.47 \times 10^{-2}$	$-1.16 \times 10^{-2}$	$31.88 \times 10^{-2}$
Typical effect (Cohen's $d$ ) of recalibration on AJE vs. baserates	0.30	0.21	0.30	-0.14	1.05
Prop. of samples in which recalibration improved mean AJE vs. baserates	84%	100%	28%	1%	100%

Table 38

*[PHL data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced AJE vs. baserates) beyond chance.*

Outcome Measure	$H_0$	$Prop. Mass > H_0$	Stat. ( $V$ )	p-value
Typical proportion of judgments for which recalibration improved AJE vs. baserates	Mdn. = 0.5	74%	2346	<.001***
Typical pairwise change in AJE vs. baserates (pre - post), due to recalibration	Mdn. = 0	84%	2553	<.001***
Typical effect (Cohen's $d$ ) of recalibration on AJE vs. baserates	Mdn. = 0	84%	2560	<.001***
Prop. of samples in which recalibration improved mean AJE vs. baserates	Mdn. = 0.5	84%	2490	<.001***

Significance levels: \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$

**Discussion.** Despite the fact that participants in Study 3c had only lay knowledge of air temperatures in Philadelphia, the results of Analysis 3c.ii demonstrate that a basic willingness to attend to one's errors (and a small sample of historical outcomes) is all a

decision maker needs to systematically improve SPJs. Perhaps even more remarkably, Analysis 3c.ii demonstrates that this “true” improvement in judgmental accuracy (i.e., a systematic decrease in the differences between SPJs and historical baserates) can be achieved with (a) a realistic amount of data; (b) only modest knowledge of empirical outcomes; and (c) no knowledge of “objective” probabilities, whatsoever. Thus, as anticipated, Analysis 3c.ii provides a definitive example of a case where a simple credibility model (here, the linear credibility framework) can be used to improve real-world judgments.

Specifically, the results of Analysis 3c.ii demonstrate that shrinking SPJs towards crowdsourced *estimated optima* was able to systematically improve the accuracy of SPJs, relative to historical baserates. Depending on how AJE vs. baserates was measured, the study-wide expected effects of recalibration were to reduce AJE in 60% of individual judgments, each by an average of 5.13 points on the probability scale. This corresponded to an average, pairwise Cohen’s  $d$  of 0.30, and an improvement in mean AJE in 84% of bootstrap samples, on average. Though these effects were predictably smaller than the corresponding reduction in AJE vs. crowd aggregates, all four effects on AJE vs. baserates were significantly greater than chance, suggesting that credibility-based recalibration could have been expected to improve AJE vs. baserates in Study 3c, in general. Consequently, the results of Analysis 3c.ii provide clear evidence that the linear credibility framework was able to meet the first two criteria of a useful credibility model (identifying stable, generalizable, and detrimental relationships) when fit to the Philadelphia air temperature data.

**Analysis 3c.iii: typical effects of recalibration on ALE (PHL recal., ALE).** In Analysis 3c.iii, I examined the typical effects of recalibration on absolute linear error (ALE). As with other recalibration studies, a reliable, positive effect of recalibration on ALE would indicate that the linear credibility framework can be useful for identifying (and correcting) genuine errors and biases in judgment. In addition, if the effects of recalibration on ALE are suitably large, then the argument can be made that the linear credibility framework met all three criteria of a useful credibility model — suggesting, perhaps, that “usefulness” is a common feature of empirical models of credibility.

To examine this possibility, Analysis 3c.iii examined the typical effects of credibility-based recalibration on ALE, across forecasters. The results of this analysis can be seen in the figures and tables below. Figures 52-55 show the empirical distributions of the four ALE-based DVs across forecasters, and Figure 56 shows a visual comparison of mean ALEs before and after recalibration. Table 39 provides descriptive statistics for each of the distributions represented in Figures 52-55, and Table 40 shows the results of Wilcoxon signed-rank tests examining the null hypothesis that the median of each distribution does not differ from chance (represented by red dotted lines in Figures 52-55).

Figure 52

*[PHL data]: Typical proportion of judgments for which recalibration improved (reduced) ALE.*

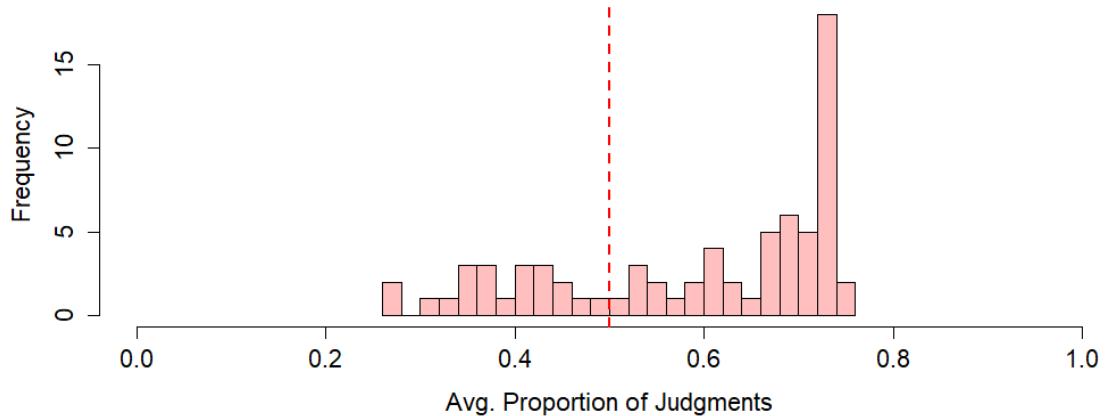
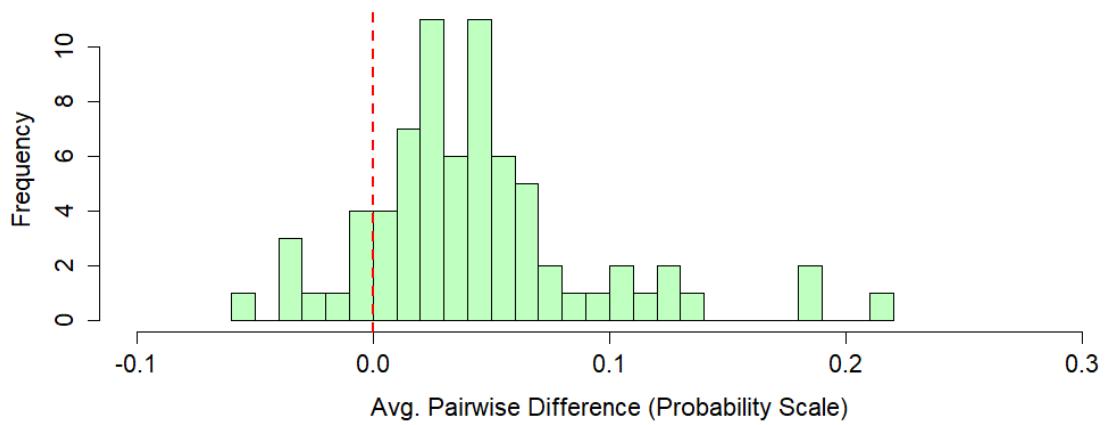


Figure 53

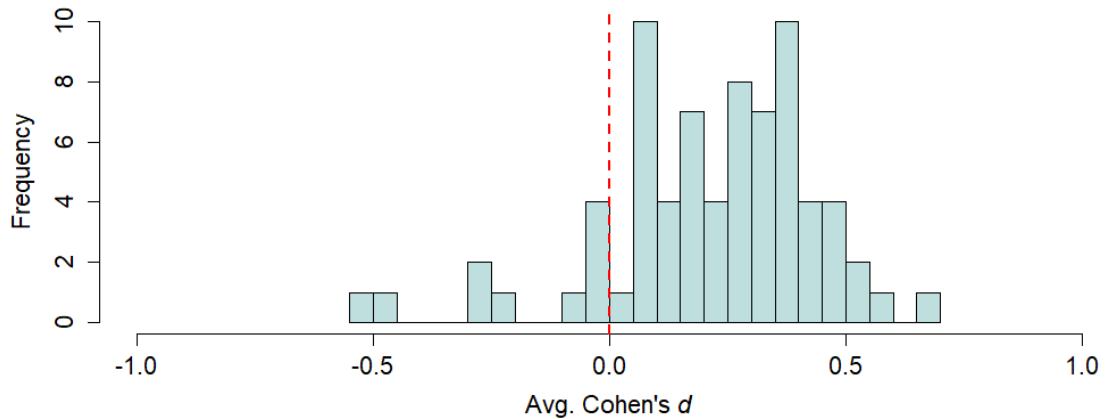
*[PHL data]: Typical pairwise change in ALE (pre – post), due to recalibration.*



Note: positive values indicate an improvement (reduction) in ALE.

Figure 54

*[PHL data]: Typical effect-size (Cohen's d) of recalibration on ALE.*



Note: positive values indicate an improvement (reduction) in ALE.

Figure 55

[PHL data]: Proportion of samples in which recalibration improved (reduced) mean ALE.

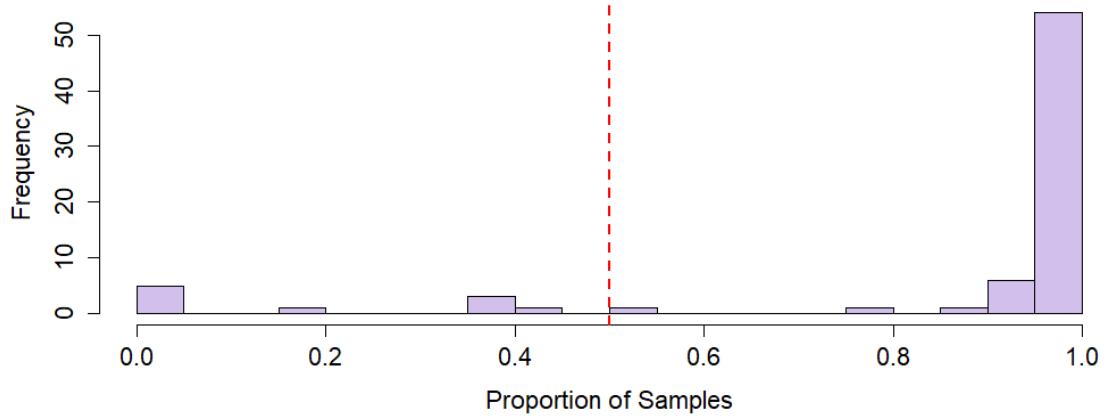
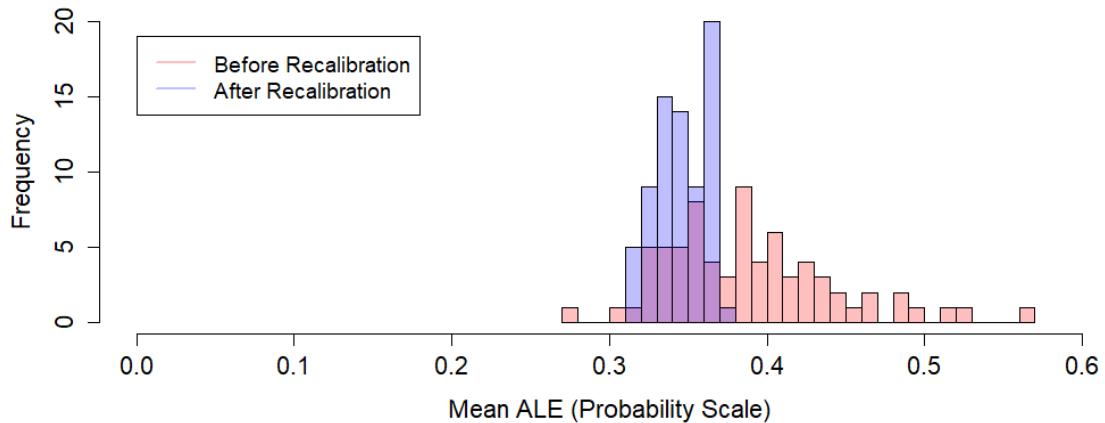


Figure 56

[PHL data]: Mean ALE, before and after recalibration.



Note: smaller values indicate more accurate judgements (smaller errors), on average.

Table 39

*[PHL data]: Typical effects of credibility-based recalibration on ALE, summarized across forecasters.*

Outcome Measure	Between-Subjects Summary Statistics				
	Mean	Mdn.	SD	Min.	Max.
Typical proportion of judgments for which recalibration improved ALE	59%	64%	15%	27%	75%
Typical pairwise change in ALE (pre - post), due to recalibration	4.37 $\times 10^{-2}$	3.55 $\times 10^{-2}$	4.88 $\times 10^{-2}$	-5.42 $\times 10^{-2}$	21.68 $\times 10^{-2}$
Typical effect (Cohen's $d$ ) of recalibration on ALE	0.21	0.27	0.22	-0.51	0.68
Proportion of samples in which recalibration improved mean ALE	87%	100%	29%	0%	100%

Table 40

*[PHL data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced ALE) beyond chance.*

Outcome Measure	$H_0$	<i>Prop.</i> $\text{Mass} > H_0$	Stat. ( $V$ )	<i>p-value</i>
Typical proportion of judgments for which recalibration improved ALE	Mdn. = 0.5	71%	2198	<.001 ***
Typical pairwise change in ALE (pre - post), due to recalibration	Mdn. = 0	86%	2508	<.001 ***
Typical effect (Cohen's $d$ ) of recalibration on ALE	Mdn. = 0	86%	2437	<.001 ***
Proportion of samples in which recalibration improved mean ALE	Mdn. = 0.5	86%	2468.5	<.001 ***

Significance levels: \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$

**Discussion.** Similar to previous findings in Study 3c (PHL recalibration), the results of Analysis 3c.iii demonstrate that credibility-based recalibration typically led to non-trivial improvements in ALE. As with both variants of AJE, this conclusion was supported by all four ALE-based outcome measures, with the average participant in the Philadelphia air temperature study being able to expect 59% of their individual judgments to improve, each by an average of 4.37 points on the probability scale. Overall, this corresponded to an average, pairwise Cohen's  $d$  of 0.21, and a better (smaller) mean ALE in 87% of a forecaster's prediction samples, on average. Once again, this corresponded to distributions of typical effect sizes that were significantly greater than chance, suggesting that credibility-based recalibration would likely have improved ALE in Study 3c, in general. Critically, all four distributions of DVs suggested that (a) the effects of recalibration were more frequently positive than they were negative (the odds of a positive effect were roughly 7:3 for the lowest performing DV, and greater than 17:3 for

the other three); and (b) that the practical effects of recalibration were not insubstantial (the study-wide average improvement in ALE was 4 points on the probability scale). Thus, the results of Analysis 3c.iii suggest that the linear credibility framework fulfilled all three criteria for a useful credibility model when applied to judgments from Study 3c.

**Analysis 3c.iv: typical effects of recalibration on reliability (PHL recal., reliability).** Finally, in Analysis 3c.iv, I examined the typical effects of recalibration on *reliability* (see: Murphy, 1973). As with ALE, a systematic improvement in *reliability* due to recalibration would suggest that shrinking SPJs towards *estimated optima* had reduced errors and biases in judgment. In principle, this analysis was conceptually parallel to the analysis carried-out in Analysis 3c.ii (PHL recal., AJE vs. baserates), though the baserates observed here were those that were observed in-sample, rather than over the past ten years.

As a final test of the linear credibility framework, therefore, Analysis 3c.iv examined the typical effects of credibility-based recalibration on *reliability*, across forecasters. The results of this analysis can be seen in the figures and tables below. Figures 57 and 58 show the empirical distributions of the two *reliability*-based DVs across forecasters. Table 41 provides descriptive statistics for each of the distributions represented in Figures 57 and 58, and Table 42 shows the results of Wilcoxon signed-rank tests examining the null hypothesis that the median of each distribution does not differ from chance (represented by red dotted lines in Figures 57 and 58).

Figure 57

[PHL data]: Proportion of samples in which recalibration improved (reduced) reliability.

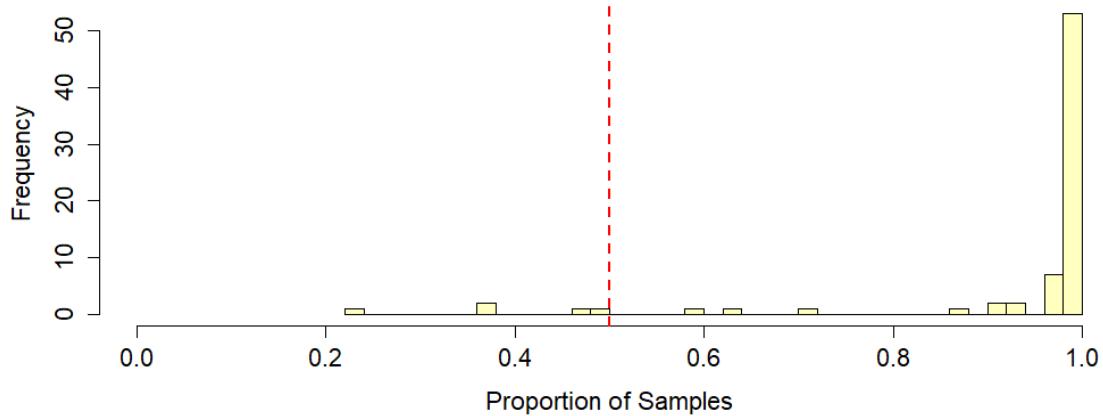
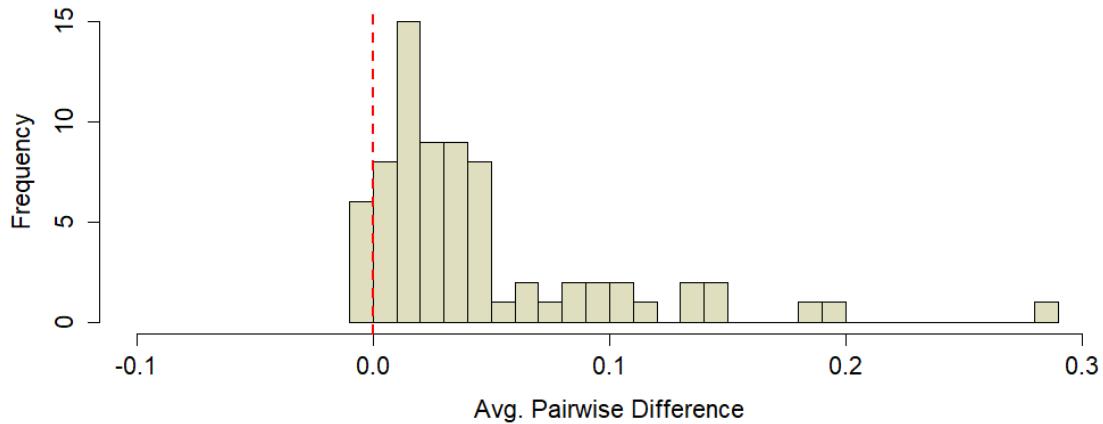


Figure 58

[PHL data]: Typical pairwise change in reliability (pre – post), due to recalibration.



Note: positive values indicate an improvement (reduction) in reliability.

Table 41

[PHL data]: Typical effects of credibility-based recalibration on reliability, summarized across forecasters.

Between-Subjects Summary Statistics					
Outcome Measure	Mean	Mdn.	SD	Min.	Max.
Proportion of samples in which recalibration improved <i>reliability</i>	94%	100%	17%	23%	100%
Typical pairwise change in <i>reliability</i> (pre - post), due to recalibration	4.54 $\times 10^{-2}$	2.60 $\times 10^{-2}$	5.33 $\times 10^{-2}$	-0.27 $\times 10^{-2}$	28.62 $\times 10^{-2}$

Table 42

*[PHL data]: Wilcoxon signed-rank tests, assessing whether credibility-based recalibration typically improved judgments (i.e., reduced reliability) beyond chance.*

Outcome Measure	$H_0$	Prop. Mass $> H_0$	Stat. ( $V$ )	p-value
Proportion of samples in which recalibration improved <i>reliability</i>	Mdn. = 0.5	93%	2613	<.001***
Typical pairwise change in <i>reliability</i> (pre - post), due to recalibration	Mdn. = 0	92%	2675	<.001***

Significance levels: \* p <.05; \*\* p <.01; \*\*\* p <.001

**Discussion.** Consistent with the other findings in Study 3c (PHL recalibration), the results of Analysis 3c.iv demonstrate that credibility-based recalibration had a typically beneficial effect on *reliability* in the Philadelphia air temperature data. Indeed, for an average participant in Study 3c, recalibration was expected to improve (reduce) *reliability* in a massive 94% of each forecaster's bootstrap trials, on average (and 100% of trials at the median). As expected, this effect was accompanied by an average pairwise improvement (reduction) in *reliability* across samples. In both cases, the typical effects of

recalibration on *reliability* were greater than chance for more than 92% of forecasters, both of which corresponded to significant statistical trends. In general, therefore, the results of Analysis 3c.iv provide strong evidence to suggest that credibility-based recalibration improved both judgmental accuracy and calibration in Study 3c.

## General Discussion

Taken together, the results of Study 3c demonstrate that the linear credibility framework can be a useful model of credibility and is able to help decision makers effectively identify and ameliorate errors and biases in judgment. Furthermore, because Study 3c employed amateur participants making judgments in a domain of moderate difficulty, there is no reason to suspect that these results would not generalize to other contexts. Thus, the successes of the linear credibility framework in improving air temperature judgments suggest that even simple models of credibility may often be useful to decision makers. As discussed in Analysis 3c.ii (PHL recal., AJE vs. baserates), a particularly noteworthy finding from Study 3c is that recalibrating SPJs relative to intersubjective *estimated optima* (i.e., optimized crowd aggregates) can improve the accuracy of judgments relative to normative benchmarks such as historical baserates. As such, Study 3c is a definitive example of a case where the underlying theory of credibility is sound.

In other words, by examining the relationship between an individual's judgments and simple, crowdsourced approximations of "optimal" judgments, a decision maker can gain insight into *how* and *why* an individual's judgments tend to err. Though this sort of

examination is unlikely to provide a complete picture of such tendencies — especially when one's method for estimating credibility is relatively unsophisticated — it can still provide empirical traction on such questions in the absence of normative criteria. Consequently, if credibility can be modeled successfully, it can help narrow the gap between the judgments of an unaided individual and those of a rational observer. In Study 3c, the fact that a simple credibility model could be applied to a decidedly ordinary dataset and narrow this gap by a non-trivial degree provides grounds for optimism that the linear credibility framework — and perhaps models of credibility, more generally — may be useful in a wide variety of domains.

## Conclusions

Studies 3a-3c demonstrate that the linear credibility framework can be useful to decision makers by (a) identifying stable, generalizable relationships between an individual's SPJs and optimized crowd aggregates; and (b) improving the accuracy and calibration of an individual's SPJs by correcting for these tendencies; thereby (c) improving subjective probability judgments for a large majority of individuals (using ecologically realistic amounts of data). Strictly speaking, this was not the unanimous finding of Studies 3a-3c. In Study 3b, the linear credibility framework failed to meet two of the three criteria of a useful model of credibility. However, the results of Analysis 3b.ii (MM recal., ALE) suggests that these results can be largely attributed to the fact that *estimated optima* in Study 2b (MM recalibration) were not particularly accurate as “model” judgments, rather than because the linear credibility framework was an

(especially) ill-suited model of credibility. Consequently, the results of all three studies provide evidence to suggest that the linear credibility framework can be useful, and the results of two provide direct evidence to this effect.

In general, these results warrant two conclusions. First, in at least some situations, the linear credibility framework can be useful as a model of credibility and can help decision makers identify and ameliorate errors and biases in judgment. Second, because it is generally unlikely that the “true” relationship between an individual’s judgments and any set of *estimated optima* is a simple, atheoretical line, it is reasonable to assume that the usefulness of the linear credibility framework represents an empirical lower bound for the usefulness of credibility models, more generally. In a wide variety of contexts, therefore, decision makers may find that credibility is worth examining.

## References

- Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision*, 3(1), 20-39.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. D. Kahneman, P. Slovic, and A. Tverski (Eds.) *Judgement under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 595-600.

## CHAPTER 4

### ON THE ROBUSTNESS OF EMPIRICAL MODELS OF CREDIBILITY

#### **Abstract:**

Modeling credibility represents an attempt to evaluate subjective probability judgments using simple, practical procedures. However, the absence of normative criteria makes it difficult to carry-out these evaluations rigorously and places constraints on the validity of empirical credibility models. Fortunately, as I have shown in previous chapters, decision makers can use exploratory research methods to examine the robustness of their models. To demonstrate this principle, Chapter 4 will revisit the weakest analyses from Chapters 2 & 3 and demonstrate how basic applications of the scientific method can be used to improve the performance of the linear credibility framework. In Study 4, I demonstrate that the reliability of credibility estimates can be improved by separately examining heterogeneous question types in the Philadelphia air temperature study. In Study 5, I demonstrate that improving the accuracy of *estimated optima* can improve the effects of credibility-based recalibration on absolute linear error in the March Madness study.

#### **Introduction**

Despite the successes of the linear credibility framework, examining credibility is a messy business. In large part, this is because modeling credibility represents an extraordinarily difficult task: i.e., reducing the evaluation of subjective probability judgments (SPJs) to a simple, generalizable, and above all, *practical* set of empirical procedures, often in the absence of normative criteria. From a utilitarian perspective, developing tools for this type of assessment is essential to decision making in domains such as medicine (e.g., Kinnear & Jackson, 2016), public health (e.g., Cooke, Wilson, Tuomisto, Morales, Tainio, & Evans, 2007), law (e.g., Wells, 1992), finance (e.g., Jeffrey & Putnam, 2015), climate science (e.g., Budescu, Por, Broomell, & Smithson, 2014), politics (e.g., Gill & Walker, 2005; Tetlock & Lebow, 2001), and military intelligence

(e.g., Johnston, 2005; Kent, 1964). Yet, from a scientific perspective, the difficulty of identifying “objective” probabilities places severe constraints on the rigor with which SPJs can be evaluated. As a result, there will always be limits to the validity of empirical models of credibility, and even the most well-fitted model is unlikely to be a panacea.

To be a productive exercise, therefore, examining credibility requires a decision maker to understand (a) the conditions under which credibility estimates are likely to be informative; and (b) the extent to which such estimates should be interpreted as *measures* of error and bias in judgment vs. predictors or correlates of the same. In other words, gleaning insight from empirical models of credibility requires decision makers to have a firm grasp on the conditions under which a given model is expected to be robust and the conditions under which it is not. In the case of the linear credibility framework, for example, interpreting credibility estimates requires a variety of inferences about the fit and appropriateness of the model, including (but not limited to):

- The extent to which a pool of judges can be expected to possess diagnostic information about the events in question;
- The validity of Baron and colleague’s method for identifying *estimated optima* (Baron, Mellers, Tetlock, Stone & Ungar, 2014);
- The likely agreement between *estimated optima* and “objective” probabilities in a given domain (or, perhaps more clearly, the degree of irreducible uncertainty in the decision environment; see: Baron et al., 2014);
- The descriptive fit of main-effects linear regression to the relationship between an individual’s judgments and *estimated optima*;

- The generalizability of this fit to other judgments (out-of-sample, but within-person);
- The stability of this relationship over time;
- And the construct-validity of this fit when decomposed into components of *bias*, *expertise*, and *consistency*.

Fortunately, the research presented in Chapters 1-3 demonstrates that it is possible for empirical models of credibility to achieve a reasonable degree of “fit” or “appropriateness” on all of the items listed above. Indeed, given that the linear credibility framework was designed to undercut several of these factors, I have argued that models of credibility—and especially those that take care to provide a reasonable descriptive fit to the data at hand—may often be robust to departures from ideal conditions. In general, therefore, the ill-specified nature of any empirical model of credibility is both its greatest strength and its greatest weakness. On the one hand, by relaxing the rigorous inferential standards demanded by research on judgment under uncertainty, models like the linear credibility framework can allow decision makers to probe subjective knowledge and draw weak (although generally useful) inferences about the ways in which an individual’s judgments tend to err. On the other hand, by relying on weak inferential standards such as intersubjective agreement and average (i.e., historical) tendencies toward error, models such as the linear credibility framework can come at the cost of certainty in one’s evaluations. Thus, while often informative, examining credibility requires a decision maker to put a substantial degree of faith in her model.

At first glance, it is tempting to conclude that these strengths and weaknesses are typical of the “useful, but risky” types of tradeoffs associated with heuristic modes of evaluation (for a summary, see: Kahneman & Slovic, 2005). As I have demonstrated in Chapters 2 and 3, however, decision makers can use exploratory research methods to evaluate the reliability, validity, and utility of a given model of credibility (in a given context). Thus, in many cases, it is reasonable to assume that the risks associated with empirical models of credibility can be reduced (or at the very least, characterized) with a little bit of planning and some thoughtful calibration. For example, because the performance of credibility estimates is contingent on a variety of factors (including latent properties of the decision environment), it may often be advisable for decision makers to compare the performance of several models of credibility, and to be well-versed in the weakness, assumptions, and boundary conditions of the model that they ultimately choose. If this is done ahead of time, then when a decision maker encounters difficulty with her model (as I did at several points throughout this dissertation), she can leverage her familiarity with its strengths and weaknesses to begin diagnosing how and why it is not living-up to expectations.

To demonstrate how this sort of diagnosis might occur, Studies 4 and 5 will present variations on two of the weakest analyses from Chapters 2 and 3. In both cases, the purpose of these analyses will be to (a) follow-up on my suppositions about how and why the linear credibility framework failed to meet my expectations; and (b) demonstrate how these issues might be ameliorated. Because these analyses were both conducted post-hoc, they should not be taken as evidence for the value or applicability of the linear

credibility framework. Instead, I present these studies for the sole purpose of demonstrating that—while models of credibility are necessarily imperfect—they can also be tested, tailored, and improved with a little bit of exploration. In Study 4, therefore, I will reexamine Analysis 2c.i (PHL reliability) and demonstrate that the reliability of credibility estimates derived from the Philadelphia air temperature study can be improved by considering different subsets of judgments separately. Then, in Study 5, I will return to Analysis 3b.ii to demonstrate that the typical effects of credibility-based recalibration on absolute linear error (ALE) can be dramatically improved in the March Madness data by identifying a more strongly defensible set of *estimated optima*.

#### **Study 4: The Effect of Question Heterogeneity on Reliability (PHL reliability redux)**

In Analysis 2c.i (PHL reliability), non-bootstrapped estimates of *bias* and *expertise* from the Philadelphia air temperature study were less reliable than one might have hoped. As shown in Figure 14 (reported in Analysis 2c.i), intraclass correlations for these estimates never surpassed the conventional thresholds for “poor” ( $ICC < 0.40$ ) and “fair” ( $0.40 \leq ICC \leq 0.59$ ), respectively (Ciccetti, 1994). Fortunately, these shortcomings were not sufficient to undermine the results of Study 2c (PHL reliability/validity), as the bootstrapping process was able to compensate for the observed lack of reliability. In principle, however, a lack of reliability in non-bootstrapped estimates of *bias* and *expertise* has important implications for the linear credibility framework. Specifically, this shortcoming is likely to produce an unnecessary degree of variance in credibility estimates from one trial to the next. Thus, in cases where a decision maker does not have

the luxury of bootstrapping, estimates of *bias* and *expertise* are unlikely likely to provide an accurate indication of an individual’s “skill” or “proficiency” in subjective probability judgment and could result in ineffective or detrimental recalibration.

Following Study 2c, my best guess as to why the reliabilities of non-bootstrapped estimates of *bias* and *expertise* were suboptimal was that individuals were prone to different degrees of error and bias across different types of questions (e.g., predictions about temperatures in January vs. predictions about temperatures in July). The decision to include three types of questions in the Philadelphia air temperature study was intended to increase the speed with which data could be gathered. However, in hindsight, it is apparent that this was a risky methodological choice. Thus, to follow-up on whether credibility varied across question types, I conducted a series of post-hoc analyses in which I examined the reliabilities of different subsets of the Philadelphia air temperature data separately.

## Method

**Procedure.** In Study 4, data from the Philadelphia air temperature study were divided into five non-mutually-exclusive<sup>29</sup> categories: (a) predictions concerning probe temperatures from the month of January; (b) predictions concerning probe temperatures from the month of July; (c) predictions concerning the probability of observing a

---

<sup>29</sup> In the ideal case, predictions would have been divided into six mutually exclusive categories according to a  $2 \times 3$  factorial design — i.e., month = {January, July}  $\times$  question type = {plus five, minus five, more extreme}. However, because this would have yielded only 40 SPJs per participant per cell, this design would have made it impossible to replicate the procedure used in Analysis 2c.i (PHL reliability).

temperature five degrees warmer than the probe temperature; (d) predictions concerning the probability of observing a temperature five degrees cooler than the probe temperature; and (e) predictions concerning the probability of observing a temperature more extreme than the probe temperature (i.e., colder in January, warmer in July). For each subset of SPJs, I then evaluated the reliability of linear credibility estimates using the same procedure as Analysis 2c.i (PHL reliability).

As a reminder, this procedure examined the reliability (i.e., intraclass correlation) of non-bootstrapped credibility estimates at five levels of calibration sample size:  $n_{cal} = \{10, 20, \dots, 50\}$ . At each level, reliability was estimated across 30 bootstrap trials. As described above, the purpose of this analysis was to determine whether the reliability of non-bootstrapped estimates would be higher when data were divided according to question type (vs. pooled), presumably because individuals exhibited inconsistent tendencies towards error and bias across different types of questions.

## Results

The results of Study 4 are presented below. In Figures 59-61, reliabilities of non-bootstrapped estimates of *bias*, *expertise*, and *consistency* are graphed in separate plots. Figure 59 shows the reliability of non-bootstrapped estimates of *bias* ( $\hat{\alpha}_{in}^*$ ); Figure 60 shows the reliability of non-bootstrapped estimates of *expertise* ( $\hat{\beta}_{in}^*$ ); and Figure 61 shows the reliability of non-bootstrapped estimates of *consistency* ( $\hat{\sigma}_{in}^*$ ). In each of these plots, the y-axis represents reliability (i.e., intraclass correlation, or ICC); the x-axis

represents calibration sample size ( $n_{cal}$ ); and separate curves represent different subsets of the Philadelphia air temperature data.

Figure 59

[PHL data]: Reliability of non-bootstrapped estimates of bias ( $\hat{\alpha}_{in}^*$ ), varying by calibration sample size ( $n_{cal}$ ) and data subset.

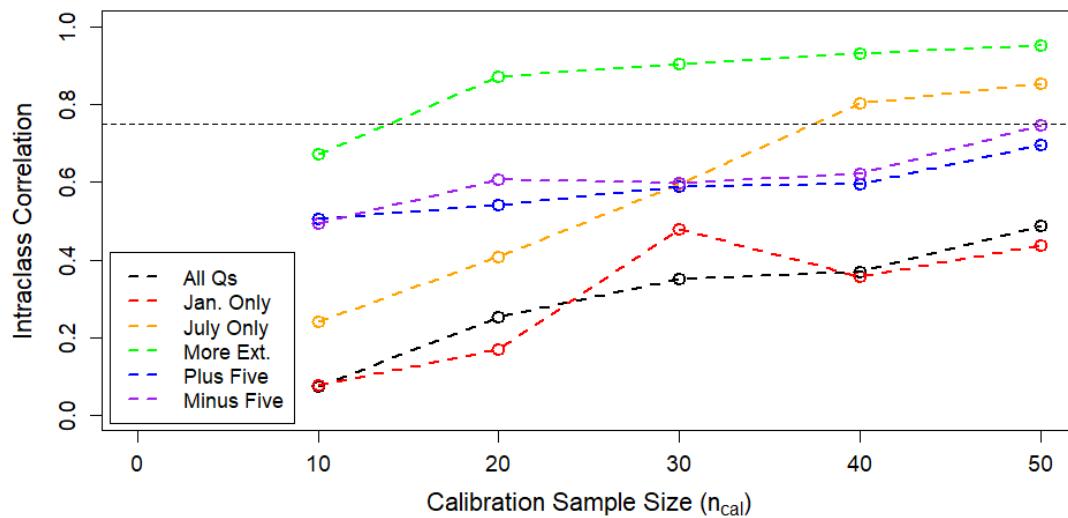


Figure 60

[PHL data]: Reliability of non-bootstrapped estimates of expertise ( $\hat{\beta}_{in}^*$ ), varying by calibration sample size ( $n_{cal}$ ) and data subset.

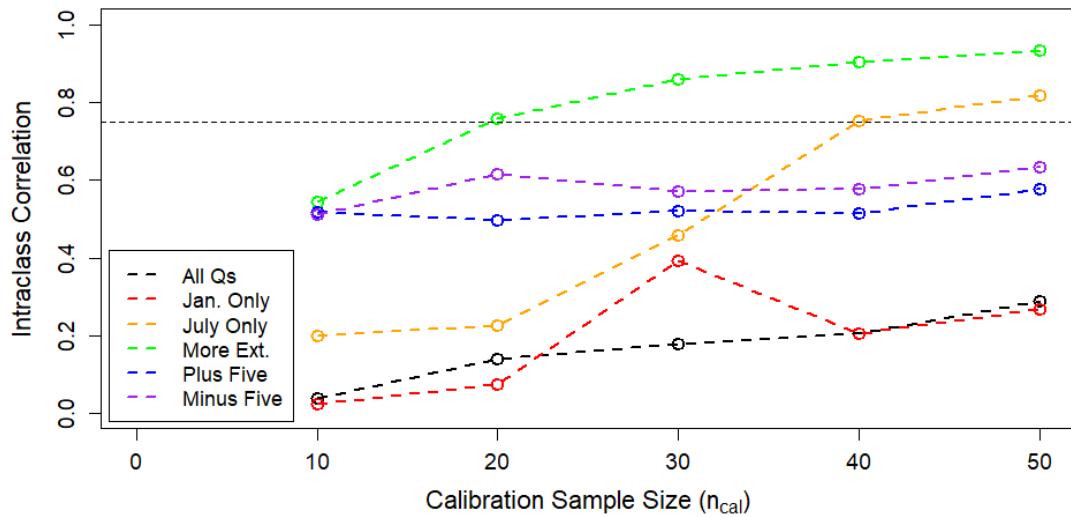
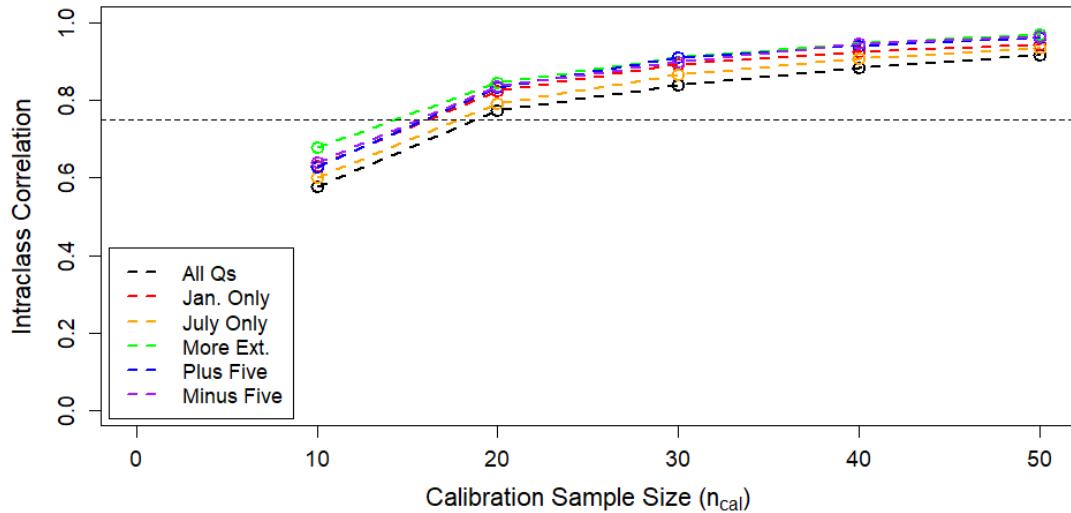


Figure 61

[PHL data]: Reliability of non-bootstrapped estimates of consistency ( $\hat{\sigma}_{in}^*$ ), varying by calibration sample size ( $n_{cal}$ ) and data subset.



## Discussion

As expected, the results of Study 4 demonstrate that the reliability of non-bootstrapped estimates of *bias* and *expertise* were heterogeneous across question types. In general, these results suggest that the poor performance of the linear credibility framework in Analysis 2c.i (PHL reliability) can be attributed to the fact that individuals' *bias* and *expertise* tended to vary across question types (rather than the alternative, which would suggest that the linear credibility framework was fundamentally ill-suited to modeling credibility in these data). As can be seen in Figures 59-61, this explanation is further supported by the fact that the reliabilities of estimates derived from subsets of the Philadelphia air temperature data were typically higher than those derived from the pooled set of all questions.

In the case of both *bias* and *expertise*, however, the reliability of estimates derived from SPJs pertaining to January probe temperatures do not follow this trend. Fortunately, there is a plausible explanation for this pattern of results. In general, winter air temperatures in the North-Eastern United States tend to be less consistent with temperatures experienced elsewhere in the United States (on the same dates) than summer temperatures. Thus, it is conceivable that participants in the Philadelphia air temperature study may have varied widely in their domain knowledge of January air temperatures in Philadelphia, but not July air temperatures. In this case, one would expect errors and biases associated with January SPJs to be less consistent than those associated with July (or those associated with subsets that pool across both months)—and thus, for the reliability of January credibility estimates to be lower. Broadly speaking, therefore, the results of Study 4 are consistent with the explanation that reliabilities in Analysis 2c.i

(PHL reliability) were suppressed by the fact that individuals' relative tendencies towards error and bias (i.e., credibility) were inconsistent across question types.

### **Study 5: The Impact of *Estimated Optima* Accuracy on the Effects of Recalibration (MM recalibration, ALE redux)**

In Analysis 3b.ii (MM recal., ALE), credibility-based recalibration of SPJs in the March Madness study failed to produce significant improvements in outcomes defined in terms of absolute linear error (ALE). Indeed, regardless of how outcomes were operationalized, the typical effects of recalibration tended to be *detrimental* to ALE in this analysis (see: Figures 35-39 and Tables 31-32, reported in Chapter 3). In principle, these results can be explained in one of two ways. First, it is conceivable that main-effects linear regression provided a prohibitively poor fit to data in the March Madness study. Second, it is possible that recalibration did not improve ALE in Analysis 3b.ii (MM recal., ALE) because recalibration tended to shrink individuals' judgments towards *estimated optima* that were not particularly accurate. Because the results of Analyses 3b.i (MM recal., AJE) and 3b.iii (MM recal., *reliability*) both indicated that recalibration tended to have a positive effect on outcomes related to absolute judgment error (AJE) and *reliability* (i.e., a measure similar to what the forecasting literature calls *calibration*; Lichtenstein, Fischhoff, & Phillips, 1982), my general conclusion from Study 3b (MM recalibration) was that the latter explanation was more likely—especially given the difficulty of March Madness predictions.

In line with this explanation, a post-hoc examination of *estimated optima* in Study 3b revealed that Baron et al.'s (2014) procedure for optimized aggregation did not dramatically improve ALE, relative to individual judgments. To put this in perspective, consider the performance of individual forecasters. Across all 143 participants in the March Madness study, mean ALEs ranged from 0.65 to 0.25 (where lower values indicate smaller errors, on average), with a study-wide mean of 0.39 and a standard deviation of 0.06. In contrast to this distribution, the mean ALE associated with Study 3b's *estimated optima* was 0.37—a value that represents an improvement of only 0.47 standard deviations over the average individual forecaster, and which was outperformed by a full 31% of individuals (i.e., 44/143 participants). From a statistical perspective, this difference corresponds to a significant improvement in mean ALE ( $t(142) = 5.67, p < 0.001$ ). However, upon closer inspection, it is apparent that (a) the practical effect-size of this improvement is small; and (b) that while Baron et al.'s (2014) procedure provided an improvement in ALE *at the average*, there is a substantial proportion of forecasters for whom optimized aggregation would have reduced accuracy (i.e., increased mean ALE).

Critically, however, the results of these post-hoc analyses also revealed that the predictive signal in the March Madness data could be amplified by deriving *estimated optima* from top-performing forecasters. For example, by limiting the optimized aggregation procedure to SPJs provided by the top 10% of forecasters (as measured by average Brier scores), the mean ALE of *estimated optima* could be improved to 0.31—a value which represents an improvement of 1.46 standard deviations over the average individual, and which was outperformed by only 4% of forecasters (i.e., 6/143).

participants). Though I would not have known to limit Baron et al's (2014) procedure in this way *ex ante*, the results of these exploratory analyses provided an opportunity to reexamine the performance of the linear credibility framework when fit to a more informative set of *estimated optima*. Thus, in Study 5 I reexamined the results of Analysis 3b.ii (MM recal., ALE) when credibility functions are fit to *estimated optima* derived from top-performing forecasters (i.e., the top 10% of participants, as measured by average Brier scores).

## Method

**Procedure.** The procedure for Study 5 was identical to that used in Analysis 3b.ii (MM recal., ALE), with the single exception that *estimated optima* were derived from only those SPJs provided by the top 10% of participants, as measured by forecast accuracy (average Brier scores). For simplicity, I will refer to this group of forecasters as *top performers*. To refresh the reader's memory, the procedure used in Analysis 3b.ii consisted of a single iteration of the general credibility estimation procedure. As in Analysis 3b.ii, the analytic parameters for this procedure were a calibration sample size of  $n_{\text{cal}} = 50$ ; a minimum prediction sample size of  $n_{\text{pred}} = 1$  (to maximize statistical power); and  $n_{\text{boot}} = 100$  bootstrap trials.

## Results

The results of Study 5 are shown in Figures 62-65, below. In all cases, Figures 62-65 contrast the typical effects of recalibration when credibility functions were fit to

*estimated optima* derived from all forecasters (represented by the pink bars) to the typical effects of recalibration when credibility functions were fit to *estimated optima* derived from top performers. Figure 62 contrasts distributions of the typical proportion of judgments for which recalibration improved (reduced) ALE; Figure 63 contrasts distributions of the typical pairwise change in ALE (pre – post), due to recalibration; Figure 64 contrasts distributions of the typical effect-size (Cohen's *d*) of recalibration on ALE; and Figure 65 contrasts distributions of the proportion of samples in which recalibration improved (reduced) mean ALE. In all cases, I opt not to report statistical tests concerning these distributions, as analyses were post-hoc and intended to be illustrative rather than explanatory. However, I do include a representation of the typical effect-size that would be expected by chance (as indicated by the red dotted-line in each figure).

Figure 62

*[MM data]: Comparison of the typical proportion of judgments for which recalibration improved (reduced) ALE when estimated optima are derived from the SPJs of all forecasters vs. top performers.*

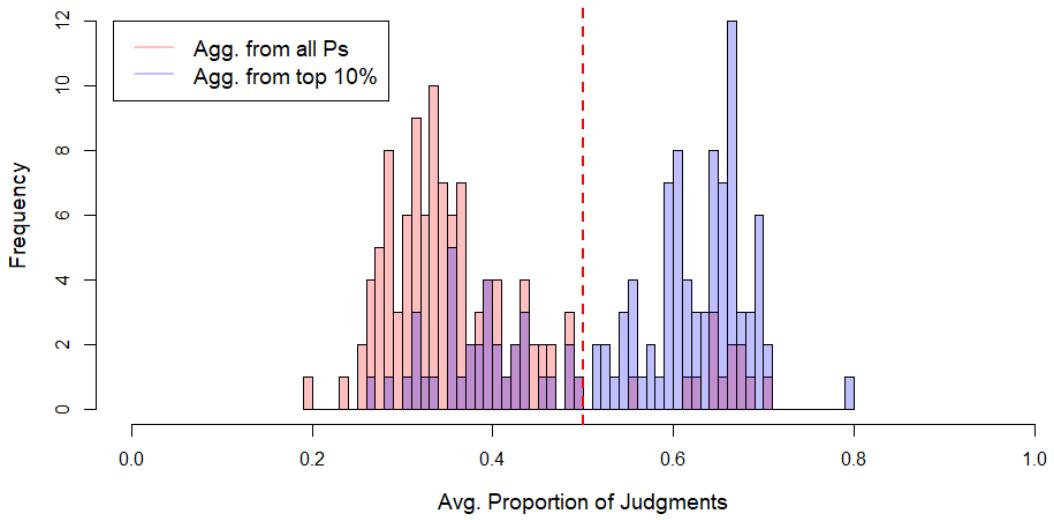
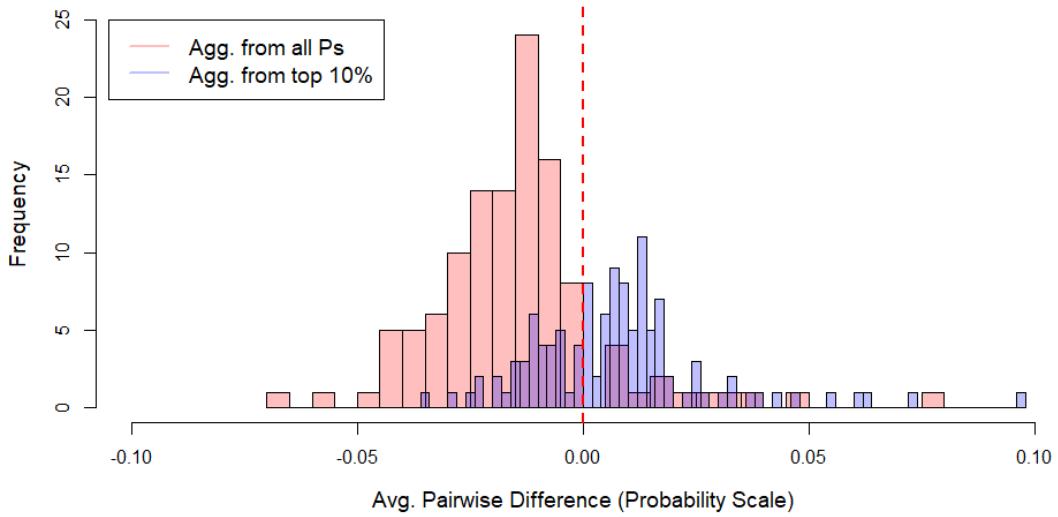


Figure 63

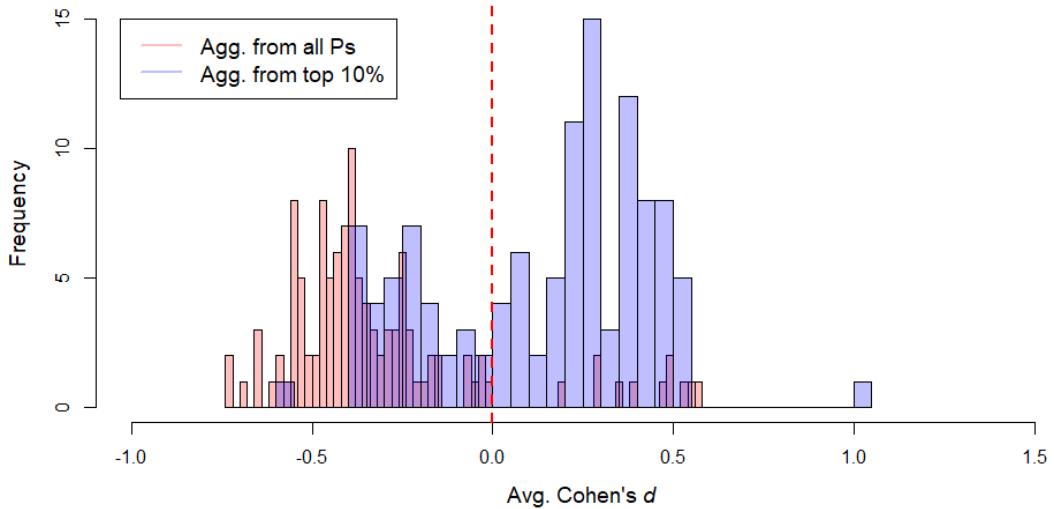
*[MM data]: Comparison of the typical pairwise change in ALE (pre – post), due to recalibration when estimated optima are derived from the SPJs of all forecasters vs. top performers.*



*Note:* positive values indicate an improvement (reduction) in ALE.

Figure 64

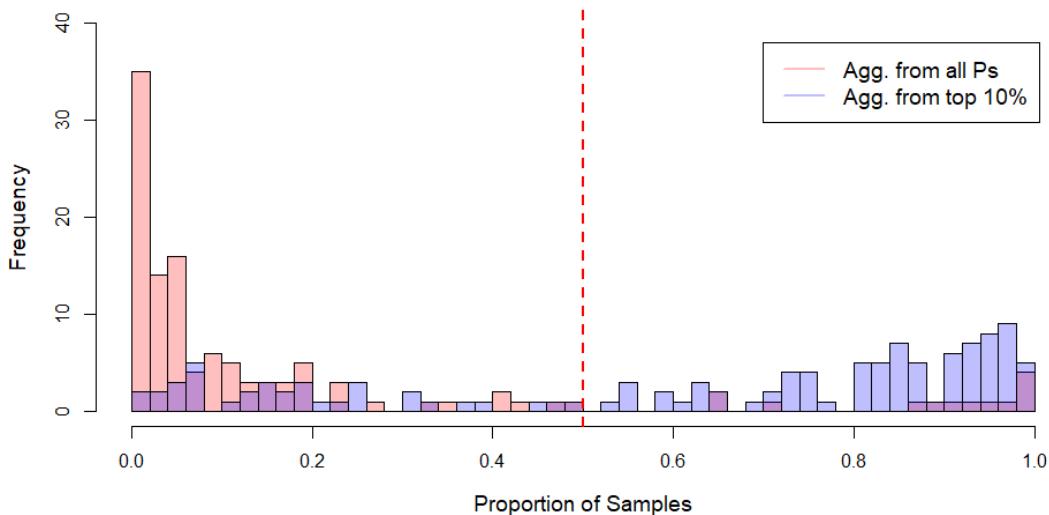
[MM data]: Comparison of the typical effect-size (Cohen's  $d$ ) of recalibration on ALE when estimated optima are derived from the SPJs of all forecasters vs. top performers.



Note: positive values indicate an improvement (reduction) in ALE.

Figure 65

[MM data]: Comparison of the proportion of samples in which recalibration improved (reduced) mean ALE when estimated optima are derived from the SPJs of all forecasters vs. top performers.



## Discussion

As predicted, the results of Study 5 demonstrate that the principal reason credibility-based recalibration did not typically improve ALE in Analysis 3b.ii (MM recal., ALE) is that *estimated optima* in the March Madness study were not particularly informative. As can be seen in Figures 62-65, fitting credibility functions to *estimated optima* derived from top performers improved the typical effects of recalibration on all four outcomes associated with ALE. In principle, of course, this pattern of results is not surprising. By shrinking individuals' SPJs toward more accurate *estimated optima* in Study 5, one might assume that it is a foregone conclusion that recalibration would yield more accurate results in Study 5 than it did in Analysis 3b.ii (MM recal., ALE). Critically, however, while it is tempting to dismiss these results as a case where I have sampled on the dependent variable, it is instructive to remember that all results presented in Study 5 (and elsewhere in this dissertation) reflect improvements in ALE *out-of-*

*sample*. As such, the results of Study 5 cannot be accurately summarized by saying that shrinking SPJs towards accurate judgments makes them more accurate. Instead, the results above demonstrate that individuals' SPJs tend to be less accurate than *estimated optima* for predictable, generalizable reasons. Thus, the better a decision maker can identify these reasons (e.g., by fine-tuning her model of credibility), the more effectively she can account for them in the future.

## **Summary**

If there is one thing that this research makes clear, it is that there is unlikely to be a one-size-fits-all approach to modeling credibility. Thus, in many cases, it may be prudent for decision makers to use exploratory research methods to examine the robustness of empirical models of credibility. Though far from an exhaustive treatment of the subject, this dissertation demonstrates that even simple models of credibility can be reasonably robust to violations of the (often unrealistic) assumptions that are necessary to operationalize credibility as an empirical construct. In addition, the present chapter argues that simple exploratory methods can help a decision maker test and observe a model's robustness in cases where there is reason to suspect it may be weak, vulnerable, or prone to fail. Using specific examples drawn from Chapters 2 and 3, Studies 4 and 5 demonstrate how such methods might be applied to empirical models of credibility and suggest that models such as the linear credibility framework can be tailored and improved with basic applications of the scientific method. Thus, while I refrain from drawing formal conclusions about the studies presented in Chapter 4, these studies provide reason

for optimism about the robustness, utility, and applicability of empirical models of credibility.

## References

- Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2), 133-145.
- Budescu, D. V., Por, H. H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, 4(6), 508-512.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284.
- Cooke, R. M., Wilson, A. M., Tuomisto, J. T., Morales, O., Tainio, M., & Evans, J. S. (2007). A probabilistic characterization of the relationship between fine particulate matter and mortality: Elicitation of European experts. *Environmental Science & Technology*, 41(18), 6598-6605.
- Gill, J., & Walker, L. D. (2005). Elicited priors for Bayesian model specifications in political science research. *The Journal of Politics*, 67(3), 841-872.
- Jeffrey, J. H., & Putman, A. O. (2015). Subjective probability in behavioral economics and finance: A radical reformulation. *Journal of Behavioral Finance*, 16(3), 231-249.
- Johnston, R. (2005). *Analytic culture in the U.S. intelligence community: An ethnographic study*. Washington, D. C.: Center for the Study of Intelligence.
- Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. K. J. Holyoak, & R. G. Morrison (Eds.). *The Cambridge handbook of thinking and reasoning*. New York: Cambridge University Press.
- Kent, S. (1964). Words of estimative probability. *Studies in Intelligence*, 8(4), 49-65.
- Kinnear, J., & Jackson, R. (2017). Constructing diagnostic likelihood: Clinical decisions using subjective versus statistical probability. *Postgraduate Medical Journal*, 93(1101), 425-429.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. D. Kahneman, P. Slovic, & A. Tverski (Eds.) *Judgement under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Tetlock, P. E., & Lebow, R. N. (2001). Poking counterfactual holes in covering laws: Cognitive styles and historical reasoning. *American Political Science Review*, 95(4), 829-843.

Wells, G. L. (1992). Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology*, 62(5), 739-752.

## GENERAL CONCLUSIONS

Across nine studies, I have demonstrated that the linear credibility framework is a robust (albeit imperfect) model of credibility that can provide reliable, valid, and useful estimates of an individual's tendencies towards error and bias in judgment. By design, these studies reflect the performance of a relatively weak model of credibility that can be implemented with ecologically realistic types (and amounts) of data across a wide variety of decision environments. As a result, I argue that these results represent a lower-bound for the performance of empirical models of credibility, more generally. Consequently, I conclude that (a) examining credibility may often provide decision makers with a practical method for evaluating the “quality” or relative validity of subjective probability judgments (SPJs); and (b) that even simple models of credibility can often allow them to do so in a “cost-effective” (i.e., net-beneficial) way. In practice, of course, the assumptions underlying the linear credibility framework (and other, similar models of credibility) make the examination of credibility an imperfect science. However, as I have demonstrated, the risks associated with credibility information can often be anticipated and examined, thereby mitigating their impact on decision making.

Following from these conclusions, it stands to reason that decision makers may often benefit from examining credibility. Indeed, while the scope and magnitude of such benefits are likely to vary from one decision environment to the next, the theoretical foundations of credibility suggest that credibility information should be useful for at least two broad purposes: (a) identifying high- and low-performing judges with relatively little

up-front costs; and (b) assessing the “quality” or relative validity of the information that such judges provide. When considered in the context of real-world decision making, therefore, the success of the linear credibility framework suggests a broad range of applications.

When used to identify high- and low-performing judges, for example, credibility information is likely to allow decision makers to accomplish a variety of desirable tasks (or, at the very least, to improve their performance on such tasks). Though far from an exhaustive list, these tasks include: identifying expert judges and/or those individuals with the most informative judgments<sup>30</sup> (perhaps for the purposes of reducing the number of judges necessary for a given task); evaluating the performance of judges and/or decision makers in public and private spheres (e.g., medicine, finance, public policy, intelligence analysis, etc.); holding decision makers accountable for their judgments; and targeting underperforming individuals for training, debiasing, or other remedial interventions. Indeed, in principle, credibility information ought to be useful in any context where a decision maker would benefit from making concrete claims about the relative skill of various judges. Given the performance of the linear credibility framework in the present research, I argue that empirical models of credibility may often be able to live up to this standard, in practice.

---

<sup>30</sup> Note that the most “informative” judgements are not necessarily the “best” judgments, as measured by conventional means (e.g., accuracy). Consider, for example, a judge whose SPJs are highly *biased* and *inexpert*, but perfectly *consistent* with respect to *estimated optima*. Though the SPJs provided by such a judge are unlikely to be informative in their own right, a decision maker with knowledge of her credibility would be able to use her SPJs to perfectly reproduce *estimated optima*. Thus, a decision maker might gain more information by examining the SPJs of this “poor” (but highly consistent) judge than those provided by a more accurate (but less consistent) alternative.

Furthermore, when used to assess the “quality” or relative validity of the judgements at one’s disposal, credibility information is likely to help decision makers make substantive claims about (or, at the very least, place meaningful constraints on) the degree of uncertainty associated with a subjective belief. When used for this purpose, credibility information might allow decision makers to accomplish tasks such as: recalibrating SPJs by “correcting for” historical tendencies towards error and bias in judgment (as was done in this dissertation); estimating credible intervals around an individual’s SPJs (i.e., expressing SPJs as a point estimate  $\pm$  an individual’s “expected degree of error”); combining this type of interval-based SPJ across individuals to identify an informative posterior distribution of beliefs; and eliminating (or reducing) individual-level errors and biases prior to aggregation. From a statistical perspective, that is, credibility information ought to provide decision makers with information about the reliability and validity of an individual’s SPJs as measures of “optimal” judgments. And once again, given the performance of the linear credibility framework, the results of the present research suggest that — while imperfect — empirical models of credibility can deliver this type of information in a useful, real-world capacity.

When considered in general, therefore, the results of this research suggest that examining credibility is likely to provide decision makers with a wealth of information about what they know (and what they *don’t* know) about the “quality” or relative validity of an individual’s subjective probability judgments. Furthermore, the results of this research suggest that such information can be extracted by decision makers with a relatively small degree of investment and effort. From the perspective of a real-world

decision maker, both findings are cause for optimism. Thus, there is strong scientific basis for additional research on empirical models of credibility and their applications to forecasting, risk assessment, and decision analysis.

## APPENDIX A

Appendix A contains plain-text printouts of the R code used to conduct the GJP analyses reported in this dissertation. Similar scripts were used to conduct the parallel analyses for the March Madness study and the Philadelphia air-temperature study. To prevent the accidental introduction of errors, these scripts have been minimally altered from the versions used by the author in his original analyses. As a result, the organization of these scripts does not match the organization of the research presented in Chapters 2 & 3 precisely. Also, in some cases, the terminology used in these scripts is altered from that used in the main body of this dissertation (e.g., *consistency* is represented by the name “xi” in these scripts, whereas it is designated by the Greek letter sigma,  $\sigma$ , elsewhere). Finally, there are some analyses that appear in these scripts that are not reported in Chapters 2 & 3 (these discrepancies are most prominent in the “SPJ Recalibration” script). In general, these analyses were not reported because there were redundant with those presented in the main body of this dissertation and/or because they represented a less defensible approach to answering the research question at hand. In the author’s judgment, these rejected analyses do not substantively change, weaken, or otherwise contradict the results reported therein.

Notably, to meet the necessary formatting constraints, the following scripts include automatic text-wrapping, and can be difficult read in some places. In addition, the often refer to (.csv or .dat) data files that could not be included in this appendix. For copies of the original (.R) script documents or to request permission to use the associated data files, please contact the author directly.

```

#####
# Assessing Credibility in Subjective Probability Judgment #
# Reliability Experiments, Good Judgment Data #
#
# Josh Baker #
# jbak@sas.upenn.edu #
#
#####

#####
# DATA CLEANING AND PREPARATION
#
#####

##### Set Working directory and load relevant libraries #####
library(psych)
library(plotrix)

setwd("~/yourPath")

#####
# Import and Prepare Good Judgment Data for Use in Recalibration #####
fcasts <- read.csv("fcasts_small.file.csv")

## Convert factors to numeric or character data types (and trim last two characters from each element of "ifp_id")
fcasts$ifp_id <- as.numeric(substr(as.character(fcasts$ifp_id),1,4))
fcasts$answer_option <- as.character(fcasts$answer_option)
fcasts$date <- as.character(fcasts$date)
fcasts$timestamp <- as.character(fcasts$timestamp)
fcasts$outcome <- as.character(fcasts$outcome)
fcasts$g.tnt <- as.character(fcasts$g.tnt)
fcasts$mod.tag <- as.character(fcasts$mod.tag)
fcasts$date.closed <- as.character(fcasts$date.closed)
fcasts$time.horiz <- as.character(fcasts$time.horiz)

## Subset fccasts to answer_option=="a" (only forecasts for focal outcomes, not complements)
fcasts <- subset(fccasts, answer_option=="a")

## Set aside fccasts metadata that are irrelevant to recalibration analyses
fcasts_w.metadata <- fccasts
fcasts <- fccasts[which(is.element(colnames(fccasts),c("ifp_id",
                                                 "user_id",
                                                 "outcome",
                                                 "val.unrounded",
                                                 "bs.unrounded")))]]

## Rename some columns for consistency/expediency
colnames(fccasts)[which(colnames(fccasts)=="ifp_id")] <- "ifp.id"
colnames(fccasts)[which(colnames(fccasts)=="user_id")] <- "gjp.id"
colnames(fccasts)[which(colnames(fccasts)=="val.unrounded")] <- "prob"
colnames(fccasts)[which(colnames(fccasts)=="bs.unrounded")] <- "bs"

## Recode "outcome" to binary ("a" = 1; "b" = 0)
fccasts$outcome <- ifelse(fccasts$outcome=="a",1,0)

```

```

## Round all prob values to two places to fix a floating point rounding issue
fcasts$prob <- round(fcasts$prob,2)

## Extract GJP IDs as a vector
gjp.ids <- unique(fcasts$gjp.id)
gjp.ids <- gjp.ids[order(gjp.ids)] #Sort in ascending order

#####
##### Elementary Functions for credibility estimation #####
#####

## Function for calculating the number of unique elements in a vector #####
n.unique <- function(vector){length(unique(vector))}

## Create a data.frame for individual.difference data (here, only used to track n.ifps) #####
ind.diffs <- data.frame(gjp.id=gjp.ids)
ind.diffs$gjp.id <- gjp.ids
fcasts <- fcasts[order(fcasts$gjp.id),]
ind.diffs <- cbind(ind.diffs, n.ifps=as.vector(tapply(fcasts$ifp.id,fcasts$gjp.id,n.unique)))

### Brier score function, for optimization #####
bs.opt <- function(x,outcome){sum((temp.outcomes - ptrans(temp.probs,x))^2)}

### Brier score function, for scoring #####
calc.bs <- function(x,outcome){(x-outcome)^2}

## Function for converting a probability to log-odds
calc.logodds <- function(x){log(x/(1 - x))}

## Extremizing function used to correct for "Regression" towards 50% (see: Baron et al., 2014)
ptrans <- function(p,a){p^a/(p^a + (1 - p)^a)}

## General function for calculating sum of squared errors
calc.sse <- function(obj,est){sum((obj-est)^2)}

#####
##### Function for generating optimized, aggregate judgments, given a sample size criterion #####
#####

gen.agg_gjp <- function(sample.size, method=c("mean","median"), log.odds=TRUE){

  ## Subset to only those forecasters who addressed a sufficient number of IFPs
  keepers <- gjp.ids[which(ind.diffs$n.ifps >= (sample.size + 30))]          #Identify forecasters
  with enough data to recalibrate 30+ out of sample judgments
  working.data <- subset(fcasts, is.element(fcasts$gjp.id, keepers))           #Subset to only relevant
  forecasters
  working.data$prob <- ifelse(working.data$prob==0,0.01,working.data$prob) #Adjust extreme values
  to prevent (-)Inf, once transformed to log-odds
  working.data$prob <- ifelse(working.data$prob==1,0.99,working.data$prob) #Adjust extreme values
  to prevent (-)Inf, once transformed to log-odds
  working.data <- working.data                                         #Save working data to
  .GlobalEnv
  temp.outcomes <- as.vector(tapply(working.data$outcome, working.data$ifp.id, mean)) #Save
  relevant outcomes to .GlobalEnv

  ## Aggregate judgments and save to .GolbalEnv
  if(method=="mean"){
    temp.probs <- as.vector(tapply(working.data$prob, working.data$ifp.id, mean))
  }
  if(method=="median"){
    temp.probs <- as.vector(tapply(working.data$prob, working.data$ifp.id, median))
  }
}

```

```

}

## Calculate optimized extremizing coefficient
a <- as.numeric(optimise(bs.opt, interval=c(0,20)))[1]

## Apply extremizing coefficient to simple aggregatees to calculate optimized aggregate estimates
agg.judge <- ptrans(temp.probs,a)

## If desired, transform agg.judge to log-odds
if(log.odds){
  agg.judge <- calc.logodds(agg.judge)
}

return(agg.judge)

#rm(sample.size,method,log.odds,keepers,working.data,temp.outcomes,temp.probs,a,agg.judge)
}
#####
#####

##### Function for credibility-based recalibration and out of sample prediction (oosp) using GJP
Data #####
oosp_gjp <- function(user.id, sample.size, method=c("mean","median"), log.odds=TRUE, n.resample){

  ## Create an empty data.frame for results
  data.out <- data.frame(gjp.id=0, est.alpha=0, est.beta=0, est.ser=0) #Unique identifier for each forecaster #Estimated CF intercept (bias) #Estimated CF slope (expertise) #Standard error of the CF regression = "xi"
  (consistency)

  ## Generate optimized aggregates
  agg <- gen.agg_gjp(sample.size,method,log.odds)
  agg_ifp.ids <- as.vector(tapply(working.data$ifp.id, working.data$ifp.id, mean)) #Identify IFPs for which there exists an optimized aggregate

  ## Prepare individual forecaster's data
  user.data <- subset(working.data, gjp.id==user.id) #Subset to individual ("user")
  user.data <- user.data[order(user.data$ifp.id),] #Sort user.obs by ifp.id
  user.data$prob <- ifelse(user.data$prob==0, 0.01, user.data$prob) #Adjust extreme values to prevent (-)Inf, once transformed to log-odds
  user.data$prob <- ifelse(user.data$prob==1, 0.99, user.data$prob) #Adjust extreme values to prevent (-)Inf, once transformed to log-odds
  user.ifps <- unique(user.data$ifp.id) #Generate list of ifps that user responded to

  agg_user.matched.sample <- agg[which(is.element(agg_ifp.ids,user.ifps))]
  #Select matching optimized aggregates
  outcomes_user.matched.sample <- as.vector(tapply(user.data$outcome, user.data$ifp.id, mean))
  #select matching outcomes

  ## Calculate mean or median individual judgments for each ifp
  if(method=="mean"){
    user.obs <- as.vector(tapply(user.data$prob, user.data$ifp.id, mean))
    user.obs <- calc.logodds(user.obs)
  }
  if(method=="median"){
    user.obs <- as.vector(tapply(user.data$prob, user.data$ifp.id, median))
    user.obs <- calc.logodds(user.obs)
  }

  ## Fit credibility function, and conduct recalibration n.resample times for each individual
  i <- 1 #Resample index

  while(i <= n.resample){
    # Set/reset skip indicator for unusable loop iterations

```

```

flag <- FALSE

# Split data into calibration sample and prediction sample
cal.sample.indices <- sample.int(length(user.obs), sample.size, replace=FALSE)
pred.sample.indices <- setdiff(c(1:length(user.obs)), cal.sample.indices)

user.obs_cal.sample <- user.obs[cal.sample.indices]
user.obs_pred.sample <- user.obs[pred.sample.indices]

agg_cal.sample <- agg_user.matched.sample[cal.sample.indices]
agg_pred.sample <- agg_user.matched.sample[pred.sample.indices]

outcomes_cal.sample <- outcomes_user.matched.sample[cal.sample.indices]
outcomes_pred.sample <- outcomes_user.matched.sample[pred.sample.indices]

# Estimate credibility function (re-fit on a new calibration sample each iteration)
cf <- lm(agg_cal.sample ~ user.obs_cal.sample)

# Check for inestimable credibility functions, and re-set loop if necessary
if(is.na(cf$coeff[1]) | is.na(cf$coeff[2])){
  flag <- TRUE
  i <- i-1
}

# If CF is usable, calculate effects of recalibration
if(!(flag)){

  ## Compile output into a new row and add to data.out
  newrow <- data.frame(gjp.id=user.id,
                        est.alpha=cf$coeff[1],
                        est.beta=cf$coeff[2],
                        est.ser=summary(cf)$sigma)

  data.out <- rbind(data.out,newrow)
  flag <- FALSE
}

i <- i+1
}

## Clean-up and Return Results
data.out <- data.out[-1,]          #Remove dummy first row of data.out
rownames(data.out) <- seq.int(1:nrow(data.out)) #As a general precaution, reset rownames of
data.out

return(data.out)

#rm(user.id,sample.size,method,log.odds,n.reample,agg,agg_ifp.ids,user.data,user.ifps,agg_user.match
ed.sample,
#outcomes_user.matched.sample,user.obs,i,flag,cal.sample.indices,pred.sample.indices,user.obs_cal.sa
mple,
#user.obs_pred.sample,agg_cal.sample,agg_pred.sample,outcomes_cal.sample,outcomes_pred.sample,cf,cor
rected,
#mod,data.out,rmse.pre_lo,rmse.post_lo,rmse.impr,aje.pre_lo,aje.post_lo,aje.pre_p,aje.post_p,diffs.a
je_lo,
#diffs.aje_p,prop.aje.impr,mean.aje.pre_lo,mean.aje.post_lo,diff.mean.aje_lo,mean.aje.pre_p,mean.aje
.post_p,
#diff.mean.aje_p,mean.aje.impr,median.aje.pre_lo,median.aje.post_lo,diff.median.aje_lo,median.aje.pr
e_p,

```

```

#median.aje.post_p,diff.median.aje_p,median.aje.impr,coh.d_aje.pre.post_lo,coh.d_aje.pre.post_p,mean
.diff.aje_lo,
#median.diff.aje_lo,mean.diff.aje_p,median.diff.aje_p,ale.pre,ale.post,diffs.ale,prop.ale.impr,mean.
ale.pre,
#mean.ale.post,diff.mean.ale,mean.ale.impr,median.ale.pre,median.ale.post,diff.median.ale,median.ale
.impr,
#coh.d_ale.pre.post,mean.diff.ale,median.diff.ale,bs.pre,bs.post,diffs.bs,prop.bs.impr,mean.bs.pre,m
ean.bs.post,
#diff.mean.bs,mean.bs.impr,median.bs.pre,median.bs.post,diff.median.bs,median.bs.impr,coh.d_bs.pre.p
ost,
#mean.diff.bs,median.diff.bs,rel.pre,rel.post,diff.rel,rel.impr,res.pre,res.post,diff.res,res.impr,u
ncertainty)
}
#####
##### Function to run oosp_gjp for all eligible forecasters, given a desired sample.size
#####
run.oosp_gjp <- function(sample.size, method=c("mean","median"), log.odds=TRUE, n.resample){

  data.out <- data.frame(gjp.id=0,                               # AOS = average over samples
                        aos.alpha=0,
                        sd.alpha=0,
                        aos.beta=0,
                        sd.beta=0,
                        aos.xi=0,
                        sd.xi=0)

  ## Subset to only those forecasters with enough data to recalibrate 30+ out of sample predictions
  keepers <- gjp.ids[which(ind.diff$n.ifps >= (sample.size + 30))]

  ## Run oosp_gjp for each forecaster in keepers
  n <- 1

  while(n <= length(keepers)){
    temp <- oosp_gjp(user.id=keepers[n],sample.size,method,log.odds,n.resample)

    ## Compile data in a new row and add to data.out
    newrow <- data.frame(gjp.id=temp$gjp.id[1],
                          aos.alpha=mean(temp$est.alpha),
                          sd.alpha=sd(temp$est.alpha),
                          aos.beta=mean(temp$est.beta),
                          sd.beta=sd(temp$est.beta),
                          aos.xi=mean(temp$est.ser),
                          sd.xi=sd(temp$est.ser))

    data.out <- rbind(data.out,newrow)

    n <- n+1
  }

  data.out <- data.out[-1,]
  return(data.out)

  #rm(sample.size,method,log.odds,n.resample,data.out,keepers,n,temp,newrow)
}

#####
#####

```

```

##### Function to generate a reliability curve for non-bootstrapped credibility estimates
#####
gen.rc_non.boot <- function(method=c("mean","median"), log.odds=TRUE, n.resample){
  require(psych)

  sample.sizes <- c(10,20,30,40,50)
  rel.curve.alpha <- vector()
  rel.curve.beta <- vector()
  rel.curve.xi <- vector()
  n <- 1

  while(n <= length(sample.sizes)){
    keepers <- gjp.ids[which(ind.diff$n.ifps >= (sample.sizes[n] + 30))]

    alpha.out <- matrix(nrow=length(keepers),ncol=n.resample)
    beta.out <- matrix(nrow=length(keepers),ncol=n.resample)
    xi.out <- matrix(nrow=length(keepers),ncol=n.resample)
    i <- 1

    while(i <= length(keepers)){
      temp <-
      oosp_gjp(user.id=keepers[i],sample.size=sample.sizes[n],method="mean",log.odds=T,n.resample=n.resample)
      alpha.out[i,] <- temp$est.alpha
      beta.out[i,] <- temp$est.beta
      xi.out[i,] <- temp$est.ser

      i <- i+1
    }

    icc.alpha <- ICC(alpha.out)
    icc.beta <- ICC(beta.out)
    icc.xi <- ICC(xi.out)

    rel.curve.alpha <- append(rel.curve.alpha, icc.alpha[[1]]$ICC[1])
    rel.curve.beta <- append(rel.curve.beta, icc.beta[[1]]$ICC[1])
    rel.curve.xi <- append(rel.curve.xi, icc.xi[[1]]$ICC[1])

    n <- n+1
  }

  data.out <- data.frame(n.cal=sample.sizes,
                         rel.curve.alpha=rel.curve.alpha,
                         rel.curve.beta=rel.curve.beta,
                         rel.curve.xi=rel.curve.xi)

  return(data.out)

  #rm(method,log.odds,sample.sizes,rel.curve.alpha,rel.curve.beta,rel.curve.xi,n,temp,alpha.out,
  #beta.out,xi.out,i,data.out,icc.alpha,icc.beta,icc.xi)
}
#####
##### Function to generate a reliability curve for bootstrapped credibility estimates
#####
gen.rc_boot <- function(sample.size, method=c("mean","median"), log.odds=TRUE){
  require(psych)

  n.bootstraps <- c(10,20,30,40,50,60,70,80,90,100,200,250)
  trials <- c("a","b","c")
  n <- 1
  t <- 1

  while(n <= length(n.bootstraps)){
    while(t <= length(trials)){
      temp <- run.oosp_gjp(sample.size, method="mean", log.odds=TRUE, n.resample=n.bootstraps[n])
    }
  }
}

```

```

colnames(temp) <- paste(colnames(temp), "_", n.bootstraps[n], trials[t], sep="")

if(n==1 & t==1){
  data.out <- temp
} else{
  data.out <- cbind(data.out, temp)
}

t <- t+1
}

t <- 1
n <- n+1
}

#### Gather data for reliability calculations
aos.alphas_10 <- cbind(data.out$aos.alpha_10a,data.out$aos.alpha_10b,data.out$aos.alpha_10c)
aos.alphas_20 <- cbind(data.out$aos.alpha_20a,data.out$aos.alpha_20b,data.out$aos.alpha_20c)
aos.alphas_30 <- cbind(data.out$aos.alpha_30a,data.out$aos.alpha_30b,data.out$aos.alpha_30c)
aos.alphas_40 <- cbind(data.out$aos.alpha_40a,data.out$aos.alpha_40b,data.out$aos.alpha_40c)
aos.alphas_50 <- cbind(data.out$aos.alpha_50a,data.out$aos.alpha_50b,data.out$aos.alpha_50c)
aos.alphas_60 <- cbind(data.out$aos.alpha_60a,data.out$aos.alpha_60b,data.out$aos.alpha_60c)
aos.alphas_70 <- cbind(data.out$aos.alpha_70a,data.out$aos.alpha_70b,data.out$aos.alpha_70c)
aos.alphas_80 <- cbind(data.out$aos.alpha_80a,data.out$aos.alpha_80b,data.out$aos.alpha_80c)
aos.alphas_90 <- cbind(data.out$aos.alpha_90a,data.out$aos.alpha_90b,data.out$aos.alpha_90c)
aos.alphas_100 <- cbind(data.out$aos.alpha_100a,data.out$aos.alpha_100b,data.out$aos.alpha_100c)
aos.alphas_200 <- cbind(data.out$aos.alpha_200a,data.out$aos.alpha_200b,data.out$aos.alpha_200c)
aos.alphas_250 <- cbind(data.out$aos.alpha_250a,data.out$aos.alpha_250b,data.out$aos.alpha_250c)

aos.betas_10 <- cbind(data.out$aos.beta_10a,data.out$aos.beta_10b,data.out$aos.beta_10c)
aos.betas_20 <- cbind(data.out$aos.beta_20a,data.out$aos.beta_20b,data.out$aos.beta_20c)
aos.betas_30 <- cbind(data.out$aos.beta_30a,data.out$aos.beta_30b,data.out$aos.beta_30c)
aos.betas_40 <- cbind(data.out$aos.beta_40a,data.out$aos.beta_40b,data.out$aos.beta_40c)
aos.betas_50 <- cbind(data.out$aos.beta_50a,data.out$aos.beta_50b,data.out$aos.beta_50c)
aos.betas_60 <- cbind(data.out$aos.beta_60a,data.out$aos.beta_60b,data.out$aos.beta_60c)
aos.betas_70 <- cbind(data.out$aos.beta_70a,data.out$aos.beta_70b,data.out$aos.beta_70c)
aos.betas_80 <- cbind(data.out$aos.beta_80a,data.out$aos.beta_80b,data.out$aos.beta_80c)
aos.betas_90 <- cbind(data.out$aos.beta_90a,data.out$aos.beta_90b,data.out$aos.beta_90c)
aos.betas_100 <- cbind(data.out$aos.beta_100a,data.out$aos.beta_100b,data.out$aos.beta_100c)
aos.betas_200 <- cbind(data.out$aos.beta_200a,data.out$aos.beta_200b,data.out$aos.beta_200c)
aos.betas_250 <- cbind(data.out$aos.beta_250a,data.out$aos.beta_250b,data.out$aos.beta_250c)

aos.xis_10 <- cbind(data.out$aos.xi_10a,data.out$aos.xi_10b,data.out$aos.xi_10c)
aos.xis_20 <- cbind(data.out$aos.xi_20a,data.out$aos.xi_20b,data.out$aos.xi_20c)
aos.xis_30 <- cbind(data.out$aos.xi_30a,data.out$aos.xi_30b,data.out$aos.xi_30c)
aos.xis_40 <- cbind(data.out$aos.xi_40a,data.out$aos.xi_40b,data.out$aos.xi_40c)
aos.xis_50 <- cbind(data.out$aos.xi_50a,data.out$aos.xi_50b,data.out$aos.xi_50c)
aos.xis_60 <- cbind(data.out$aos.xi_60a,data.out$aos.xi_60b,data.out$aos.xi_60c)
aos.xis_70 <- cbind(data.out$aos.xi_70a,data.out$aos.xi_70b,data.out$aos.xi_70c)
aos.xis_80 <- cbind(data.out$aos.xi_80a,data.out$aos.xi_80b,data.out$aos.xi_80c)
aos.xis_90 <- cbind(data.out$aos.xi_90a,data.out$aos.xi_90b,data.out$aos.xi_90c)
aos.xis_100 <- cbind(data.out$aos.xi_100a,data.out$aos.xi_100b,data.out$aos.xi_100c)
aos.xis_200 <- cbind(data.out$aos.xi_200a,data.out$aos.xi_200b,data.out$aos.xi_200c)
aos.xis_250 <- cbind(data.out$aos.xi_250a,data.out$aos.xi_250b,data.out$aos.xi_250c)

## Run ICC function for each data.frame
icc_aos.alphas_10 <- ICC(aos.alphas_10)
icc_aos.alphas_20 <- ICC(aos.alphas_20)
icc_aos.alphas_30 <- ICC(aos.alphas_30)
icc_aos.alphas_40 <- ICC(aos.alphas_40)
icc_aos.alphas_50 <- ICC(aos.alphas_50)
icc_aos.alphas_60 <- ICC(aos.alphas_60)
icc_aos.alphas_70 <- ICC(aos.alphas_70)
icc_aos.alphas_80 <- ICC(aos.alphas_80)
icc_aos.alphas_90 <- ICC(aos.alphas_90)
icc_aos.alphas_100 <- ICC(aos.alphas_100)
icc_aos.alphas_200 <- ICC(aos.alphas_200)
icc_aos.alphas_250 <- ICC(aos.alphas_250)

```

```

icc_aos.betas_10 <- ICC(aos.betas_10)
icc_aos.betas_20 <- ICC(aos.betas_20)
icc_aos.betas_30 <- ICC(aos.betas_30)
icc_aos.betas_40 <- ICC(aos.betas_40)
icc_aos.betas_50 <- ICC(aos.betas_50)
icc_aos.betas_60 <- ICC(aos.betas_60)
icc_aos.betas_70 <- ICC(aos.betas_70)
icc_aos.betas_80 <- ICC(aos.betas_80)
icc_aos.betas_90 <- ICC(aos.betas_90)
icc_aos.betas_100 <- ICC(aos.betas_100)
icc_aos.betas_200 <- ICC(aos.betas_200)
icc_aos.betas_250 <- ICC(aos.betas_250)

icc_aos.xis_10 <- ICC(aos.xis_10)
icc_aos.xis_20 <- ICC(aos.xis_20)
icc_aos.xis_30 <- ICC(aos.xis_30)
icc_aos.xis_40 <- ICC(aos.xis_40)
icc_aos.xis_50 <- ICC(aos.xis_50)
icc_aos.xis_60 <- ICC(aos.xis_60)
icc_aos.xis_70 <- ICC(aos.xis_70)
icc_aos.xis_80 <- ICC(aos.xis_80)
icc_aos.xis_90 <- ICC(aos.xis_90)
icc_aos.xis_100 <- ICC(aos.xis_100)
icc_aos.xis_200 <- ICC(aos.xis_200)
icc_aos.xis_250 <- ICC(aos.xis_250)

## Tabulate ICCs for reliability curves
rel.curve_aos.alpha <- data.frame(n.resample=n.bootstraps,
                                    icc_aos.alpha=c(icc_aos.alphas_10[[1]]$ICC[1],
                                                   icc_aos.alphas_20[[1]]$ICC[1],
                                                   icc_aos.alphas_30[[1]]$ICC[1],
                                                   icc_aos.alphas_40[[1]]$ICC[1],
                                                   icc_aos.alphas_50[[1]]$ICC[1],
                                                   icc_aos.alphas_60[[1]]$ICC[1],
                                                   icc_aos.alphas_70[[1]]$ICC[1],
                                                   icc_aos.alphas_80[[1]]$ICC[1],
                                                   icc_aos.alphas_90[[1]]$ICC[1],
                                                   icc_aos.alphas_100[[1]]$ICC[1],
                                                   icc_aos.alphas_200[[1]]$ICC[1],
                                                   icc_aos.alphas_250[[1]]$ICC[1])))

rel.curve_aos.beta <- data.frame(icc_aos.beta=c(icc_aos.betas_10[[1]]$ICC[1],
                                                 icc_aos.betas_20[[1]]$ICC[1],
                                                 icc_aos.betas_30[[1]]$ICC[1],
                                                 icc_aos.betas_40[[1]]$ICC[1],
                                                 icc_aos.betas_50[[1]]$ICC[1],
                                                 icc_aos.betas_60[[1]]$ICC[1],
                                                 icc_aos.betas_70[[1]]$ICC[1],
                                                 icc_aos.betas_80[[1]]$ICC[1],
                                                 icc_aos.betas_90[[1]]$ICC[1],
                                                 icc_aos.betas_100[[1]]$ICC[1],
                                                 icc_aos.betas_200[[1]]$ICC[1],
                                                 icc_aos.betas_250[[1]]$ICC[1])))

rel.curve_aos.xi <- data.frame(icc_aos.xi=c(icc_aos.xis_10[[1]]$ICC[1],
                                              icc_aos.xis_20[[1]]$ICC[1],
                                              icc_aos.xis_30[[1]]$ICC[1],
                                              icc_aos.xis_40[[1]]$ICC[1],
                                              icc_aos.xis_50[[1]]$ICC[1],
                                              icc_aos.xis_60[[1]]$ICC[1],
                                              icc_aos.xis_70[[1]]$ICC[1],
                                              icc_aos.xis_80[[1]]$ICC[1],
                                              icc_aos.xis_90[[1]]$ICC[1],
                                              icc_aos.xis_100[[1]]$ICC[1],
                                              icc_aos.xis_200[[1]]$ICC[1],
                                              icc_aos.xis_250[[1]]$ICC[1],
                                              icc_aos.xis_300[[1]]$ICC[1]))

```

```

    icc_aos.xis_250[[1]]$ICC[1]))}

out <- cbind(rel.curve_aos.alpha, rel.curve_aos.beta, rel.curve_aos.xi)

return(out)

#rm(sample.size,method,log.odds,n.resample,n.bootstraps,trials,n,t,temp,data.out,aos.alphas_10,
#aos.alphas_20,aos.alphas_30,aos.alphas_40,aos.alphas_50,aos.alphas_60,aos.alphas_70,aos.alphas_80,
#aos.alphas_90,aos.alphas_100,aos.alphas_200,aos.alphas_250,aos.betas_10,aos.betas_20,aos.betas_30,
#aos.betas_40,aos.betas_50,aos.betas_60,aos.betas_70,aos.betas_80,aos.betas_90,aos.betas_100,
#aos.betas_200,aos.betas_250,aos.xi_10,aos.xi_20,aos.xi_30,aos.xi_40,aos.xi_50,aos.xi_60,aos.xi_70,
#aos.xi_80,aos.xi_90,aos.xi_100,aos.xi_200,aos.xi_250,rel.curve_aos.alpha,rel.curve_aos.beta,
#rel.curve_aos.xi)
}
#####
##### Calculate reliability curves for non-bootstrapped cred. estimates #####
#####

## Set seed to ensure replicability

## Number of trials == 10
set.seed(826)
rc.nonboot_nt.10 <- gen.rc_non.boot("mean", TRUE, n.resample=10)

## Number of trials == 20
set.seed(826)
rc.nonboot_nt.20 <- gen.rc_non.boot("mean", TRUE, n.resample=20)

## Number of trials == 30
set.seed(826)
rc.nonboot_nt.30 <- gen.rc_non.boot("mean", TRUE, n.resample=30)

#####
##### Calculate reliability curves for bootstrapped cred. estimates #####
#####

## Sample size == 10
set.seed(826)
rc.boot_ss.10 <- gen.rc_boot(10, "mean", TRUE)

## Sample size == 20
set.seed(826)
rc.boot_ss.20 <- gen.rc_boot(20, "mean", TRUE)

## Sample size == 30
set.seed(826)
rc.boot_ss.30 <- gen.rc_boot(30, "mean", TRUE)

## Sample size == 40
set.seed(826)
rc.boot_ss.40 <- gen.rc_boot(40, "mean", TRUE)

## Sample size == 50
set.seed(826)
rc.boot_ss.50 <- gen.rc_boot(50, "mean", TRUE)

```

```

#####
##### Plot Non-Bootstrapped Reliability Curves
#####
par(mar=c(5,5,1.5,3))

plot(x=rc.nonboot_nt.30$n.cal, y=rc.nonboot_nt.30$rel.curve.alpha,
      xlim=c(0,50), ylim=c(0,1), col="red", lty=1, lwd=2,
      xlab=expression(paste("Calibration Sample Size (", n[cal], ")"), sep="")),
      ylab="Intraclass Correlation",
      #main="Reliability of Non-Bootstrapped Credibility Estimates (GJP Data)",
      cex.lab=1.4, cex.axis=1.2, cex.main=1.2, cex=1.5)

points(x=rc.nonboot_nt.30$n.cal, y=rc.nonboot_nt.30$rel.curve.beta,
       xlim=c(0,50), ylim=c(0,1), col="blue", lwd=2, cex=1.5)

points(x=rc.nonboot_nt.30$n.cal, y=rc.nonboot_nt.30$rel.curve.xi,
       xlim=c(0,50), ylim=c(0,1), col="green", lwd=2, cex=1.5)

lines(x=rc.nonboot_nt.30$n.cal, y=rc.nonboot_nt.30$rel.curve.alpha, col="red", lty=2, lwd=2)
lines(x=rc.nonboot_nt.30$n.cal, y=rc.nonboot_nt.30$rel.curve.beta, col="blue", lty=2, lwd=2)
lines(x=rc.nonboot_nt.30$n.cal, y=rc.nonboot_nt.30$rel.curve.xi, col="green", lty=2, lwd=2)

abline(h=0.75, col="black", lty=2, lwd=1)

legend(x=35, y=0.42, col=c("red","blue","green"), lty=2,lwd=2, cex=1.2, y.intersp=1.5,
       legend=c(expression(paste("Est. Bias (", hat(alpha)[ "in"]^"**", ")"), sep="")),
       expression(paste("Est. Expertise (", hat(beta)[ "in"]^"**", ")"), sep="")),
       expression(paste("Est. Consistency (", hat(sigma)[ "in"]^"**", ")"), sep="")))

#####
##### Plot Bootstrapped Reliability Curves
#####

## AOS Alpha
par(mar=c(5,5,1.5,3))

plot(x=rc.boot_ss.10$n.resample, y=rc.boot_ss.10$icc_aos.alpha,
      xlim=c(0,250), ylim=c(0.73,1), col="red", lty=1, lwd=2,
      xlab=expression(paste("Number of Bootstrap Trials (", n[boot], ")"), sep="")),
      ylab="Intraclass Correlation",
      #main=expression(paste("Reliability of Bootstrapped Estimates of Bias (", hat(alpha)[i], "),
      GJP Data")),
      cex.lab=1.4, cex.axis=1.2, cex.main=1.5, cex=1.5)

points(x=rc.boot_ss.20$n.resample, y=rc.boot_ss.20$icc_aos.alpha,
       xlim=c(0,250), ylim=c(0.73,1), col="orange", lwd=2, cex=1.5)

points(x=rc.boot_ss.30$n.resample, y=rc.boot_ss.30$icc_aos.alpha,
       xlim=c(0,250), ylim=c(0.73,1), col="green", lwd=2, cex=1.5)

points(x=rc.boot_ss.40$n.resample, y=rc.boot_ss.40$icc_aos.alpha,
       xlim=c(0,250), ylim=c(0.73,1), col="blue", lwd=2, cex=1.5)

points(x=rc.boot_ss.50$n.resample, y=rc.boot_ss.50$icc_aos.alpha,
       xlim=c(0,250), ylim=c(0.73,1), col="purple", lwd=2, cex=1.5)

lines(x=rc.boot_ss.10$n.resample, y=rc.boot_ss.10$icc_aos.alpha, col="red", lty=2, lwd=2)
lines(x=rc.boot_ss.20$n.resample, y=rc.boot_ss.20$icc_aos.alpha, col="orange", lty=2, lwd=2)
lines(x=rc.boot_ss.30$n.resample, y=rc.boot_ss.30$icc_aos.alpha, col="green", lty=2, lwd=2)
lines(x=rc.boot_ss.40$n.resample, y=rc.boot_ss.40$icc_aos.alpha, col="blue", lty=2, lwd=2)
lines(x=rc.boot_ss.50$n.resample, y=rc.boot_ss.50$icc_aos.alpha, col="purple", lty=2, lwd=2)

```

```

abline(h=0.75, col="black", lty=2, lwd=1)
axis.break(axis=2, breakpos=0.73, style="slash", brw=0.02)

legend(x=155, y=0.87, col=c("red","orange","green","blue","purple"), lty=2, lwd=2, cex=1.2,
       legend=c(expression(paste("Cal. sample size (", n[cal], ") = 10", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 20", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 30", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 40", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 50", sep=""))))

## AOS Beta
par(mar=c(5,5,1.5,3))

plot(x=rc.boot_ss.10$n.resample, y=rc.boot_ss.10$icc_aos.beta,
      xlim=c(0,250), ylim=c(0.73,1), col="red", lty=1, lwd=2,
      xlab=expression(paste("Number of Bootstrap Trials (", n[boot], ")"), sep=""),
      ylab="Intraclass Correlation",
      #main=expression(paste("Reliability of Bootstrapped Estimates of Bias (", hat(beta)[i], "), GJP
      Data")),
      cex.lab=1.4, cex.axis=1.2, cex.main=1.5, cex=1.5)

points(x=rc.boot_ss.20$n.resample, y=rc.boot_ss.20$icc_aos.beta,
       xlim=c(0,250), ylim=c(0.73,1), col="orange", lwd=2, cex=1.5)

points(x=rc.boot_ss.30$n.resample, y=rc.boot_ss.30$icc_aos.beta,
       xlim=c(0,250), ylim=c(0.73,1), col="green", lwd=2, cex=1.5)

points(x=rc.boot_ss.40$n.resample, y=rc.boot_ss.40$icc_aos.beta,
       xlim=c(0,250), ylim=c(0.73,1), col="blue", lwd=2, cex=1.5)

points(x=rc.boot_ss.50$n.resample, y=rc.boot_ss.50$icc_aos.beta,
       xlim=c(0,250), ylim=c(0.73,1), col="purple", lwd=2, cex=1.5)

lines(x=rc.boot_ss.10$n.resample, y=rc.boot_ss.10$icc_aos.beta, col="red", lty=2, lwd=2)
lines(x=rc.boot_ss.20$n.resample, y=rc.boot_ss.20$icc_aos.beta, col="orange", lty=2, lwd=2)
lines(x=rc.boot_ss.30$n.resample, y=rc.boot_ss.30$icc_aos.beta, col="green", lty=2, lwd=2)
lines(x=rc.boot_ss.40$n.resample, y=rc.boot_ss.40$icc_aos.beta, col="blue", lty=2, lwd=2)
lines(x=rc.boot_ss.50$n.resample, y=rc.boot_ss.50$icc_aos.beta, col="purple", lty=2, lwd=2)

abline(h=0.75, col="black", lty=2, lwd=1)
axis.break(axis=2, breakpos=0.73, style="slash", brw=0.02)

legend(x=155, y=0.87, col=c("red","orange","green","blue","purple"), lty=2, lwd=2, cex=1.2,
       legend=c(expression(paste("Cal. sample size (", n[cal], ") = 10", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 20", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 30", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 40", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 50", sep=""))))

## AOS Xi
par(mar=c(5,5,1.5,3))

plot(x=rc.boot_ss.10$n.resample, y=rc.boot_ss.10$icc_aos.xi,
      xlim=c(0,250), ylim=c(0.73,1), col="red", lty=1, lwd=2,
      xlab=expression(paste("Number of Bootstrap Trials (", n[boot], ")"), sep=""),
      ylab="Intraclass Correlation",
      #main=expression(paste("Reliability of Bootstrapped Estimates of Bias (", hat(xi)[i], "), GJP
      Data")),
      cex.lab=1.4, cex.axis=1.2, cex.main=1.5, cex=1.5)

points(x=rc.boot_ss.20$n.resample, y=rc.boot_ss.20$icc_aos.xi,
       xlim=c(0,250), ylim=c(0.73,1), col="orange", lwd=2, cex=1.5)

points(x=rc.boot_ss.30$n.resample, y=rc.boot_ss.30$icc_aos.xi,
       xlim=c(0,250), ylim=c(0.73,1), col="green", lwd=2, cex=1.5)

points(x=rc.boot_ss.40$n.resample, y=rc.boot_ss.40$icc_aos.xi,
       xlim=c(0,250), ylim=c(0.73,1), col="blue", lwd=2, cex=1.5)

points(x=rc.boot_ss.50$n.resample, y=rc.boot_ss.50$icc_aos.xi,
       xlim=c(0,250), ylim=c(0.73,1), col="purple", lwd=2, cex=1.5)

```

```

xlim=c(0,250), ylim=c(0.73,1), col="green", lwd=2, cex=1.5)
points(x=rc.boot_ss.40$n.resample, y=rc.boot_ss.40$icc_aos.xi,
       xlim=c(0,250), ylim=c(0.73,1), col="blue", lwd=2, cex=1.5)

points(x=rc.boot_ss.50$n.resample, y=rc.boot_ss.50$icc_aos.xi,
       xlim=c(0,250), ylim=c(0.73,1), col="purple", lwd=2, cex=1.5)

lines(x=rc.boot_ss.10$n.resample, y=rc.boot_ss.10$icc_aos.xi, col="red", lty=2, lwd=2)
lines(x=rc.boot_ss.20$n.resample, y=rc.boot_ss.20$icc_aos.xi, col="orange", lty=2, lwd=2)
lines(x=rc.boot_ss.30$n.resample, y=rc.boot_ss.30$icc_aos.xi, col="green", lty=2, lwd=2)
lines(x=rc.boot_ss.40$n.resample, y=rc.boot_ss.40$icc_aos.xi, col="blue", lty=2, lwd=2)
lines(x=rc.boot_ss.50$n.resample, y=rc.boot_ss.50$icc_aos.xi, col="purple", lty=2, lwd=2)

abline(h=0.75, col="black", lty=2, lwd=1)

axis.break(axis=2, breakpos=0.73, style="slash", brw=0.02)

legend(x=155, y=0.87, col=c("red","orange","green","blue","purple"), lty=2, lwd=2, cex=1.2,
       legend=c(expression(paste("Cal. sample size (", n[cal], ") = 10", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 20", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 30", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 40", sep="")),
                expression(paste("Cal. sample size (", n[cal], ") = 50", sep=""))))

```

```

#####
# Assessing Credibility in Subjective Probability Judgment #
# Correlates of Credibility, Good Judgment Data #
# (GJP Validity; GJP enrichment vs. credibility) #
# Josh Baker #
# jbak@sas.upenn.edu #
#####

#####
# DATA CLEANING AND PREPARATION
#
#####

##### Set Working directory and load relevant libraries #####
library(haven)
library(psych)
library(glmnet)

setwd("~/yourPath")

#####

##### Import and Clean Forecast Data #####
fcasts <- read.csv("fcasts_small.file.csv")

## Convert factors to numeric or character data types (and trim last two characters from each element of "ifp_id")
fcasts$ifp_id <- as.numeric(substr(as.character(fccasts$ifp_id),1,4))
fcasts$answer_option <- as.character(fccasts$answer_option)
fcasts$date <- as.character(fccasts$date)
fcasts$timestamp <- as.character(fccasts$timestamp)
fcasts$outcome <- as.character(fccasts$outcome)
fcasts$g.tnt <- as.character(fccasts$g.tnt)
fcasts$mod.tag <- as.character(fccasts$mod.tag)
fcasts$date.closed <- as.character(fccasts$date.closed)
fcasts$time.horiz <- as.character(fccasts$time.horiz)

## Subset fccasts to answer_option=="a" (only forecasts for focal outcomes, not complements)
fcasts <- subset(fccasts, answer_option=="a")

## Subset to columns that are relevant to correlational analyses
fcasts <- fccasts[which(is.element(colnames(fccasts),c("ifp_id",
                                                       "user_id",
                                                       "date",
                                                       "g.tnt",
                                                       "mod.tag",
                                                       "outcome",
                                                       "val.unrounded",
                                                       "bs.unrounded")))]]

## Rename some columns for consistency/expediency
colnames(fccasts)[which(colnames(fccasts)=="ifp_id")] <- "ifp.id"
colnames(fccasts)[which(colnames(fccasts)=="user_id")] <- "gjp.id"
colnames(fccasts)[which(colnames(fccasts)=="val.unrounded")] <- "prob"
colnames(fccasts)[which(colnames(fccasts)=="bs.unrounded")] <- "bs"

```

```

## Recode "outcome" to binary ("a" = 1; "b" = 0)
fcasts$outcome <- ifelse(fccasts$outcome=="a",1,0)

## Round all prob values to two places to fix a floating point rounding issue
fcasts$prob <- round(fccasts$prob,2)

## Reorder fccasts by gjp.id
fcasts <- fccasts[order(fccasts$gjp.id),]

## Re-order columns, for clarity
fcasts <- fccasts[c(2,5,3,1,7,8,4,6)]

## Extract GJP IDs as a vector
gjp.ids <- unique(fccasts$gjp.id)
gjp.ids <- gjp.ids[order(gjp.ids)] #Sort in ascending order

#####
##### Import Recal_50 Data #####
#####

recal_50 <- read.csv("recal_50.csv")

## Extract recal_50 IDs as a vector
r50.ids <- unique(recal_50$gjp.id)
r50.ids <- r50.ids[order(r50.ids)] #Sort in ascending order

#####
##### Import and Clean Individual Differences Data (recovered from Jeff) #####
#####

ind.diffs_jeff <- read_dta("ind.diffs.dta")

## Subset to only those columns that are relevant to correlates of credibility analyses
ind.diffs_jeff <- ind.diffs_jeff[which(is.element(colnames(ind.diffs_jeff), c("gjpid",
                                                                           "nifpsaddressed",
                                                                           "avgnfcastsperifp",
                                                                           "granscore",
                                                                           "ed_yr3",
                                                                           "ed_yr1_yr2",
                                                                           "y4education",
                                                                           "numeracy_score_yr3",
                                                                           "numeracy_score_yr1_yr2",
                                                                           "y4numeracyscore",
                                                                           "raven_adjusted_score",
                                                                           "y4ravenscore",
                                                                           "crt_score_yr3",
                                                                           "y4crtscore",
                                                                           "foxhedge_single",
                                                                           "nfcog_score_yr1_yr2",
                                                                           "y4nfccogscore",
                                                                           "thresh"))]

## Rename ind.diffs_jeff columns for clarity/consistency
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="gjpid")] <- "gjp.id"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="nifpsaddressed")] <- "n.ifps"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="avgnfcastsperifp")] <- "avg.n.updates"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="granscore")] <- "prop.fine.grained"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="ed_yr3")] <- "ed.y3"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="ed_yr1_yr2")] <- "ed.y1.y2"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="y4education")] <- "ed.y4"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="numeracy_score_yr3")] <- "numeracy.y3"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="numeracy_score_yr1_yr2")] <- "numeracy.y1.y2"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="y4numeracyscore")] <- "numeracy.y4"

```

```

colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="raven_adjusted_score")] <- "adj.ravens.y3"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="y4ravenscore")] <- "ravens.y4"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="crt_score_yr3")] <- "crt.y3"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="y4crtscore")] <- "crt.y4"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="foxhedge_single")] <- "fox.hedgehog"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="nfcog_score_yr1_yr2")] <- "nfcog.y1.y2"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="y4nfcogscore")] <- "nfcog.y4"
colnames(ind.diffs_jeff)[which(colnames(ind.diffs_jeff)=="thresh")] <- "b.star"

## Subset to only those forecasters in recal_50
ind.diffs_jeff <- subset(ind.diffs_jeff, is.element(ind.diffs_jeff$gjp.id, recal_50$gjp.id))

## Sort by gjp.id
ind.diffs_jeff <- ind.diffs_jeff[order(ind.diffs_jeff$gjp.id),]

## Convert ind.diffs_jeff to a simple data frame (vs. a multi-class tibble object)
ind.diffs_jeff <- as.data.frame(ind.diffs_jeff)

#####
# Import and Clean Individual Differences Data from GJP
#####

ind.diffs_gjp <- read.csv("individual.diffs_across.yrs.csv", stringsAsFactors=F)

## Subset to only those columns that are relevant to correlates of credibility analyses
ind.diffs_gjp <- ind.diffs_gjp[which(is.element(colnames(ind.diffs_gjp), c("user_id",
  "age",
  "gender",
  "aomt_score")))]

## Rename ind.diffs_gjp columns for clarity/consistency
colnames(ind.diffs_gjp)[which(colnames(ind.diffs_gjp)=="user_id")] <- "gjp.id"
colnames(ind.diffs_gjp)[which(colnames(ind.diffs_gjp)=="gender")] <- "male"
colnames(ind.diffs_gjp)[which(colnames(ind.diffs_gjp)=="aomt_score")] <- "aomt.y3"

## Subset to only those forecasters in recal_50
ind.diffs_gjp <- subset(ind.diffs_gjp, is.element(ind.diffs_gjp$gjp.id, recal_50$gjp.id))

## Sort by gjp.id
ind.diffs_gjp <- ind.diffs_gjp[order(ind.diffs_gjp$gjp.id),]

#####
# Elementary Functions for merging, subsetting, and extracting individual difference data
#####

## Function for calculating the number of unique elements in a vector #####
n.unique <- function(vector){length(unique(vector))}

## Function for returning the first element in a vector #####
first <- function(vector){vector[1]}

## Function for identifying complete data.frame rows (i.e., rows without missingness)
is.complete <- function(data.in){
  i <- 1
  out <- vector()

  while(i <= nrow(data.in)){
    out[i] <- 1 - (as.numeric(NA %in% as.matrix(data.in[i,])))
    i <- i + 1
  }

  return(out)
}

#rm(data.in,i,out)
}

```

```

## Function to mimic the functionality of excel's vlookup
vlookup <- function(source.df, target.df, match.col.name, return.col.name){
  source_match.col.num <- as.integer(which(colnames(source.df)==match.col.name))
  target_match.col.num <- as.integer(which(colnames(target.df)==match.col.name))
  target_return.col.num <- as.integer(which(colnames(target.df)==return.col.name))

  i <- 1
  out <- vector()

  while(i <= nrow(source.df)){
    if(is.element(source.df[i,source_match.col.num],target.df[,target_match.col.num])){
      out[i] <-
      target.df[which(target.df[,target_match.col.num]==source.df[i,source_match.col.num]),target_return.col.num]
    } else{
      out[i] <- NA
    }

    i <- i+1
  }

  return(out)
}

#rm(source.df,target.df,match.col.name,return.col.name,source_match.col.num,target_match.col.num,tar
get_return.col.num,i,out)
}

## Function for effects coding categorical variables
ec.expand <- function(data.in, category.name, ref=NULL){
  levels <- levels(as.factor(data.in))

  if(is.null(ref)){
    ref <- levels[1]
  }

  ec.levels <- setdiff(levels, ref)

  data.out <- matrix(nrow=length(data.in),ncol=length(ec.levels))

  i <- 1
  while(i <= length(ec.levels)){
    data.out[,i] <- ifelse(data.in==ec.levels[i],1,ifelse(data.in==ref,-1,0))
    i <- i+1
  }

  data.out <- as.data.frame(data.out)
  colnames(data.out) <- paste(category.name, ec.levels, "ec", sep=".")

  return(data.out)
}

#rm(data.in,category.name,ref,levels,ec.levels,data.out,i)
}

## Function for conducting mean imputation on data.frame columns
impute.means <- function(data.in){
  data.out <- as.matrix(data.in)
  i <- 1

  while(i <= ncol(data.out)){
    data.out[,i] <- ifelse(is.na(data.out[,i]),mean(data.out[,i], na.rm=T), data.out[,i])
    i <- i+1
  }

  data.out <- as.data.frame(data.out)

  return(data.out)
}

```

```

    #rm(data.in,data.out)
}

#####
##### Create a raw, master version of ind.diffs
#####

## Combine data from ind.diffs_jeff and ind.diffs_gjp
ind.diffs_raw <- ind.diffs_jeff
ind.diffs_raw <- cbind(ind.diffs_raw, aomt.y3=vlookup(ind.diffs_jeff, ind.diffs_gjp, "gjp.id",
"aomt.y3"))
ind.diffs_raw <- cbind(ind.diffs_raw, age=vlookup(ind.diffs_jeff, ind.diffs_gjp, "gjp.id", "age"))
ind.diffs_raw <- cbind(ind.diffs_raw, male=vlookup(ind.diffs_jeff, ind.diffs_gjp, "gjp.id", "male"))

## Calculate mean and median brier scores for each person, and merge into ind.diffs_raw
sub <- subset(fcasts, is.element(fcasts$gjp.id, recal_50$gjp.id))

i <- 1
avg.bs <- vector()
med.bs <- vector()

while(i <= length(r50.ids)){
  sub.fc <- subset(sub, gjp.id==r50.ids[i])
  sub.fc <- sub.fc[order(sub.fc$ifp.id),]
  user.ifps <- unique(sub.fc$ifp.id)
  j <- 1
  ifp.agg <- vector()

  while(j <= length(user.ifps)){
    sub.ifp <- subset(sub.fc, ifp.id==user.ifps[j])
    ifp.agg[j] <- mean(as.vector(by(sub.ifp$bs, sub.ifp$date, mean)))
    j <- j+1
  }

  avg.bs[i] <- mean(ifp.agg)
  med.bs[i] <- median(ifp.agg)
  i <- i+1
}

ind.diffs_raw <- cbind(ind.diffs_raw, avg.bs=avg.bs, med.bs=med.bs)
rm(i,avg.bs,med.bs,sub.fc,user.ifps,j,ifp.agg,sub.ifp)

## Merge-in and effects code teaming, training, and super data from fcasts
ind.diffs_raw <- cbind(ind.diffs_raw, g.tnt=as.vector(by(sub$g.tnt, sub$gjp.id, first)))
ind.diffs_raw$g.tnt <- levels(as.factor(fcasts$g.tnt))[ind.diffs_raw$g.tnt]
ind.diffs_raw$g.tnt <- sub("\\.", "", ind.diffs_raw$g.tnt)
ind.diffs_raw <- cbind(ind.diffs_raw, ec.expand(ind.diffs_raw$g.tnt, category.name="condition",
ref="int"))

## Merge education data across years to reduce missingness, and recode
temp <-
as.vector(apply(ind.diffs_raw[which(is.element(colnames(ind.diffs_raw),c("ed.y3","ed.y1.y2","ed.y4"))], 1, max, na.rm=T))
temp <- ifelse(temp==Inf,0,temp)
temp <- as.character(temp)
temp <- ifelse(temp=="0","other",temp)
temp <- ifelse(temp=="1","bachelors",temp)
temp <- ifelse(temp=="2","masters",temp)
temp <- ifelse(temp=="3","doctorate",temp)
ind.diffs_raw <- cbind(ind.diffs_raw, educ=temp)

## Effects code education
ind.diffs_raw <- cbind(ind.diffs_raw, ec.expand(ind.diffs_raw$educ, category.name="educ",
ref="other"))

## Merge-in credibility data

```

```

ind.diffs_raw <- cbind(aos.alpha=vlookup(ind.diffs_raw, recal_50, "gjp.id", "aos.alpha"),
                       sd.alpha=vlookup(ind.diffs_raw, recal_50, "gjp.id", "sd.alpha"),
                       aos.beta=vlookup(ind.diffs_raw, recal_50, "gjp.id", "aos.beta"),
                       sd.beta=vlookup(ind.diffs_raw, recal_50, "gjp.id", "sd.beta"),
                       aos.xi=vlookup(ind.diffs_raw, recal_50, "gjp.id", "aos.xi"),
                       sd.xi=vlookup(ind.diffs_raw, recal_50, "gjp.id", "sd.xi"),
                       ind.diffs_raw)

## Adjust mos.alpha and mos.beta to reflect difference from normative values
ind.diffs_raw$aos.alpha <- abs(ind.diffs_raw$aos.alpha - 0)
ind.diffs_raw$aos.beta <- abs(ind.diffs_raw$aos.beta - 1)
colnames(ind.diffs_raw)[which(colnames(ind.diffs_raw)=="aos.alpha")] <- "ane.aos.alpha" #ANE =
absolute normative error
colnames(ind.diffs_raw)[which(colnames(ind.diffs_raw)=="aos.beta")] <- "ane.aos.beta" #ANE =
absolute normative error

## Cleanup and reorganize
rm(sub,temp)
ind.diffs_raw <-
ind.diffs_raw[c(7,1:6,28:29,33,31,32,34,24,8:10,14,13,20,15,21,16,22,17,18,23,25:27,12,11,19,35,36,3
8,37)]

##### Create four versions of ind.diffs, varying mean imputation and averaging across years
#####

##### No averaging (primarily y3 data), no mean imputation #####
ind.diffs_na.ni <- ind.diffs_raw[c(1:17,19,21,23,25,26,28:30,35:37)]
ind.diffs_na.ni <- cbind(ind.diffs_na.ni, complete=is.complete(ind.diffs_na.ni))
ind.diffs_na.ni <- subset(ind.diffs_na.ni, complete==1)
ind.diffs_na.ni <- ind.diffs_na.ni[-which(colnames(ind.diffs_na.ni)=="complete")]

##### No averaging (primarily y3 data), with mean imputation #####
ind.diffs_na.wi <- ind.diffs_raw[c(1:17,19,21,23,25,26,28:30,35:37)]
ind.diffs_na.wi <- impute.means(ind.diffs_na.wi)

##### Averaging across data from all available years, no mean imputation #####
ind.diffs_wa.ni <- ind.diffs_raw

# Average numeracy scores across years
temp <-
as.vector(apply(ind.diffs_wa.ni[which(is.element(colnames(ind.diffs_wa.ni),c("numeracy.y1.y2","numer
acy.y3","numeracy.y4")))], 1, mean, na.rm=T))
temp <- ifelse(is.nan(temp),NA,temp)
ind.diffs_wa.ni <- cbind(ind.diffs_wa.ni, avg.numeracy=temp)

# Adjust y4 Raven's scores and average across years
ind.diffs_wa.ni$ravens.y4 <- ind.diffs_wa.ni$ravens.y4 / 12
colnames(ind.diffs_wa.ni)[which(colnames(ind.diffs_wa.ni)=="ravens.y4")] <- "adj.ravens.y4"
temp <-
as.vector(apply(ind.diffs_wa.ni[which(is.element(colnames(ind.diffs_wa.ni),c("adj.ravens.y3","adj.ra
vens.y4")))], 1, mean, na.rm=T))
temp <- ifelse(is.nan(temp),NA,temp)
ind.diffs_wa.ni <- cbind(ind.diffs_wa.ni, avg.adj.ravens=temp)

# Average CRT scores across years (y2 excluded because not extended CRT)
temp <-
as.vector(apply(ind.diffs_wa.ni[which(is.element(colnames(ind.diffs_wa.ni),c("crt.y3","crt.y4")))],
1, mean, na.rm=T))
temp <- ifelse(is.nan(temp),NA,temp)
ind.diffs_wa.ni <- cbind(ind.diffs_wa.ni, avg.crt=temp)

# Average nfcog scores across years
temp <-
as.vector(apply(ind.diffs_wa.ni[which(is.element(colnames(ind.diffs_wa.ni),c("nfcog.y1.y2","nfcog.y4
")))], 1, mean, na.rm=T))
temp <- ifelse(is.nan(temp),NA,temp)

```

```

ind.diffs_wa.ni <- cbind(ind.diffs_wa.ni, avg.nfcog=temp)

# Subset to final, merged/averaged data
ind.diffs_wa.ni <- ind.diffs_wa.ni[c(1:17,38:40,25,41,28:30,35:37)]

# Subset to only complete cases
ind.diffs_wa.ni <- cbind(ind.diffs_wa.ni, complete=is.complete(ind.diffs_wa.ni))
ind.diffs_wa.ni <- subset(ind.diffs_wa.ni, complete==1)
ind.diffs_wa.ni <- ind.diffs_wa.ni[-which(colnames(ind.diffs_wa.ni)=="complete")]

##### Averaging across data from all available years, with mean imputation #####
ind.diffs_wa.wi <- ind.diffs_raw
ind.diffs_wa.wi <- ind.diffs_wa.wi[c(1:30,35:37)]
ind.diffs_wa.wi <- impute.means(ind.diffs_wa.wi)

# Average numeracy scores across years
temp <-
as.vector(apply(ind.diffs_wa.wi[which(is.element(colnames(ind.diffs_wa.wi),c("numeracy.y1.y2","numeracy.y3","numeracy.y4")))], 1, mean, na.rm=T))
temp <- ifelse(is.nan(temp),NA,temp)
ind.diffs_wa.wi <- cbind(ind.diffs_wa.wi, avg.numeracy=temp)

# Adjust y4 Raven's scores and average across years
ind.diffs_wa.wi$ravens.y4 <- ind.diffs_wa.wi$ravens.y4 / 12
colnames(ind.diffs_wa.wi)[which(colnames(ind.diffs_wa.wi)=="ravens.y4")] <- "adj.ravens.y4"
temp <-
as.vector(apply(ind.diffs_wa.wi[which(is.element(colnames(ind.diffs_wa.wi),c("adj.ravens.y3","adj.ravens.y4")))], 1, mean, na.rm=T))
temp <- ifelse(is.nan(temp),NA,temp)
ind.diffs_wa.wi <- cbind(ind.diffs_wa.wi, avg.adj.ravens=temp)

# Average CRT scores across years (y2 excluded because not extended CRT)
temp <-
as.vector(apply(ind.diffs_wa.wi[which(is.element(colnames(ind.diffs_wa.wi),c("crt.y3","crt.y4")))], 1, mean, na.rm=T))
temp <- ifelse(is.nan(temp),NA,temp)
ind.diffs_wa.wi <- cbind(ind.diffs_wa.wi, avg.crt=temp)

# Average nfcog scores across years
temp <-
as.vector(apply(ind.diffs_wa.wi[which(is.element(colnames(ind.diffs_wa.wi),c("nfcog.y1.y2","nfcog.y4")))], 1, mean, na.rm=T))
temp <- ifelse(is.nan(temp),NA,temp)
ind.diffs_wa.wi <- cbind(ind.diffs_wa.wi, avg.nfcog=temp)

# Subset to final, merged/averaged data
ind.diffs_wa.wi <- ind.diffs_wa.wi[c(1:17,34:37,25,28:33)]
rm(temp)

#####
##### Standardize all non-categorical variables in each set of ind.diffs
#####

## ind.diffs_na.ni (no averaging, no imputation)
ind.diffs_na.ni$ane-aos.alpha <- scale(ind.diffs_na.ni$ane-aos.alpha)
ind.diffs_na.ni$sd.alpha <- scale(ind.diffs_na.ni$sd.alpha)
ind.diffs_na.ni$ane-aos.beta <- scale(ind.diffs_na.ni$ane-aos.beta)
ind.diffs_na.ni$sd.beta <- scale(ind.diffs_na.ni$sd.beta)
ind.diffs_na.ni$aos.xi <- scale(ind.diffs_na.ni$aos.xi)
ind.diffs_na.ni$sd.xi <- scale(ind.diffs_na.ni$sd.xi)
ind.diffs_na.ni$avg.bs <- scale(ind.diffs_na.ni$avg.bs)
ind.diffs_na.ni$med.bs <- scale(ind.diffs_na.ni$med.bs)
ind.diffs_na.ni$b.star <- scale(ind.diffs_na.ni$b.star)
ind.diffs_na.ni$n.ifps <- scale(ind.diffs_na.ni$n.ifps)
ind.diffs_na.ni$avg.n.updates <- scale(ind.diffs_na.ni$avg.n.updates)
ind.diffs_na.ni$prop.fine.grained <- scale(ind.diffs_na.ni$prop.fine.grained)
ind.diffs_na.ni$numeracy.y3 <- scale(ind.diffs_na.ni$numeracy.y3)
```

```

ind.diffs_na.ni$adj.ravens.y3 <- scale(ind.diffs_na.ni$adj.ravens.y3)
ind.diffs_na.ni$crt.y3 <- scale(ind.diffs_na.ni$crt.y3)
ind.diffs_na.ni$fox.hedgehog <- scale(ind.diffs_na.ni$fox.hedgehog)
ind.diffs_na.ni$nfcog.y1.y2 <- scale(ind.diffs_na.ni$nfcog.y1.y2)
ind.diffs_na.ni$aomt.y3 <- scale(ind.diffs_na.ni$aomt.y3)
ind.diffs_na.ni$age <- scale(ind.diffs_na.ni$age)

## ind.diffs_na.wi (no averaging, with imputation)
ind.diffs_na.wi$ane-aos.alpha <- scale(ind.diffs_na.wi$ane-aos.alpha)
ind.diffs_na.wi$sd.alpha <- scale(ind.diffs_na.wi$sd.alpha)
ind.diffs_na.wi$ane-aos.beta <- scale(ind.diffs_na.wi$ane-aos.beta)
ind.diffs_na.wi$sd.beta <- scale(ind.diffs_na.wi$sd.beta)
ind.diffs_na.wi$aos.xi <- scale(ind.diffs_na.wi$aos.xi)
ind.diffs_na.wi$sd.xi <- scale(ind.diffs_na.wi$sd.xi)
ind.diffs_na.wi$avg.bs <- scale(ind.diffs_na.wi$avg.bs)
ind.diffs_na.wi$med.bs <- scale(ind.diffs_na.wi$med.bs)
ind.diffs_na.wi$b.star <- scale(ind.diffs_na.wi$b.star)
ind.diffs_na.wi$n.ifps <- scale(ind.diffs_na.wi$n.ifps)
ind.diffs_na.wi$avg.n.updates <- scale(ind.diffs_na.wi$avg.n.updates)
ind.diffs_na.wi$prop.fine.grained <- scale(ind.diffs_na.wi$prop.fine.grained)
ind.diffs_na.wi$numeracy.y3 <- scale(ind.diffs_na.wi$numeracy.y3)
ind.diffs_na.wi$adj.ravens.y3 <- scale(ind.diffs_na.wi$adj.ravens.y3)
ind.diffs_na.wi$crt.y3 <- scale(ind.diffs_na.wi$crt.y3)
ind.diffs_na.wi$fox.hedgehog <- scale(ind.diffs_na.wi$fox.hedgehog)
ind.diffs_na.wi$nfcog.y1.y2 <- scale(ind.diffs_na.wi$nfcog.y1.y2)
ind.diffs_na.wi$aomt.y3 <- scale(ind.diffs_na.wi$aomt.y3)
ind.diffs_na.wi$age <- scale(ind.diffs_na.wi$age)

## ind.diffs_wa.ni (with averaging, no imputation)
ind.diffs_wa.ni$ane-aos.alpha <- scale(ind.diffs_wa.ni$ane-aos.alpha)
ind.diffs_wa.ni$sd.alpha <- scale(ind.diffs_wa.ni$sd.alpha)
ind.diffs_wa.ni$ane-aos.beta <- scale(ind.diffs_wa.ni$ane-aos.beta)
ind.diffs_wa.ni$sd.beta <- scale(ind.diffs_wa.ni$sd.beta)
ind.diffs_wa.ni$aos.xi <- scale(ind.diffs_wa.ni$aos.xi)
ind.diffs_wa.ni$sd.xi <- scale(ind.diffs_wa.ni$sd.xi)
ind.diffs_wa.ni$avg.bs <- scale(ind.diffs_wa.ni$avg.bs)
ind.diffs_wa.ni$med.bs <- scale(ind.diffs_wa.ni$med.bs)
ind.diffs_wa.ni$b.star <- scale(ind.diffs_wa.ni$b.star)
ind.diffs_wa.ni$n.ifps <- scale(ind.diffs_wa.ni$n.ifps)
ind.diffs_wa.ni$avg.n.updates <- scale(ind.diffs_wa.ni$avg.n.updates)
ind.diffs_wa.ni$prop.fine.grained <- scale(ind.diffs_wa.ni$prop.fine.grained)
ind.diffs_wa.ni$avg.numeracy <- scale(ind.diffs_wa.ni$avg.numeracy)
ind.diffs_wa.ni$avg.adj.ravens <- scale(ind.diffs_wa.ni$avg.adj.ravens)
ind.diffs_wa.ni$avg.crt <- scale(ind.diffs_wa.ni$avg.crt)
ind.diffs_wa.ni$avg.nfcog <- scale(ind.diffs_wa.ni$avg.nfcog)
ind.diffs_wa.ni$fox.hedgehog <- scale(ind.diffs_wa.ni$fox.hedgehog)
ind.diffs_wa.ni$aomt.y3 <- scale(ind.diffs_wa.ni$aomt.y3)
ind.diffs_wa.ni$age <- scale(ind.diffs_wa.ni$age)

## ind.diffs_wa.wi (with averaging, with imputation)
ind.diffs_wa.wi$ane-aos.alpha <- scale(ind.diffs_wa.wi$ane-aos.alpha)
ind.diffs_wa.wi$sd.alpha <- scale(ind.diffs_wa.wi$sd.alpha)
ind.diffs_wa.wi$ane-aos.beta <- scale(ind.diffs_wa.wi$ane-aos.beta)
ind.diffs_wa.wi$sd.beta <- scale(ind.diffs_wa.wi$sd.beta)
ind.diffs_wa.wi$aos.xi <- scale(ind.diffs_wa.wi$aos.xi)
ind.diffs_wa.wi$sd.xi <- scale(ind.diffs_wa.wi$sd.xi)
ind.diffs_wa.wi$avg.bs <- scale(ind.diffs_wa.wi$avg.bs)
ind.diffs_wa.wi$med.bs <- scale(ind.diffs_wa.wi$med.bs)
ind.diffs_wa.wi$b.star <- scale(ind.diffs_wa.wi$b.star)
ind.diffs_wa.wi$n.ifps <- scale(ind.diffs_wa.wi$n.ifps)
ind.diffs_wa.wi$avg.n.updates <- scale(ind.diffs_wa.wi$avg.n.updates)
ind.diffs_wa.wi$prop.fine.grained <- scale(ind.diffs_wa.wi$prop.fine.grained)
ind.diffs_wa.wi$avg.numeracy <- scale(ind.diffs_wa.wi$avg.numeracy)
ind.diffs_wa.wi$avg.adj.ravens <- scale(ind.diffs_wa.wi$avg.adj.ravens)
ind.diffs_wa.wi$avg.crt <- scale(ind.diffs_wa.wi$avg.crt)
ind.diffs_wa.wi$avg.nfcog <- scale(ind.diffs_wa.wi$avg.nfcog)
ind.diffs_wa.wi$fox.hedgehog <- scale(ind.diffs_wa.wi$fox.hedgehog)
ind.diffs_wa.wi$aomt.y3 <- scale(ind.diffs_wa.wi$aomt.y3)
ind.diffs_wa.wi$age <- scale(ind.diffs_wa.wi$age)

```

```

#####
##### PAIRWISE CORRELATIONS
#
#####
##### Elementary functions for constructing correlation tables
#####

## Function for assigning stars on the basis of statistical significance
assign.stars <- function(p.value,include.ns=TRUE){
  if(p.value > 0.05){
    if(include.ns){
      out <- "n.s."
    } else{
      out <- ""
    }
  }
  if(p.value <= 0.05 & p.value > 0.01){
    out <- "*"
  }
  if(p.value <= 0.01 & p.value > 0.001){
    out <- "**"
  }
  if(p.value <= 0.001){
    out <- "***"
  }
}

return(out)
}

## Function for building a correlation table
gen.cor.table <- function(data, print.p=TRUE, include.ns=TRUE){
  require(psych)

  out <- matrix(nrow=nrow(data), ncol=ncol(data))
  rownames(out) <- colnames(data)
  colnames(out) <- colnames(data)

  i <- 1
  j <- 1

  while(i <= ncol(data)){
    while(j <= i){
      temp <- corr.test(as.matrix(data[,i]), as.matrix(data[,j]))
      est <- round(temp$r,3)
      p <- round(temp$p,4)
      sig <- assign.stars(p, include.ns=include.ns)
      p.sig <- paste(p,sig,sep=", ")
      if(print.p){
        out[i,j] <- paste(est, p.sig, sep=", ")
      } else{
        out[i,j] <- paste(est, sig, sep="")
      }

      j <- j+1
    }

    j <- 1
    i <- i+1
  }

  return(out)
}

```

```

    #rm(data,print.p,include.ns,out,i,j,temp,est,p,sig,p.sig,out)
}

#####
##### Generate and export correlation tables for each set of ind.diffs
#####

## No averaging, no imputation
cor.table_na.ni <- gen.cor.table(ind.diffs_na.ni, print.p=TRUE, include.ns=TRUE)
write.csv(cor.table_na.ni, "cor.table_na.ni.csv")

## No averaging, with imputation
cor.table_na.wi <- gen.cor.table(ind.diffs_na.wi, print.p=TRUE, include.ns=TRUE)
write.csv(cor.table_na.wi, "cor.table_na.wi.csv")

## With averaging, no imputation
cor.table_wa.ni <- gen.cor.table(ind.diffs_wa.ni, print.p=TRUE, include.ns=TRUE)
write.csv(cor.table_wa.ni, "cor.table_wa.ni.csv")

## With averaging, with imputation
cor.table_wa.wi <- gen.cor.table(ind.diffs_wa.wi, print.p=TRUE, include.ns=TRUE)
write.csv(cor.table_wa.wi, "cor.table_wa.wi.csv")

#####
##### EXPLORATORY REGRESSIONS
#
# Elementary functions for exploratory regression analyses
#####

## Function to condense lm output ##
condense.lm <- function(mod){
  summ <- summary(mod)
  out <- paste(round(as.vector(summ$coefficients[,1]),2), "
  (",round(as.vector(summ$coefficients[,2]),2), ")
  ",lapply(as.vector(summ$coefficients[,4]),assign.stars,include.ns=FALSE),sep="")
  out <- append(out, c(round(summ$r.squared,3), round(summ$adj.r.squared,3),
  round(sqrt(mean(summ$residuals^2)),3), round(AIC(mod),2), round(BIC(mod),2)))
  return(out)
}

## Function to create three-model table ("tmt") comparinf kitchen-sink, reduced, and ridge
regression approaches
tmt <- function(data.all,mod.ks,mod.red,mod.ridge){

  ## Extract coefficients from each model ##
  coeffs.all <- append(colnames(data.all), c("mult.r2","adj.r2","RMSE","AIC","BIC"))
  coeffs.all[1] <- "(Intercept)"
  coeffs.ks <- append(rownames(summary(mod.ks)$coefficients),
  c("mult.r2","adj.r2","RMSE","AIC","BIC"))
  coeffs.red <- append(rownames(summary(mod.red)$coefficients),
  c("mult.r2","adj.r2","RMSE","AIC","BIC"))
  coeffs.ridge <- append(rownames(summary(mod.ridge)$coefficients),
  c("mult.r2","adj.r2","RMSE","AIC","BIC"))

  ## Create spine for output ##

```

```

out <- data.frame(coeff=coeffs.all)

## Create contents for each table cell ##
items.ks <- condense.lm(mod.ks)
items.red <- condense.lm(mod.red)
items.ridge <- condense.lm(mod.ridge)

## Identify which table rows should be full for each model ##
slots.ks <- which(is.element(coeffs.all,coeffs.ks))
slots.red <- which(is.element(coeffs.all,coeffs.red))
slots.ridge <- which(is.element(coeffs.all,coeffs.ridge))

i <- 1
ks <- 1
red <- 1
ridge <- 1
col.ks <- vector()
col.red <- vector()
col.ridge <- vector()

while(i <= nrow(out)) {

  ## Fill kitchen sink model
  if(is.element(i,slots.ks)){
    col.ks[i] <- items.ks[ks]
    ks <- ks+1
  } else{
    col.ks[i] <- "-----"
  }

  ## Fill reduced model
  if(is.element(i,slots.red)){
    col.red[i] <- items.red[red]
    red <- red+1
  } else{
    col.red[i] <- "-----"
  }

  ## Fill ridge model
  if(is.element(i,slots.ridge)){
    col.ridge[i] <- items.ridge[ridge]
    ridge <- ridge+1
  } else{
    col.ridge[i] <- "-----"
  }

  i <- i+1
}

out <- cbind(out,
            kitchen.sink=col.ks,
            reduced=col.red,
            ridge.best_lambda.1se=col.ridge)

return(out)

#rm(data.all,mod.ks,mod.red,mod.ridge,coeffs.all,coeffs.ks,coeffs.red,coeffs.ridge,out,items.ks,item
s.red,
  #items.ridge,slots.ks,slots.red,slots.ridge,i,ks,red,ridge,col.ks,col.red,col.ridge)
}

## Function to fit baseline, kitchen sink, reduced, and ridge models, and generate a tmt table
gen.tmt <- function(dv, data.in, predictor.cols){
  require(glmnet)
}

```

```

if(is.numeric(dv)){
  outcome.col <- dv
}

if(is.character(dv)){
  outcome.col <- which(colnames(data.in)==dv)
}

## Gather kitchen sink data ##
data.ks <- data.in[c(outcome.col,predictor.cols)]

## Check for dv as duplicate in predictor columns, and remove if necessary
if(is.element(paste(colnames(data.ks)[1],".1",sep=""),colnames(data.ks)[2:ncol(data.ks)])){
  dupe <- which(colnames(data.ks)==paste(colnames(data.ks)[1],".1",sep=""))
  data.ks <- data.ks[-dupe]
}

## Fit kitchen sink model ##
mod.ks <- lm(as.formula(paste(eval(colnames(data.ks)[1]), "~ .", sep="")), data=data.ks)

## Gather ridge model data ##
cv <- cv.glmnet(as.matrix(data.ks[c(2:ncol(data.ks))]),as.vector(data.ks[,1]))
data.ridge <- data.ks[which(coef(cv,s="lambda.1se")[2:length(coef(cv,s="lambda.1se"))]!=0) + 1]
data.ridge <- cbind(data.ks[c(1)],data.ridge)

## Fit ridge regression model ##
mod.ridge <- lm(as.formula(paste(eval(colnames(data.ks)[1]), "~ .", sep="")), data=data.ridge)

## Gather reduced model data ##
data.red <-
data.ks[which(as.vector(summary(mod.ks)$coefficients[2:nrow(summary(mod.ks)$coefficients),][,4])<=0.05) + 1]
data.red <- cbind(data.ks[c(1)],data.red)

## Fit reduced model ##
mod.red <- lm(as.formula(paste(eval(colnames(data.ks)[1]), "~ .", sep="")), data=data.red)

## Generate Output ##
out <- tmt(data.ks,mod.ks,mod.red,mod.ridge)
return(out)

#rm(dv,data.in,predictor.cols,data.ks,mod.ks,data.red,mod.red,data.ridge,mod.ridge)
}

##### Generate and save tmt tables for credibility estimates (ind.diffs_wa.wi)
#####
tmt_ane-aos-alpha-wa.wi <- gen.tmt(dv="ane-aos-alpha", data.in=ind.diffs_wa.wi,
predictor.cols=c(8,10:28))
write.csv(tmt_ane-aos-alpha-wa.wi, "tmt_ane-aos-alpha-wa.wi.csv")

tmt_ane-aos-beta-wa.wi <- gen.tmt(dv="ane-aos-beta", data.in=ind.diffs_wa.wi,
predictor.cols=c(8,10:28))
write.csv(tmt_ane-aos-beta-wa.wi, "tmt_ane-aos-beta-wa.wi.csv")

tmt-aos-xi-wa.wi <- gen.tmt(dv="aos-xi", data.in=ind.diffs_wa.wi, predictor.cols=c(8,10:28))
write.csv(tmt-aos-xi-wa.wi, "tmt-aos-xi-wa.wi.csv")

##### Generate and save tmt tables for avg.bs (ind.diffs_wa.wi)
#####
tmt_avg-bs-wa.wi <- gen.tmt(dv="avg.bs", data.in=ind.diffs_wa.wi, predictor.cols=c(2,4,6,10:28))
write.csv(tmt_avg-bs-wa.wi, "tmt_avg-bs-wa.wi.csv")

```

```

#####
##### Linear Hypothesis Testing: Overlap Between Enhanced Environment and Credibility
#
#####
##### Estimate the relevant models for comparison
#####
mod.b <- lm(avg.bs ~ ., data=ind.diffs_wa.wi[c(8,14:28)])      #Baseline model
mod.e <- lm(avg.bs ~ ., data=ind.diffs_wa.wi[c(8,10:28)])      #Environmental vars only
mod.c <- lm(avg.bs ~ ., data=ind.diffs_wa.wi[c(2,4,6,8,14:28)]) #Credibility vars only
mod.f <- lm(avg.bs ~ ., data=ind.diffs_wa.wi[c(2,4,6,8,10:28)]) #Full model

#####
##### Construct a table to compare the various models
#####

## Manually assign ind.diffs_wa.wi to data.all (to minimize departures from tmt code)
data.all <- ind.diffs_wa.wi

## Extract coefficients from each model ##
coeffs.all <- append(colnames(data.all), c("mult.r2","adj.r2","RMSE","AIC","BIC"))
coeffs.all[1] <- "(Intercept)"
coeffs.b <- append(rownames(summary(mod.b)$coefficients), c("mult.r2","adj.r2","RMSE","AIC","BIC"))
coeffs.e <- append(rownames(summary(mod.e)$coefficients), c("mult.r2","adj.r2","RMSE","AIC","BIC"))
coeffs.c <- append(rownames(summary(mod.c)$coefficients), c("mult.r2","adj.r2","RMSE","AIC","BIC"))
coeffs.f <- append(rownames(summary(mod.f)$coefficients), c("mult.r2","adj.r2","RMSE","AIC","BIC"))

## Create spine for output ##
env.vs.cred.table <- data.frame(coeff=coeffs.all)

## Create contents for each table cell ##
items.b <- condense.lm(mod.b)
items.e <- condense.lm(mod.e)
items.c <- condense.lm(mod.c)
items.f <- condense.lm(mod.f)

## Identify which table rows should be full for each model ##
slots.b <- which(is.element(coeffs.all,coeffs.b))
slots.e <- which(is.element(coeffs.all,coeffs.e))
slots.c <- which(is.element(coeffs.all,coeffs.c))
slots.f <- which(is.element(coeffs.all,coeffs.f))

i <- 1
b <- 1
e <- 1
c <- 1
f <- 1
col.b <- vector()
col.e <- vector()
col.c <- vector()
col.f <- vector()

while(i <= nrow(env.vs.cred.table)){

  ## Fill baseline model
  if(is.element(i,slots.b)){
    col.b[i] <- items.b[b]
    b <- b+1
  } else{
    col.b[i] <- "----"
  }

  ## Fill environment model
  if(is.element(i,slots.e)){
    col.e[i] <- items.e[e]
  }
}

```

```

    e <- e+1
} else{
  col.e[i] <- "----"
}

## Fill credibility model
if(is.element(i,slots.c)){
  col.c[i] <- items.c[c]
  c <- c+1
} else{
  col.c[i] <- "----"
}

## Fill full model
if(is.element(i,slots.f)){
  col.f[i] <- items.f[f]
  f <- f+1
} else{
  col.f[i] <- "----"
}

i <- i+1
}

env.vs.cred.table <- cbind(env.vs.cred.table,
                           baseline=col.b,
                           environment.only=col.e,
                           credibility.only=col.c,
                           full.model=col.f)

rm(data.all,coeffs.all,coeffs.b,coeffs.e,coeffs.c,coeffs.f,items.b,items.e,items.c,items.f,slots.b,slots.e,
  slots.c,slots.f,i,b,e,c,f,col.b,col.e,col.c,col.f)

write.csv(env.vs.cred.table, "env.vs.cred.table.csv")

##### Formal hypothesis tests
#####
## Baseline vs. Environment
anova(mod.b,mod.e)

#Res.Df   RSS Df Sum of Sq      F      Pr(>F)
#1     738 528.35
#2     734 454.28  4    74.065 29.917 < 2.2e-16 ***
#  ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Baseline vs. Credibility
anova(mod.b,mod.c)

#Res.Df   RSS Df Sum of Sq      F      Pr(>F)
#1     738 528.35
#2     735 234.84  3    293.51 306.21 < 2.2e-16 ***
#  ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Environment vs. Full
anova(mod.e,mod.f)

#Res.Df   RSS Df Sum of Sq      F      Pr(>F)
#1     734 454.28

```

```
#2      731 227.10  3     227.18 243.75 < 2.2e-16 ***
#  ---
#  Signif. codes:  0 '****' 0.001 '***' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Credibility vs. Full
anova(mod.c,mod.f)

#Res.Df   RSS Df Sum of Sq      F    Pr(>F)
#1      735 234.84
#2      731 227.10  4     7.7338 6.2234 6.304e-05 ***
#  ---
#  Signif. codes:  0 '****' 0.001 '***' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

#####
# Assessing Credibility in Subjective Probability Judgment #
# SPJ Recalibration Scripts, Good Judgment Data #
# (GJP Recalibration) #
# Josh Baker #
# jbak@sas.upenn.edu #
#####

#####
# DATA CLEANING AND PREPARATION
#
#####

##### Set Working directory and load relevant libraries #####
library(effsize)
setwd("~/yourPath")

#####
# Import and Prepare Good Judgment Data for Use in Recalibration #####
fcasts <- read.csv("fcasts_small.file.csv")

## Convert factors to numeric or character data types (and trim last two characters from each element of "ifp_id")
fcasts$ifp_id <- as.numeric(substr(as.character(fcasts$ifp_id),1,4))
fcasts$answer_option <- as.character(fcasts$answer_option)
fcasts$date <- as.character(fcasts$date)
fcasts$timestamp <- as.character(fcasts$timestamp)
fcasts$outcome <- as.character(fcasts$outcome)
fcasts$g.tnt <- as.character(fcasts$g.tnt)
fcasts$mod.tag <- as.character(fcasts$mod.tag)
fcasts$date.closed <- as.character(fcasts$date.closed)
fcasts$time.horiz <- as.character(fcasts$time.horiz)

## Subset fcasts to answer_option=="a" (only forecasts for focal outcomes, not complements)
fcasts <- subset(fcasts, answer_option=="a")

## Set aside fcasts metadata that are irrelevant to recalibration analyses
fcasts_w.metadata <- fcasts
fcasts <- fcasts[which(is.element(colnames(fcasts),c("ifp_id",
                                                 "user_id",
                                                 "outcome",
                                                 "val.unrounded",
                                                 "bs.unrounded")))]]

## Rename some columns for consistency/expediency
colnames(fcasts)[which(colnames(fcasts)=="ifp_id")] <- "ifp.id"
colnames(fcasts)[which(colnames(fcasts)=="user_id")] <- "gjp.id"
colnames(fcasts)[which(colnames(fcasts)=="val.unrounded")] <- "prob"
colnames(fcasts)[which(colnames(fcasts)=="bs.unrounded")] <- "bs"

## Recode "outcome" to binary ("a" = 1; "b" = 0)
fcasts$outcome <- ifelse(fcasts$outcome=="a",1,0)

```

```

## Round all prob values to two places to fix a floating point rounding issue
fcasts$prob <- round(fcasts$prob,2)

## Extract GJP IDs as a vector
gjp.ids <- unique(fcasts$gjp.id)
gjp.ids <- gjp.ids[order(gjp.ids)] #Sort in ascending order

#####
##### Elementary Functions for credibility estimation and results calculation
#####

## Function for calculating the number of unique elements in a vector #####
n.unique <- function(vector){length(unique(vector))}

## Create a data.frame for individual.difference data (here, only used to track n.ifps) #####
ind.diffs <- data.frame(gjp.id=gjp.ids)
fcasts <- fcasts[order(fcasts$gjp.id),]
ind.diffs <- cbind(ind.diffs, n.ifps=as.vector(tapply(fcasts$ifp.id,fcasts$gjp.id,n.unique)))

### BS function for optimization ###
bs.opt <- function(x,outcome){sum((temp.outcomes - ptrans(temp.probs,x))^2)}

### BS function for scoring ###
calc.bs <- function(x,outcome){(x-outcome)^2}

## Function for converting a probability to log-odds
calc.logodds <- function(x){log(x/(1 - x))}

## Extremizing function used to correct for "Regression" towards 50% (see: Baron et al., 2014)
ptrans <- function(p,a){p^a/(p^a + (1 - p)^a)}

## General function for calculating SSE
calc.sse <- function(obj,est){sum((obj-est)^2)}

## Function for calculating reliability (one component of Brier Score decomposition)
calc.rel <- function(probs,outcomes){
  probs <- round(probs,2)                                     #Round probability values to fix
  issues with floating point representation

  rel.data <- data.frame(prob=probs,outcome=outcomes)        #Arguments structured as data.frame
  for use with "by()"
  rel.data <- rel.data[order(rel.data$prob),]                  #rel.data, sorted in ascending prob
  order

  n <- as.numeric(nrow(rel.data))                            #Total number of forecasts
  fk <- unique(rel.data$prob)                                #Vector of unique forecast values
  nk <- as.vector(by(rel.data$prob, rel.data$prob, length)) #Number of forecasts associated
  with each "bin" (unique forecast value)
  obark <- as.vector(by(rel.data$outcome, rel.data$prob, mean)) #Baserate associated with each
  "bin" (unique forecast value)

  rel <- (1/n) * sum((nk * ((fk - obark)^2)))  #Reliability = mean weighted SSE between fcast values
  and baserates across "bins"

  return(rel)
}

#rm(probs,outcomes,probs.1,probs.0,out.1,out.0,rel.data,n,fk,nk,obark,rel)
}

```

```

## Function for calculating resolution (one component of Brier Score decomposition)
calc.res <- function(probs,outcomes){
  res.data <- data.frame(prob=probs,outcome=outcomes) #Arguments structured as data.frame
  for use with "by()"
    res.data <- res.data[order(res.data$prob),] #res.data, sorted in ascending prob
  order

  n <- nrow(res.data) #Total number of forecasts
  nk <- as.vector(by(res.data$prob, res.data$prob, length)) #Number of forecasts associated
  with each "bin" (unique forecast value)
  obar <- mean(res.data$outcome) #Overall baserate
  obark <- as.vector(by(res.data$outcome, res.data$prob, mean)) #Baserate associated with each
  "bin" (unique forecast value)

  res <- (1/n) * sum((nk * ((obark - obar)^2))) #Resolution = mean weighted SSE between "bin"
  baserates and overall baserates across "bins"

  return(res)
}

#rm(probs,outcomes,probs.1,probs.0,out.1,out.0,res.data,n,nk,obar,obark,res)
}

## Function for calculating uncertainty (one component of Brier Score decomposition)
calc.unc <- function(outcomes){
  obar <- mean(outcomes) #Overall baserate for event occurrence

  unc <- obar*(1-obar)

  return(unc)
}

#rm(outcomes,obar)
}

#####
# Function for generating optimized, aggregate judgments, given a sample size criterion
#####
gen.agg_gjp <- function(sample.size, method=c("mean","median"), log.odds=TRUE){

  ## Subset to only those forecasters who addressed a sufficient number of IFPs
  keepers <- gjp.ids[which(ind.diffs$n.ifps >= (sample.size + 30))] #Identify forecasters
  with enough data to recalibrate 30+ out of sample judgments
  working.data <- subset(fcasts, is.element(fcasts$gjp.id, keepers)) #Subset to only relevant
  forecasters
  working.data$prob <- ifelse(working.data$prob==0,0.01,working.data$prob) #Adjust extreme values
  to prevent (-)Inf, once transformed to log-odds
  working.data$prob <- ifelse(working.data$prob==1,0.99,working.data$prob) #Adjust extreme values
  to prevent (-)Inf, once transformed to log-odds
  working.data <- working.data #Save working data to
  .GlobalEnv
  temp.outcomes <- as.vector(tapply(working.data$outcome, working.data$ifp.id, mean)) #Save
  relevant outcomes to .GlobalEnv

  ## Aggregate judgments and save to .GolbalEnv
  if(method=="mean"){
    temp.probs <- as.vector(tapply(working.data$prob, working.data$ifp.id, mean))
  }
  if(method=="median"){
    temp.probs <- as.vector(tapply(working.data$prob, working.data$ifp.id, median))
  }

  ## Calculate optimized extremizing coefficient
  a <- as.numeric(optimize(bs.opt, interval=c(0,20)))[1]

  ## Apply extremizing coefficient to simple aggrgeates to calculate optimized aggregate estimates
  agg.judge <- ptrans(temp.probs,a)
}

```

```

## If desired, transform agg.judge to log-odds
if(log.odds){
  agg.judge <- calc.logodds(agg.judge)
}

return(agg.judge)

#rm(sample.size,method,log.odds,keepers,working.data,temp.outcomes,temp.probs,a,agg.judge)
}
#####
##### Function for credibility-based recalibration and out of sample prediction using GJP Data
#####
oosp_gjp <- function(user.id, sample.size, method=c("mean","median"), log.odds=TRUE, n.resample){

  ## Load "effsize" package
  require(effsize)

  ## Create an empty data.frame for results
  data.out <- data.frame(gjp.id=0,
                         cal.sample.size=0,
                         sample
                         (recalibrated) sample
                         (consistency)
                         before recalibration
                         after recalibration
                         optimized aggregates
                         reduced AJE
                         (pre-post; log-odds)
                         (pre-post; prob)
                         (pre-post; log-odds)
                         (pre-post; prob)
                         (log-odds)
                         (prob)
                         recalibration (log-odds)
                         recalibration (log-odds)
                         recalibration (prob))

  pred.sample.size=0,          #Number of observations in each CF calibration
  est.alpha=0,                 #Estimated CF intercept (bias)
  est.beta=0,                  #Estimated CF slope (expertise)
  est.ser=0,                   #Standard error of the CF regression = "xi"
  rmse.pre_lo=0,               #Root mean squared error (judgements vs. opt.agg),
  rmse.post_lo=0,              #Root mean squared error (judgements vs. opt.agg),
  rmse.impr=0,                 #Did recalibration reduce RMSE?
  ## AJE = Absolute Judgment Error = abs. diff. between judgments and
  prop.aje.impr=0,             #Proportion of judgments for which recalibration
  mean.aje.pre_lo=0,            #Mean AJE before recalibration (log-odds)
  mean.aje.post_lo=0,           #Mean AJE after recalibration (log-odds)
  diff.mean.aje_lo=0,           #Change in mean AJE as a result of recalibration
  mean.aje.pre_p=0,             #Mean AJE before recalibration (prob)
  mean.aje.post_p=0,             #Mean AJE after recalibration (prob)
  diff.mean.aje_p=0,             #Change in mean AJE as a result of recalibration
  mean.aje.impr=0,              #Did recalibration reduce mean AJE?
  median.aje.pre_lo=0,           #Median AJE before recalibration (log-odds)
  median.aje.post_lo=0,          #Median AJE after recalibration (log-odds)
  diff.median.aje_lo=0,          #Change in median AJE as a result of recalibration
  median.aje.pre_p=0,             #Median AJE before recalibration (prob)
  median.aje.post_p=0,             #Median AJE after recalibration (prob)
  diff.median.aje_p=0,             #Change in median AJE as a result of recalibration
  median.aje.impr=0,              #Did recalibration reduce median AJE?
  coh.d_aje.pre.post_lo=0,        #Effect size (cohen's D) of recalibration on AJE
  coh.d_aje.pre.post_p=0,         #Effect size (Cohen's D) of recalibration on AJE
  mean.diff.aje_lo=0,              #Mean pairwise change in AJE as a result of
  median.diff.aje_lo=0,             #Median pairwise change in AJE as a result of
  mean.diff.aje_p=0,                #Mean pairwise change in AJE as a result of
}

```

```

    median.diff.aje_p=0,      #Mean pairwise change in AJE as a result of
recalibration (prob)

    ## ALE = Absolute Linear Error = abs. diff between judgments and observed
outcomes
    prop.ale.impr=0,          #All measures in this section as similar to those
above, but for ALE (vs. AJE)
    mean.ale.pre=0,
    mean.ale.post=0,
    diff.mean.ale=0,
    mean.ale.impr=0,
    median.ale.pre=0,
    median.ale.post=0,
    diff.median.ale=0,
    median.ale.impr=0,
    coh.d_ale.pre.post=0,
    mean.diff.ale=0,
    median.diff.ale=0,

    ## BS = Brier Score = squared difference between judgments and observed
outcomes
    prop.bs.impr=0,          #All measures in this section as similar to those
above, but for BS
    mean.bs.pre=0,
    mean.bs.post=0,
    diff.mean.bs=0,
    mean.bs.impr=0,
    median.bs.pre=0,
    median.bs.post=0,
    diff.median.bs=0,
    median.bs.impr=0,
    coh.d_bs.pre.post=0,
    mean.diff.bs=0,
    median.diff.bs=0,

    ## Brier score decomposition: REL = reliability; RES = resolution
    rel.pre=0,
    rel.post=0,
    diff.rel=0,
    rel.impr=0,
    res.pre=0,
    res.post=0,
    diff.res=0,
    res.impr=0,
    uncertainty=0)

## Generate optimized aggregates
agg <- gen.agg_gjp(sample.size,method,log.odds)
agg_ifp.ids <- as.vector(tapply(working.data$ifp.id, working.data$ifp.id, mean)) #Identify IFPs
for which there exists an optimized aggregate

## Prepare individual forecaster's data
user.data <- subset(working.data, gjp.id==user.id)                      #Subset to individual ("user")
user.data <- user.data[order(user.data$ifp.id),]                            #Sort user.obs by ifp.id
user.data$prob <- ifelse(user.data$prob==0,0.01,user.data$prob)            #Adjust extreme values to prevent
(-)Inf, once transformed to log-odds
user.data$prob <- ifelse(user.data$prob==1,0.99,user.data$prob)            #Adjust extreme values to prevent
(-)Inf, once transformed to log-odds
user.ifps <- unique(user.data$ifp.id)                                       #Generate list of ifps that user
responded to

agg_user.matched.sample <- agg[which(is.element(agg_ifp.ids,user.ifps))]
#Select matching optimized aggregates
outcomes_user.matched.sample <- as.vector(tapply(user.data$outcome, user.data$ifp.id, mean))
#select matching outcomes

## Calculate mean or median individual judgments for each ifp
if(method=="mean"){

```

```

    user.obs <- as.vector(tapply(user.data$prob, user.data$ifp.id, mean))
    user.obs <- calc.logodds(user.obs)
}
if(method=="median"){
    user.obs <- as.vector(tapply(user.data$prob, user.data$ifp.id, median))
    user.obs <- calc.logodds(user.obs)
}

## Fit credibility function, and conduct recalibration n.resample times for each individual
i <- 1           #Resample index
flag <- FALSE   #Skip indicator for unusable loop iterations

while(i <= n.resample){
    # set/reset flag
    flag <- FALSE

    # Split data into calibration sample and prediction sample
    cal.sample.indices <- sample.int(length(user.obs), sample.size, replace=FALSE)
    pred.sample.indices <- setdiff(c(1:length(user.obs)), cal.sample.indices)

    user.obs_cal.sample <- user.obs[cal.sample.indices]
    user.obs_pred.sample <- user.obs[pred.sample.indices]

    agg_cal.sample <- agg_user.matched.sample[cal.sample.indices]
    agg_pred.sample <- agg_user.matched.sample[pred.sample.indices]

    outcomes_cal.sample <- outcomes_user.matched.sample[cal.sample.indices]
    outcomes_pred.sample <- outcomes_user.matched.sample[pred.sample.indices]

    # Estimate credibility function (re-fit on a new calibration sample each iteration)
    cf <- lm(agg_cal.sample ~ user.obs_cal.sample)

    # Check for inestimable credibility functions, and re-set loop if necessary
    if(is.na(cf$coeff[1]) | is.na(cf$coeff[2])){
        flag <- TRUE
        i <- i-1
    }

    # If CF is usable, calculate effects of recalibration
    if(!(flag)){

        #Apply CF to prediction sample
        corrected <- (user.obs_pred.sample * cf$coeff[2]) + cf$coeff[1]

        #Calculate RMSE outcomes
        rmse.pre_lo <-
        sqrt((calc.sse(agg_pred.sample,user.obs_pred.sample)/length(user.obs_pred.sample)))
        rmse.post_lo <- sqrt((calc.sse(agg_pred.sample,corrected)/length(corrected)))
        rmse.impr <- ifelse(rmse.post_lo < rmse.pre_lo, 1, 0)

        #Calculate AJE outcomes
        aje.pre_lo <- abs(agg_pred.sample - user.obs_pred.sample)                      #Abs. judgment
        errors, before recalibration (log-odds)
        aje.post_lo <- abs(agg_pred.sample - corrected)                                #Abs. judgment
        errors, after recalibration (log-odds)
        aje.pre_p <- abs(plogis(agg_pred.sample) - plogis(user.obs_pred.sample)) #Abs. judgment
        errors, before recalibration (prob)
        aje.post_p <- abs(plogis(agg_pred.sample) - plogis(corrected))      #Abs. judgment
        errors, after recalibration (prob)
        diff.aje_lo <- aje.pre_lo - aje.post_lo                                     #Changes in AJE
        (log-odds)
        diff.aje_p <- aje.pre_p - aje.post_p                                       #Changes in AJE
        (prob)

        prop.aje.impr <- mean(ifelse(diff.aje_lo > 0,1,0))                      #Proportion of judgments
        for which recalibration reduced AJE
    }
}

```

```

    mean.aje.pre_lo <- mean(aje.pre_lo)                                #Mean AJE before
recalibration (log-odds)
    mean.aje.post_lo <- mean(aje.post_lo)                               #Mean AJE after
recalibration (log-odds)
    diff.mean.aje_lo <- mean.aje.pre_lo - mean.aje.post_lo            #Change in mean AJE as a
result of recalibration (pre-post; log-odds)
    mean.aje.pre_p <- mean(aje.pre_p)                                   #Mean AJE before
recalibration (prob)
    mean.aje.post_p <- mean(aje.post_p)                                 #Mean AJE after
recalibration (prob)
    diff.mean.aje_p <- mean.aje.pre_p - mean.aje.post_p               #Change in mean AJE as a
result of recalibration (pre-post; prob)
    mean.aje.impr <- ifelse(diff.mean.aje_lo > 0,1,0)                #Did recalibration reduce
mean AJE?
    median.aje.pre_lo <- median(aje.pre_lo)                            #Median AJE before
recalibration (log-odds)
    median.aje.post_lo <- median(aje.post_lo)                           #Median AJE after
recalibration (log-odds)
    diff.median.aje_lo <- median.aje.pre_lo - median.aje.post_lo      #Change in median AJE as
a result of recalibration (pre-post; log-odds)
    median.aje.pre_p <- median(aje.pre_p)                             #Median AJE before
recalibration (prob)
    median.aje.post_p <- median(aje.post_p)                            #Median AJE after
recalibration (prob)
    diff.median.aje_p <- median.aje.pre_p - median.aje.post_p         #Change in median AJE as
a result of recalibration (pre-post; prob)
    median.aje.impr <- ifelse(diff.median.aje_lo > 0,1,0)           #Did recalibration reduce
median AJE?
    coh.d_aje.pre.post_lo <- cohen.d(aje.pre_lo, aje.post_lo, paired=T) #Effect size (cohen's D)
of recalibration on AJE (log-odds)
    coh.d_aje.pre.post_lo <- as.numeric(coh.d_aje.pre.post_lo$estimate) #Extract estimate of D
only
    coh.d_aje.pre.post_p <- cohen.d(aje.pre_p, aje.post_p, paired=T)   #Effect size (Cohen's D)
of recalibration on AJE (prob)
    coh.d_aje.pre.post_p <- as.numeric(coh.d_aje.pre.post_p$estimate) #Extract estimate of D
only
    mean.diff.aje_lo <- mean(diffs.aje_lo)                             #Mean pairwise change in
AJE as a result of recalibration (log-odds)
    median.diff.aje_lo <- median(diffs.aje_lo)                          #Median pairwise change
in AJE as a result of recalibration (log-odds)
    mean.diff.aje_p <- mean(diffs.aje_p)                                #Mean pairwise change in
AJE as a result of recalibration (prob)
    median.diff.aje_p <- median(diffs.aje_p)                            #Mean pairwise change in
AJE as a result of recalibration (prob)

#Calculate ALE outcomes
ale.pre <- abs(outcomes_pred.sample - plogis(user.obs_pred.sample)) #Abs. linear errors,
before recalibration (prob)
ale.post <- abs(outcomes_pred.sample - plogis(corrected))          #Abs. linear errors,
after recalibration (prob)
diffs.ale <- ale.pre - ale.post                                      #Changes in ALE (prob)

prop.ale.impr <- mean(ifelse(diffs.ale > 0,1,0))                 #All measures in this
section as similar to those above, but for ALE (vs. AJE)
mean.ale.pre <- mean(ale.pre)
mean.ale.post <- mean(ale.post)
diff.mean.ale <- mean.ale.pre - mean.ale.post
mean.ale.impr <- ifelse(diff.mean.ale > 0,1,0)
median.ale.pre <- median(ale.pre)
median.ale.post <- median(ale.post)
diff.median.ale <- median.ale.pre - median.ale.post
median.ale.impr <- ifelse(diff.median.ale > 0,1,0)
coh.d_ale.pre.post <- cohen.d(ale.pre, ale.post, paired=T)
coh.d_ale.pre.post <- as.numeric(coh.d_ale.pre.post$estimate) #Extract estimate of D only
mean.diff.ale <- mean(diffs.ale)
median.diff.ale <- median(diffs.ale)

```

```

#Calculate BS stats
bs.pre <- calc.bs(plogis(user.obs_pred.sample), outcomes_pred.sample)
bs.post <- calc.bs(plogis(corrected), outcomes_pred.sample)
diffs.bs <- bs.pre - bs.post

prop.bs.impr <- mean(ifelse(diffs.bs > 0,1,0))                                #All measures in this
section as similar to those above, but for BS
mean.bs.pre <- mean(bs.pre)
mean.bs.post <- mean(bs.post)
diff.mean.bs <- mean.bs.pre - mean.bs.post
mean.bs.impr <- ifelse(diff.mean.bs > 0,1,0)
median.bs.pre <- median(bs.pre)
median.bs.post <- median(bs.post)
diff.median.bs <- median.bs.pre - median.bs.post
median.bs.impr <- ifelse(diff.median.bs > 0,1,0)
coh.d_bs.pre.post <- cohen.d(bs.pre, bs.post, paired=T)
coh.d_bs.pre.post <- as.numeric(coh.d_bs.pre.post$estimate) #Extract estimate of D only
mean.diff.bs <- mean(diffs.bs)
median.diff.bs <- median(diffs.bs)

#Calculate Brier Score Decomposition Stats
rel.pre <- calc.rel(plogis(user.obs_pred.sample), outcomes_pred.sample)
rel.post <- calc.rel(plogis(corrected), outcomes_pred.sample)
diff.rel <- rel.pre - rel.post
rel.impr <- ifelse(diff.rel > 0,1,0)      #Smaller reliabilities indicate better performance
res.pre <- calc.res(plogis(user.obs_pred.sample), outcomes_pred.sample)
res.post <- calc.res(plogis(corrected), outcomes_pred.sample)
diff.res <- res.pre - res.post
res.impr <- ifelse(diff.res < 0,1,0)      #Larger reliabilities indicate better performance
uncertainty <- calc.unc(outcomes_pred.sample)

## Compile output into a new row and add to data.out
newrow <- data.frame(gjp.id=user.id,
                      cal.sample.size=sample.size,
                      pred.sample.size=length(user.obs_pred.sample),
                      est.alpha=cf$coeff[1],
                      est.beta=cf$coeff[2],
                      est.ser=summary(cf)$sigma,
                      rmse.pre_lo=rmse.pre_lo,
                      rmse.post_lo=rmse.post_lo,
                      rmse.impr=rmse.impr,
                      prop.aje.impr=prop.aje.impr,
                      mean.aje.pre_lo=mean.aje.pre_lo,
                      mean.aje.post_lo=mean.aje.post_lo,
                      diff.mean.aje_lo=diff.mean.aje_lo,
                      mean.aje.pre_p=mean.aje.pre_p,
                      mean.aje.post_p=mean.aje.post_p,
                      diff.mean.aje_p=diff.mean.aje_p,
                      mean.aje.impr=mean.aje.impr,
                      median.aje.pre_lo=median.aje.pre_lo,
                      median.aje.post_lo=median.aje.post_lo,
                      diff.median.aje_lo=diff.median.aje_lo,
                      median.aje.pre_p=median.aje.pre_p,
                      median.aje.post_p=median.aje.post_p,
                      diff.median.aje_p=diff.median.aje_p,
                      median.aje.impr=median.aje.impr,
                      coh.d_aje.pre.post_lo=coh.d_aje.pre.post_lo,
                      coh.d_aje.pre.post_p=coh.d_aje.pre.post_p,
                      mean.diff.aje_lo=mean.diff.aje_lo,
                      median.diff.aje_lo=median.diff.aje_lo,
                      mean.diff.aje_p=mean.diff.aje_p,
                      median.diff.aje_p=median.diff.aje_p,
                      prop.ale.impr=prop.ale.impr,
                      mean.ale.pre=mean.ale.pre,
                      mean.ale.post=mean.ale.post,
                      diff.mean.ale=diff.mean.ale,
                      mean.ale.impr=mean.ale.impr,

```

```

    median.ale.pre=median.ale.pre,
    median.ale.post=median.ale.post,
    diff.median.ale=diff.median.ale,
    median.ale.impr=median.ale.impr,
    coh.d_ale.pre.post=coh.d_ale.pre.post,
    mean.diff.ale=mean.diff.ale,
    median.diff.ale=median.diff.ale,
    prop.bs.impr=prop.bs.impr,
    mean.bs.pre=mean.bs.pre,
    mean.bs.post=mean.bs.post,
    diff.mean.bs=diff.mean.bs,
    mean.bs.impr=mean.bs.impr,
    median.bs.pre=median.bs.pre,
    median.bs.post=median.bs.post,
    diff.median.bs=diff.median.bs,
    median.bs.impr=median.bs.impr,
    coh.d_bs.pre.post=coh.d_bs.pre.post,
    mean.diff.bs=mean.diff.bs,
    median.diff.bs=median.diff.bs,
    rel.pre=rel.pre,
    rel.post=rel.post,
    diff.rel=diff.rel,
    rel.impr=rel.impr,
    res.pre=res.pre,
    res.post=res.post,
    diff.res=diff.res,
    res.impr=res.impr,
    uncertainty=uncertainty)

    data.out <- rbind(data.out,newrow)
}

i <- i+1
}

## Clean-up and Return Results
data.out <- data.out[-1,]          #Remove dummy first row of data.out
rownames(data.out) <- seq.int(1:nrow(data.out)) #As a general precaution, reset rownames of
data.out

return(data.out)

#rm(user.id,sample.size,method,log.odds,n.reample,agg,agg_ifp.ids,user.data,user.ifps,agg_user.match
ed.sample,
#outcomes_user.matched.sample,user.obs,i,flag,cal.sample.indices,pred.sample.indices,user.obs_cal.sa
mple,
#user.obs_pred.sample,agg_cal.sample,agg_pred.sample,outcomes_cal.sample,outcomes_pred.sample,cf,cor
rected,
#data.out,rmse.pre_lo,rmse.post_lo,rmse.impr,aje.pre_lo,aje.post_lo,aje.pre_p,aje.post_p,diffs.aje_l
o,
#diffs.aje_p,prop.aje.impr,mean.aje.pre_lo,mean.aje.post_lo,diff.mean.aje_lo,mean.aje.pre_p,mean.aje
.post_p,
#diff.mean.aje_p,mean.aje.impr,median.aje.pre_lo,median.aje.post_lo,diff.median.aje_lo,median.aje.pr
e_p,
#median.aje.post_p,diff.median.aje_p,median.aje.impr,coh.d_aje.pre.post_lo,coh.d_aje.pre.post_p,mean
.diff.aje_lo,
#median.diff.aje_lo,mean.diff.aje_p,median.diff.aje_p,ale.pre,ale.post,diffs.ale,prop.ale.impr,mean.
.ale.pre,
#mean.ale.post,diff.mean.ale,mean.ale.impr,median.ale.pre,median.ale.post,diff.median.ale,median.ale

```

```

.impr,
#coh.d_ale.pre.post,mean.diff.ale,median.diff.ale,bs.pre,bs.post,diffs.bs,prop.bs.impr,mean.bs.pre,
mean.bs.post,
#diff.mean.bs,mean.bs.impr,median.bs.pre,median.bs.post,diff.median.bs,median.bs.impr,coh.d_bs.pre.p
ost,
#mean.diff.bs,median.diff.bs,rel.pre,rel.post,diff.rel,rel.impr,res.pre,res.post,diff.res,res.impr,u
ncertainty)
}
#####
#####
#####

##### Function to run oosp_gjp for all eligible forecasters, given a desired sample.size
#####
run.oosp_gjp <- function(sample.size, method=c("mean","median"), log.odds=TRUE, n.resample){

  data.out <- data.frame(gjp.id=0,                                     # AOS = average over samples; MOS = median
over samples
                                cal.sample.size=0,                      # Both refer to summary stats over each
forecaster's 100 resamples
                                pred.sample.size=0,
                                aos.alpha=0,
                                mos.alpha=0,
                                sd.alpha=0,
                                aos.beta=0,
                                mos.beta=0,
                                sd.beta=0,
                                aos.xi=0,
                                mos.xi=0,
                                sd.xi=0,
                                prop.samp.rmse.impr=0,
                                aos.prop.aje.impr=0,
                                aos.mean.aje.pre_lo=0,
                                aos.mean.aje.post_lo=0,
                                aos.diff.mean.aje_lo=0,
                                aos.mean.aje.pre_p=0,
                                aos.mean.aje.post_p=0,
                                aos.diff.mean.aje_p=0,
                                prop.samp.mean.aje.impr=0,
                                aos.median.aje.pre_lo=0,
                                aos.median.aje.post_lo=0,
                                aos.diff.median.aje_lo=0,
                                aos.median.aje.pre_p=0,
                                aos.median.aje.post_p=0,
                                aos.diff.median.aje_p=0,
                                prop.samp.median.aje.impr=0,
                                aos.coh.d_aje.pre.post_lo=0,
                                aos.coh.d_aje.pre.post_p=0,
                                aos.mean.diff.aje_lo=0,
                                aos.median.diff.aje_lo=0,
                                aos.mean.diff.aje_p=0,
                                aos.median.diff.aje_p=0,
                                aos.prop.aje.impr=0,
                                aos.mean.aje.pre=0,
                                aos.mean.aje.post=0,
                                aos.diff.mean.aje=0,
                                prop.samp.mean.aje.impr=0,
                                aos.median.aje.pre=0,
                                aos.median.aje.post=0,
                                aos.diff.median.aje=0,
                                prop.samp.median.aje.impr=0,
                                aos.coh.d_ale.pre.post=0,
                                aos.mean.diff.ale=0,
                                aos.median.diff.ale=0,
                                aos.prop.bs.impr=0,

```

```

aos.mean.bs.pre=0,
aos.mean.bs.post=0,
aos.diff.mean.bs=0,
prop.samp.mean.bs.impr=0,
aos.median.bs.pre=0,
aos.median.bs.post=0,
aos.diff.median.bs=0,
prop.samp.median.bs.impr=0,
aos.coh.d_bs.pre.post=0,
aos.mean.diff.bs=0,
aos.median.diff.bs=0,
aos.rel.pre=0,
aos.rel.post=0,
aos.diff.rel=0,
prop.samp.rel.impr=0,
aos.res.pre=0,
aos.res.post=0,
aos.diff.res=0,
prop.samp.res.impr=0,
aos.uncertainty=0,
mos.prop.aje.impr=0,
mos.mean.aje.pre_lo=0,
mos.mean.aje.post_lo=0,
mos.diff.mean.aje_lo=0,
mos.mean.aje.pre_p=0,
mos.mean.aje.post_p=0,
mos.diff.mean.aje_p=0,
mos.median.aje.pre_lo=0,
mos.median.aje.post_lo=0,
mos.diff.median.aje_lo=0,
mos.median.aje.pre_p=0,
mos.median.aje.post_p=0,
mos.diff.median.aje_p=0,
mos.coh.d_aje.pre.post_lo=0,
mos.coh.d_aje.pre.post_p=0,
mos.mean.diff.aje_lo=0,
mos.median.diff.aje_lo=0,
mos.mean.diff.aje_p=0,
mos.median.diff.aje_p=0,
mos.prop.aje.impr=0,
mos.mean.ale.pre=0,
mos.mean.ale.post=0,
mos.diff.mean.ale=0,
mos.median.ale.pre=0,
mos.median.ale.post=0,
mos.diff.median.ale=0,
mos.coh.d_ale.pre.post=0,
mos.mean.diff.ale=0,
mos.median.diff.ale=0,
mos.prop.bs.impr=0,
mos.mean.bs.pre=0,
mos.mean.bs.post=0,
mos.diff.mean.bs=0,
mos.median.bs.pre=0,
mos.median.bs.post=0,
mos.diff.median.bs=0,
mos.coh.d_bs.pre.post=0,
mos.mean.diff.bs=0,
mos.median.diff.bs=0,
mos.rel.pre=0,
mos.rel.post=0,
mos.diff.rel=0,
mos.res.pre=0,
mos.res.post=0,
mos.diff.res=0,
mos.uncertainty=0)

```

## Subset to only those forecasters with enough data to recalibrate 30+ out of sample predictions

```

keepers <- gjp.ids[which(ind.diffs$n.ifps >= (sample.size + 30))]

## Run oosp_gjp for each forecaster in keepers
n <- 1

while(n <= length(keepers)){
  temp <- oosp_gjp(user.id=keepers[n],sample.size,method,log.odds,n.resample)

  ## Compile data in a new row and add to data.out
  newrow <- data.frame(gjp.id=temp$gjp.id[1],
                        cal.sample.size=temp$cal.sample.size[1],
                        pred.sample.size=temp$pred.sample.size[1],
                        aos.alpha=mean(temp$est.alpha),
                        mos.alpha=median(temp$est.alpha),
                        sd.alpha=sd(temp$est.alpha),
                        aos.beta=mean(temp$est.beta),
                        mos.beta=median(temp$est.beta),
                        sd.beta=sd(temp$est.beta),
                        aos.xi=mean(temp$est.ser),
                        mos.xi=median(temp$est.ser),
                        sd.xi=sd(temp$est.ser),
                        prop.samp.rmse.impr=mean(temp$rmse.impr),
                        aos.prop.aje.impr=mean(temp$prop.aje.impr),
                        aos.mean.aje.pre_lo=mean(temp$mean.aje.pre_lo),
                        aos.mean.aje.post_lo=mean(temp$mean.aje.post_lo),
                        aos.diff.mean.aje_lo=mean(temp$diff.mean.aje_lo),
                        aos.mean.aje.pre_p=mean(temp$mean.aje.pre_p),
                        aos.mean.aje.post_p=mean(temp$mean.aje.post_p),
                        aos.diff.mean.aje_p=mean(temp$diff.mean.aje_p),
                        prop.samp.mean.aje.impr=mean(temp$mean.aje.impr),
                        aos.median.aje.pre_lo=mean(temp$median.aje.pre_lo),
                        aos.median.aje.post_lo=mean(temp$median.aje.post_lo),
                        aos.diff.median.aje_lo=mean(temp$diff.median.aje_lo),
                        aos.median.aje.pre_p=mean(temp$median.aje.pre_p),
                        aos.median.aje.post_p=mean(temp$median.aje.post_p),
                        aos.diff.median.aje_p=mean(temp$diff.median.aje_p),
                        prop.samp.median.aje.impr=mean(temp$median.aje.impr),
                        aos.coh.d_aje.pre.post_lo=mean(temp$coh.d_aje.pre.post_lo),
                        aos.coh.d_aje.pre.post_p=mean(temp$coh.d_aje.pre.post_p),
                        aos.mean.diff.aje_lo=mean(temp$mean.diff.aje_lo),
                        aos.median.diff.aje_lo=mean(temp$median.diff.aje_lo),
                        aos.mean.diff.aje_p=mean(temp$mean.diff.aje_p),
                        aos.median.diff.aje_p=mean(temp$median.diff.aje_p),
                        aos.prop.aje.impr=mean(temp$prop.aje.impr),
                        aos.mean.ale.pre=mean(temp$mean.ale.pre),
                        aos.mean.ale.post=mean(temp$mean.ale.post),
                        aos.diff.mean.ale=mean(temp$diff.mean.ale),
                        prop.samp.mean.ale.impr=mean(temp$mean.ale.impr),
                        aos.median.ale.pre=mean(temp$median.ale.pre),
                        aos.median.ale.post=mean(temp$median.ale.post),
                        aos.diff.median.ale=mean(temp$diff.median.ale),
                        prop.samp.median.ale.impr=mean(temp$median.ale.impr),
                        aos.coh.d_ale.pre.post=mean(temp$coh.d_ale.pre.post),
                        aos.mean.diff.ale=mean(temp$mean.diff.ale),
                        aos.median.diff.ale=mean(temp$median.diff.ale),
                        aos.prop.bs.impr=mean(temp$prop.bs.impr),
                        aos.mean.bs.pre=mean(temp$mean.bs.pre),
                        aos.mean.bs.post=mean(temp$mean.bs.post),
                        aos.diff.mean.bs=mean(temp$diff.mean.bs),
                        prop.samp.mean.bs.impr=mean(temp$mean.bs.impr),
                        aos.median.bs.pre=mean(temp$median.bs.pre),
                        aos.median.bs.post=mean(temp$median.bs.post),
                        aos.diff.median.bs=mean(temp$diff.median.bs),
                        prop.samp.median.bs.impr=mean(temp$median.bs.impr),
                        aos.coh.d_bs.pre.post=mean(temp$coh.d_bs.pre.post),
                        aos.mean.diff.bs=mean(temp$mean.diff.bs),
                        aos.median.diff.bs=mean(temp$median.diff.bs),
                        aos.rel.pre=mean(temp$rel.pre),
                        aos.rel.post=mean(temp$rel.post),

```

```

aos.diff.rel=mean(temp$diff.rel),
prop.samp.rel.impr=mean(temp$rel.impr),
aos.res.pre=mean(temp$res.pre),
aos.res.post=mean(temp$res.post),
aos.diff.res=mean(temp$diff.res),
prop.samp.res.impr=mean(temp$res.impr),
aos.uncertainty=mean(temp$uncertainty),
mos.prop.aje.impr=median(temp$prop.aje.impr),
mos.mean.aje.pre_lo=median(temp$mean.aje.pre_lo),
mos.mean.aje.post_lo=median(temp$mean.aje.post_lo),
mos.diff.mean.aje_lo=median(temp$diff.mean.aje_lo),
mos.mean.aje.pre_p=median(temp$mean.aje.pre_p),
mos.mean.aje.post_p=median(temp$mean.aje.post_p),
mos.diff.mean.aje_p=median(temp$diff.mean.aje_p),
mos.median.aje.pre_lo=median(temp$median.aje.pre_lo),
mos.median.aje.post_lo=median(temp$median.aje.post_lo),
mos.diff.median.aje_lo=median(temp$diff.median.aje_lo),
mos.median.aje.pre_p=median(temp$median.aje.pre_p),
mos.median.aje.post_p=median(temp$median.aje.post_p),
mos.diff.median.aje_p=median(temp$diff.median.aje_p),
mos.coh.d_aje.pre.post_lo=median(temp$coh.d_aje.pre.post_lo),
mos.coh.d_aje.pre.post_p=median(temp$coh.d_aje.pre.post_p),
mos.mean.diff.aje_lo=median(temp$mean.diff.aje_lo),
mos.median.diff.aje_lo=median(temp$median.diff.aje_lo),
mos.mean.diff.aje_p=median(temp$mean.diff.aje_p),
mos.median.diff.aje_p=median(temp$median.diff.aje_p),
mos.prop.ale.impr=median(temp$prop.ale.impr),
mos.mean.ale.pre=median(temp$mean.ale.pre),
mos.mean.ale.post=median(temp$mean.ale.post),
mos.diff.mean.ale=median(temp$diff.mean.ale),
mos.median.ale.pre=median(temp$median.ale.pre),
mos.median.ale.post=median(temp$median.ale.post),
mos.diff.median.ale=median(temp$diff.median.ale),
mos.coh.d_ale.pre.post=median(temp$coh.d_ale.pre.post),
mos.mean.diff.ale=median(temp$mean.diff.ale),
mos.median.diff.ale=median(temp$median.diff.ale),
mos.prop.bs.impr=median(temp$prop.bs.impr),
mos.mean.bs.pre=median(temp$mean.bs.pre),
mos.mean.bs.post=median(temp$mean.bs.post),
mos.diff.mean.bs=median(temp$diff.mean.bs),
mos.median.bs.pre=median(temp$median.bs.pre),
mos.median.bs.post=median(temp$median.bs.post),
mos.diff.median.bs=median(temp$diff.median.bs),
mos.coh.d_bs.pre.post=median(temp$coh.d_bs.pre.post),
mos.mean.diff.bs=median(temp$mean.diff.bs),
mos.median.diff.bs=median(temp$median.diff.bs),
mos.rel.pre=median(temp$rel.pre),
mos.rel.post=median(temp$rel.post),
mos.diff.rel=median(temp$diff.rel),
mos.res.pre=median(temp$res.pre),
mos.res.post=median(temp$res.post),
mos.diff.res=median(temp$diff.res),
mos.uncertainty=median(temp$uncertainty))

data.out <- rbind(data.out,newrow)

n <- n+1
}

data.out <- data.out[-1,]
return(data.out)

#rm(sample.size,method,log.odds,n.resample,data.out,keepers,n,temp,newrow)
}
#####
#####
```

```

##### Run run.oosp_gjp for 20, 30,40,50,60,70,80,90,100 data points
#####
set.seed(826)
recal_20 <- run.oosp_gjp(20,method="mean",log.odds=TRUE,n.resample=100)

set.seed(826)
recal_30 <- run.oosp_gjp(30,method="mean",log.odds=TRUE,n.resample=100)

set.seed(826)
recal_40 <- run.oosp_gjp(40,method="mean",log.odds=TRUE,n.resample=100)

set.seed(826)
recal_50 <- run.oosp_gjp(50,method="mean",log.odds=TRUE,n.resample=100)

#####
##### Write recalibration results to CSVs
#####
write.csv(recal_20, "recal_20.csv", row.names=F)
write.csv(recal_30, "recal_30.csv", row.names=F)
write.csv(recal_40, "recal_40.csv", row.names=F)
write.csv(recal_50, "recal_50.csv", row.names=F)

#####
#####
#          EFFECTS OF CREDIBILITY-BASED RECALIBRATION
#
#####

#####
##### Effects of recalibration on prop.samp.rmse.impr
#####

##### Q: In what proportion of samples did recalibration reduce RMSE (judgments vs. "truth")?

## Descriptive Stats (across "forecasters")
mean(recal_50$prop.samp.rmse.impr)
#mean = 0.9605 = 96.05%

sd(recal_50$prop.samp.rmse.impr)
#sd = 0.0745 = 7.45%

median(recal_50$prop.samp.rmse.impr)
#med = 0.99 = 99%

range(recal_50$prop.samp.rmse.impr)
#range = [0.32,1] = [32%, 100%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$prop.samp.rmse.impr, xlim=c(0,1), breaks=50,
     main="Prop. of Samples in which Recalibration Improved RMSE\n(Prop. Across Each Forecaster's
100 Resamples)",
     xlab="Proportion of Resamples")

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$prop.samp.rmse.impr > 0.5, 1, 0))

```

```

##Mass of dist. > 0.5 = 0.9987 = 99.87%

wilcox.test(recal_50$prop.samp.rmse.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 284620, p < 0.001***

#####
##### Effects of recalibration on aos.prop.aje.impr
#####

#### Q: On average, what proportion of judgments saw reduced AJE as a result of recalibration?

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.prop.aje.impr)
#mean = 0.6850 = 68.50%

sd(recal_50$aos.prop.aje.impr)
#sd = 0.0545 = 5.45%

median(recal_50$aos.prop.aje.impr)
#med = 0.6871 = 68.71%

range(recal_50$aos.prop.aje.impr)
#range = [0.4671,0.8164] = [46.71%, 81.64%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.prop.aje.impr, xlim=c(0,1), breaks=50,
     main="Avg. Prop. of Judgments for which Recalibration Improved AJE\n(Prop. Across Each
Forecaster's 100 Resamples)",
     xlab="Proportion of Judgments")

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.prop.aje.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.9973 = 99.73%

wilcox.test(recal_50$aos.prop.aje.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 284620, p < 0.001***

#####
##### Effects of recalibration on aos.diff.mean.aje_lo
#####

#### Q: What was the average change in mean AJE (log-odds) as result of recalibration?
#### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.diff.mean.aje_lo)
#mean = 0.5439

sd(recal_50$aos.diff.mean.aje_lo)
#sd = 0.3347

median(recal_50$aos.diff.mean.aje_lo)
#med = 0.4939

range(recal_50$aos.diff.mean.aje_lo)
#range = [-0.0564,2.5436]

```

```

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.diff.mean.aje_lo, xlim=c(-1,3), breaks=50,
      main="Avg. Change in Mean AJE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
      xlab="Average Difference (log-odds)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.diff.mean.aje_lo > 0, 1, 0))
#Mass of dist. > 0 = 0.9894 = 98.94%

wilcox.test(recal_50$aos.diff.mean.aje_lo)
#Wilcoxon Signed-Rank Test: V = 284540, p < 0.001***



#####
##### Effects of recalibration on aos.diff.mean.aje_p
#####

##### Q: What was the average change in mean AJE (prob) as result of recalibration?
##### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.diff.mean.aje_p)
#mean = 0.0684 = 6.84%

sd(recal_50$aos.diff.mean.aje_p)
#sd = 0.0500 = 5.00%

median(recal_50$aos.diff.mean.aje_p)
#med = 0.0584 = 5.84%

range(recal_50$aos.diff.mean.aje_p)
#range = [-0.0132,0.2898] = [-1.32%, 28.98%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.diff.mean.aje_p, xlim=c(-0.4,0.4), breaks=50,
      main="Avg. Change in Mean AJE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
      xlab="Average Difference (Prob. Scale)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.diff.mean.aje_p > 0, 1, 0))
#Mass of dist. > 0 = 0.9801 = 98.01%

wilcox.test(recal_50$aos.diff.mean.aje_p)
#Wilcoxon Signed-Rank Test: V = 284360, p < 0.001***


## Visualization of Mean AJE (prob) Pre vs. Post (each observation is an individual "forecaster")
hist(recal_50$aos.mean.aje.pre_p, ylim=c(0,50), breaks=50, col=rgb(1,0,0,1/4),
      main="Mean AJE, Before and After Recalibration\n(Avg. Across Each Forecaster's 100
Resamples)",
      xlab="Mean AJE (Probability Scale)")

hist(recal_50$aos.mean.aje.post_p, ylim=c(0,50), xlim=c(0,0.6), breaks=50, col=rgb(0,0,1,1/4),
      add=T)

```

```

legend(x=0.4,y=45,legend=c("Before Recalibration","After Recalibration"),
col=c(rgb(1,0,0,1/4),rgb(0,0,1,1/4)),
lwd=2, lty=1)

#####
##### Effects of recalibration on prop.samp.mean.aje.impr
#####

#### Q: In what proportion of samples did recalibration reduce mean AJE (judgments vs. "truth")?

## Descriptive Stats (across "forecasters")
mean(recal_50$prop.samp.mean.aje.impr)
#mean = 0.9803 = 98.03%

sd(recal_50$prop.samp.mean.aje.impr)
#sd = 0.0685 = 6.85%

median(recal_50$prop.samp.mean.aje.impr)
#med = 1 = 100%

range(recal_50$prop.samp.mean.aje.impr)
#range = [0.35,1] = [35%, 100%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$prop.samp.mean.aje.impr, xlim=c(0,1), breaks=50,
     main="Prop. of Samples in which Recalibration Improved Mean AJE\n(Prop. Across Each
Forecaster's 100 Resamples)",
     xlab="Proportion of Resamples")

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$prop.samp.mean.aje.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.9934 = 99.34%

wilcox.test(recal_50$prop.samp.mean.aje.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 283110, p < 0.001***

#####
##### Effects of recalibration on aos.diff.median.aje_lo
#####

#### Q: What was the average change in Median AJE (log-odds) as result of recalibration?
#### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.diff.median.aje_lo)
#mean = 0.7310

sd(recal_50$aos.diff.median.aje_lo)
#sd = 0.4440

median(recal_50$aos.diff.median.aje_lo)
#med = 0.6855

range(recal_50$aos.diff.median.aje_lo)
#range = [-0.2482,2.8353]

```

```

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.diff.median.aje_lo, xlim=c(-1,3), breaks=50,
     main="Avg. Change in Median AJE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
     xlab="Average Difference (log-odds)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.diff.median.aje_lo > 0, 1, 0))
#Mass of dist. > 0 = 0.9735 = 97.35%

wilcox.test(recal_50$aos.diff.median.aje_lo)
#Wilcoxon Signed-Rank Test: V = 284150, p < 0.001***



#####
##### Effects of recalibration on aos.diff.median.aje_p
#####

##### Q: What was the average change in median AJE (prob) as result of recalibration?
##### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.diff.median.aje_p)
#mean = 0.1081 = 10.81%

sd(recal_50$aos.diff.median.aje_p)
#sd = 0.0811 = 8.11%

median(recal_50$aos.diff.median.aje_p)
#med = 0.0869 = 8.69%

range(recal_50$aos.diff.median.aje_p)
#range = [-0.0165,0.5178] = [-1.65%, 51.78%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.diff.median.aje_p, xlim=c(-0.1,0.6), breaks=50,
     main="Avg. Change in Median AJE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
     xlab="Average Difference (prob. scale)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.diff.median.aje_p > 0, 1, 0))
#Mass of dist. > 0 = 0.9801 = 98.01%

wilcox.test(recal_50$aos.diff.median.aje_p)
#Wilcoxon Signed-Rank Test: V = 284310, p < 0.001***


## Visualization of median AJE (prob) Pre vs. Post (each observation is an individual "forecaster")
hist(recal_50$aos.median.aje.pre_p, ylim=c(0,50), breaks=50, col=rgb(1,0,0,1/4),
     main="Median AJE, Before and After Recalibration\n(Avg. Across Each Forecaster's 100
Resamples)",
     xlab="Median AJE (Probability Scale)")

hist(recal_50$aos.median.aje.post_p, ylim=c(0,50), xlim=c(0,0.7), breaks=50, col=rgb(0,0,1,1/4),
     add=T)

```

```

legend(x=0.5,y=45,legend=c("Before Recalibration","After Recalibration"),
col=c(rgb(1,0,0,1/4),rgb(0,0,1,1/4)),
lwd=2, lty=1)

#####
##### Effects of recalibration on prop.samp.median.aje.impr
#####

#### Q: In what proportion of samples did recalibration reduce median AJE (judgments vs. "truth")?

## Descriptive Stats (across "forecasters")
mean(recal_50$prop.samp.median.aje.impr)
#mean = 0.9601 = 96.01%

sd(recal_50$prop.samp.median.aje.impr)
#sd = 0.1198 = 11.98%

median(recal_50$prop.samp.median.aje.impr)
#med = 1 = 100%

range(recal_50$prop.samp.median.aje.impr)
#range = [0.10,1] = [10%, 100%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$prop.samp.median.aje.impr, xlim=c(0,1), breaks=50,
     main="Prop. of Samples in which Recalibration Improved Median AJE\n(Prop. Across Each
Forecaster's 100 Resamples)",
     xlab="Proportion of Resamples")

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$prop.samp.median.aje.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.9761 = 97.61%

wilcox.test(recal_50$prop.samp.median.aje.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 283470, p < 0.001***

#####
##### Effects of recalibration on aos.coh.d_aje.pre.post_p
#####

#### Q: On average, what was the effect size (Cohen's D) of recalibration on AJE?

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.coh.d_aje.pre.post_p)
#mean = 0.41

sd(recal_50$aos.coh.d_aje.pre.post_p)
#sd = 0.17

median(recal_50$aos.coh.d_aje.pre.post_p)
#med = 0.41

range(recal_50$aos.coh.d_aje.pre.post_p)
#range = [-0.11, 1.05]

```

```

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.coh.d_aje.pre.post_p, xlim=c(-0.5,1.5), breaks=50,
  main="Avg. Effect (Cohen's D) of Recalibration on AJE (Prob. Scale)\n(Avg. Across Each
Forecaster's 100 Resamples)",
  xlab="Cohen's D")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.coh.d_aje.pre.post_p > 0, 1, 0))
#Mass of dist. > 0 = 0.9920 = 99.20%

wilcox.test(recal_50$aos.coh.d_aje.pre.post_p)
#Wilcoxon Signed-Rank Test: V = 284550, p < 0.001***

#####
##### Effects of recalibration on aos.mean.diff.aje_lo
#####

##### Q: On average, what was the median pairwise change in the AJE (log-odds), due to recalibration?
##### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.mean.diff.aje_lo)
#mean = 0.5439

sd(recal_50$aos.mean.diff.aje_lo)
#sd = 0.3347

median(recal_50$aos.mean.diff.aje_lo)
#med = 0.4939

range(recal_50$aos.mean.diff.aje_lo)
#range = [-0.0564, 2.5436]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.mean.diff.aje_lo, xlim=c(-1,3), breaks=50,
  main="AOS, Avg. Pairwise Difference in AJE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
  xlab="Avg. Difference (log-odds)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.mean.diff.aje_lo > 0, 1, 0))
#Mass of dist. > 0 = 0.9894 = 98.94%

wilcox.test(recal_50$aos.mean.diff.aje_lo)
#Wilcoxon Signed-Rank Test: V = 284540, p < 0.001***

#####
##### Effects of recalibration on aos.median.diff.aje_lo
#####

##### Q: On average, what was the median pairwise change in the AJE (log-odds), due to recalibration?
##### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

```

```

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.median.diff.aje_lo)
#mean = 1.0271

sd(recal_50$aos.median.diff.aje_lo)
#sd = 0.5287

median(recal_50$aos.median.diff.aje_lo)
#med = 0.9684

range(recal_50$aos.median.diff.aje_lo)
#range = [-0.0715, 2.9019]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.median.diff.aje_lo, xlim=c(-1,3), breaks=50,
     main="AOS, Med. Pairwise Difference in AJE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
     xlab="Avg. Difference (log-odds)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.median.diff.aje_lo > 0, 1, 0))
#Mass of dist. > 0 = 0.9947 = 99.47%

wilcox.test(recal_50$aos.median.diff.aje_lo)
#Wilcoxon Signed-Rank Test: V = 284600, p < 0.001***

#####
#####

##### Effects of recalibration on aos.mean.diff.aje_p
#####

#### Q: On average, what was the median pairwise change in the AJE (prob), due to recalibration?
#### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.mean.diff.aje_p)
#mean = 0.0684 = 6.84%

sd(recal_50$aos.mean.diff.aje_p)
#sd = 0.0500 = 5.00%

median(recal_50$aos.mean.diff.aje_p)
#med = 0.0584 = 5.84%

range(recal_50$aos.mean.diff.aje_p)
#range = [-0.0132, 0.2898] = [-1.32%, 28.98%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.mean.diff.aje_p, xlim=c(-0.1,0.3), breaks=50,
     main="AOS, Avg. Pairwise Difference in AJE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
     xlab="Avg. Difference (Prob. Scale)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.mean.diff.aje_p > 0, 1, 0))
#Mass of dist. > 0 = 0.9801 = 98.01%

```

```

wilcox.test(recal_50$aos.mean.diff.aje_p)
#Wilcoxon Signed-Rank Test: V = 284360, p < 0.001***

#####
##### Effects of recalibration on aos.median.diff.aje_p
#####

##### Q: On average, what was the median pairwise change in the AJE (log-odds), due to recalibration?
##### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.median.diff.aje_p)
#mean = 0.0896 = 8.96%

sd(recal_50$aos.median.diff.aje_p)
#sd = 0.0768 = 7.68%

median(recal_50$aos.median.diff.aje_p)
#med = 0.0699 = 6.99%

range(recal_50$aos.median.diff.aje_p)
#range = [-0.0012, 0.4126] = [-0.12%, 41.26%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.median.diff.aje_p, xlim=c(-0.1,0.5), breaks=50,
     main="AOS, Med. Pairwise Difference in AJE due to Recalibration (Pre - Post)\n(Avg. Across Each Forecaster's 100 Resamples)",
     xlab="Avg. Difference (Prob. Scale)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.median.diff.aje_p > 0, 1, 0))
#Mass of dist. > 0 = 0.9907 = 99.07%

wilcox.test(recal_50$aos.median.diff.aje_p)
#Wilcoxon Signed-Rank Test: V = 284580, p < 0.001***

#####
##### Effects of recalibration on aos.prop.ale.impr
#####

##### Q: On average, what proportion of judgments saw reduced ALE as a result of recalibration?

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.prop.ale.impr)
#mean = 0.7618 = 76.18%

sd(recal_50$aos.prop.ale.impr)
#sd = 0.0963 = 9.63%

median(recal_50$aos.prop.ale.impr)
#med = 0.7788 = 77.88%

range(recal_50$aos.prop.ale.impr)
#range = [0.2086, 0.9296] = [20.86%, 92.96%]

```

```

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.prop.ale.impr, xlim=c(0,1), breaks=50,
  main="Avg. Prop. of Judgments for which Recalibration Improved ALE\n(Prop. Across Each
Forecaster's 100 Resamples)",
  xlab="Proportion of Judgments")

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.prop.ale.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.9682 = 96.82%

wilcox.test(recal_50$aos.prop.ale.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 282440, p < 0.001***

#####
##### Effects of recalibration on aos.diff.mean.ale
#####

#### Q: What was the average change in mean ALE (prob) as result of recalibration?
#### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.diff.mean.ale)
#mean = 0.0707 = 7.07%

sd(recal_50$aos.diff.mean.ale)
#sd = 0.0487 = 4.87%

median(recal_50$aos.diff.mean.ale)
#med = 0.0623 = 6.23%

range(recal_50$aos.diff.mean.ale)
#range = [-0.0284, 0.2636] = [-2.84%, 26.36%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.diff.mean.ale, xlim=c(-0.1,0.3), breaks=50,
  main="Avg. Change in Mean ALE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
  xlab="Average Difference (Prob. Scale)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.diff.mean.ale > 0, 1, 0))
#Mass of dist. > 0 = 0.9695 = 96.95%

wilcox.test(recal_50$aos.diff.mean.ale)
#Wilcoxon Signed-Rank Test: V = 283730, p < 0.001***

## Visualization of mean ALE (prob) Pre vs. Post (each observation is an individual "forecaster")
hist(recal_50$aos.mean.ale.pre, ylim=c(0,50), xlim=c(0,0.6), breaks=50, col=rgb(1,0,0,1/4),
  main="Mean ALE, Before and After Recalibration\n(Avg. Across Each Forecaster's 100
Resamples)",
  xlab="Mean ALE (Probability Scale)")

hist(recal_50$aos.mean.ale.post, ylim=c(0,50), xlim=c(0,0.6), breaks=50, col=rgb(0,0,1,1/4), add=T)

legend(x=0.4,y=45,legend=c("Before Recalibration","After Recalibration"),

```

```

col=c(rgb(1,0,0,1/4),rgb(0,0,1,1/4)),
  lwd=2, lty=1)

#####
##### Effects of recalibration on prop.samp.mean.ale.impr
#####

##### Q: In what proportion of samples did recalibration reduce mean ALE (judgments vs. outcomes)?

## Descriptive Stats (across "forecasters")
mean(recal_50$prop.samp.mean.ale.impr)
#mean = 0.9545 = 95.45%

sd(recal_50$prop.samp.mean.ale.impr)
#sd = 0.1367 = 13.67%

median(recal_50$prop.samp.mean.ale.impr)
#med = 1 = 100%

range(recal_50$prop.samp.mean.ale.impr)
#range = [0, 1] = [0%, 100%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$prop.samp.mean.ale.impr, xlim=c(0,1), breaks=50,
  main="Prop. of Samples in which Recalibration Improved Mean ALE\n(Prop. Across Each
Forecaster's 100 Resamples)",
  xlab="Proportion of Resamples")

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$prop.samp.mean.ale.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.9761 = 97.61%

wilcox.test(recal_50$prop.samp.mean.ale.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 282430, p < 0.001***

#####
##### Effects of recalibration on aos.diff.median.ale
#####

##### Q: What was the average change in median ALE (prob) as result of recalibration?
##### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.diff.median.ale)
#mean = 0.1376 = 13.76%

sd(recal_50$aos.diff.median.ale)
#sd = 0.0948 = 9.48%

median(recal_50$aos.diff.median.ale)
#med = 0.1217 = 12.17%

range(recal_50$aos.diff.median.ale)
#range = [-0.1231, 0.4604] = [-12.31%, 46.04%]

```

```

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.diff.median.ale, xlim=c(-0.2,0.6), breaks=50,
      main="Avg. Change in Median ALE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
      xlab="Average Difference (Prob. Scale)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.diff.median.ale > 0, 1, 0))
#Mass of dist. > 0 = 0.9668 = 96.68%

wilcox.test(recal_50$aos.diff.median.ale)
#Wilcoxon Signed-Rank Test: V = 283260, p < 0.001***

## Visualization of median ALE (prob) Pre vs. Post (each observation is an individual "forecaster")
hist(recal_50$aos.median.ale.pre, ylim=c(0,50), xlim=c(0,0.7), breaks=50, col=rgb(1,0,0,1/4),
      main="Median ALE, Before and After Recalibration\n(Avg. Across Each Forecaster's 100
Resamples)",
      xlab="Median ALE (Probability Scale)")

hist(recal_50$aos.median.ale.post, ylim=c(0,50), xlim=c(0,0.7), breaks=50, col=rgb(0,0,1,1/4),
      add=T)

legend(x=0.5,y=45,legend=c("Before Recalibration","After Recalibration"),
       col=c(rgb(1,0,0,1/4),rgb(0,0,1,1/4)),
       lwd=2, lty=1)

#####
##### Effects of recalibration on prop.samp.mean.ale.impr
#####

##### Q: In what proportion of samples did recalibration reduce median ALE (judgments vs. outcomes)??

## Descriptive Stats (across "forecasters")
mean(recal_50$prop.samp.median.ale.impr)
#mean = 0.9632 = 96.32%

sd(recal_50$prop.samp.median.ale.impr)
#sd = 0.1545 = 15.45%

median(recal_50$prop.samp.median.ale.impr)
#med = 1 = 100%

range(recal_50$prop.samp.median.ale.impr)
#range = [0,1] = [0%, 100%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$prop.samp.median.ale.impr, xlim=c(0,1), breaks=50,
      main="Prop. of Samples in which Recalibration Improved Median ALE\n(Prop. Across Each
Forecaster's 100 Resamples)",
      xlab="Proportion of Resamples")

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$prop.samp.median.ale.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.9655 = 96.55%

wilcox.test(recal_50$prop.samp.median.ale.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 281550, p < 0.001***

```

```

#####
##### Effects of recalibration on aos.coh.d_ale.pre.post
#####

##### Q: On average, what was the effect size (Cohen's D) of recalibration on ALE?

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.coh.d_ale.pre.post)
#mean = 0.43

sd(recal_50$aos.coh.d_ale.pre.post)
#sd = 0.21

median(recal_50$aos.coh.d_ale.pre.post)
#med = 0.42

range(recal_50$aos.coh.d_ale.pre.post)
#range = [-0.30, 1.51]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.coh.d_ale.pre.post, xlim=c(-0.5,2), breaks=50,
     main="Avg. Effect (Cohen's D) of Recalibration on ALE (Prob. Scale)\n(Avg. Across Each
Forecaster's 100 Resamples)",
     xlab="Cohen's D")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.coh.d_ale.pre.post > 0, 1, 0))
#Mass of dist. > 0 = 0.9775 = 97.75%

wilcox.test(recal_50$aos.coh.d_ale.pre.post)
#Wilcoxon Signed-Rank Test: V = 283870, p < 0.001***

#####
##### Effects of recalibration on aos.mean.diff.ale
#####

##### Q: On average, what was the mean pairwise change in the ALE (prob), due to recalibration?
##### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.mean.diff.ale)
#mean = 0.0707 = 7.07%

sd(recal_50$aos.mean.diff.ale)
#sd = 0.0487 = 4.87%

median(recal_50$aos.mean.diff.ale)
#med = 0.0623 = 6.23%

range(recal_50$aos.mean.diff.ale)
#range = [-0.0284, 0.2636] = [-2.84%, 26.36%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.mean.diff.ale, xlim=c(-0.1,0.3), breaks=50,

```

```

    main="AOS, Avg. Pairwise Difference in ALE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
    xlab="Avg. Difference (Prob. Scale)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.mean.diff.ale > 0, 1, 0))
#Mass of dist. > 0 = 0.9695 = 96.95%

wilcox.test(recal_50$aos.mean.diff.ale)
#Wilcoxon Signed-Rank Test: V = 283730, p < 0.001***

#####
##### Effects of recalibration on aos.median.diff.ale
#####

#### Q: On average, what was the median pairwise change in the ALE (prob), due to recalibration?
#### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.median.diff.ale)
#mean = 0.1010 = 10.10%

sd(recal_50$aos.median.diff.ale)
#sd = 0.0826 = 8.26%

median(recal_50$aos.median.diff.ale)
#med = 0.0787 = 7.87%

range(recal_50$aos.median.diff.ale)
#range = [-0.0424, 0.4151] = [-4.24%, 41.51%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.median.diff.ale, xlim=c(-0.1,0.5), breaks=50,
     main="AOS, Med. Pairwise Difference in ALE due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
     xlab="Avg. Difference (Prob. Scale)")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.median.diff.ale > 0, 1, 0))
#Mass of dist. > 0 = 0.9629 = 96.29%

wilcox.test(recal_50$aos.median.diff.ale)
#Wilcoxon Signed-Rank Test: V = 283140, p < 0.001***

#####
##### Effects of recalibration on aos.prop.bs.impr
#####

#### Q: On average, what proportion of judgments saw reduced BS as a result of recalibration?

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.prop.bs.impr)
#mean = 0.7642 = 76.42%

```

```

sd(recal_50$aos.prop.bs.impr)
#sd = 0.0963 = 9.63%

median(recal_50$aos.prop.bs.impr)
#med = 0.7788 = 77.88%

range(recal_50$aos.prop.bs.impr)
#range = [0.2086, 0.9296] = [20.86%, 92.96%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.prop.bs.impr, xlim=c(0,1), breaks=50,
     main="Avg. Prop. of Judgments for which Recalibration Improved BS\n(Prop. Across Each
Forecaster's 100 Resamples)",
     xlab="Proportion of Judgments")

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.prop.bs.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.9682 = 96.82%

wilcox.test(recal_50$aos.prop.bs.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 282440, p < 0.001***  

#####
##### Effects of recalibration on aos.diff.mean.bs
#####

#### Q: What was the average change in mean BS as result of recalibration?
#### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.diff.mean.bs)
#mean = 0.0142

sd(recal_50$aos.diff.mean.bs)
#sd = 0.0290

median(recal_50$aos.diff.mean.bs)
#med = 0.0064

range(recal_50$aos.diff.mean.bs)
#range = [-0.0260, 0.2703]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.diff.mean.bs, xlim=c(-0.1,0.3), breaks=50,
     main="Avg. Change in Mean BS due to Recalibration (Pre - Post)\n(Avg. Across Each Forecaster's
100 Resamples)",
     xlab="Average Difference")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.diff.mean.bs > 0, 1, 0))
#Mass of dist. > 0 = 0.6525 = 65.25%  

wilcox.test(recal_50$aos.diff.mean.bs)
#Wilcoxon Signed-Rank Test: V = 222130, p < 0.001***
```

```

## Visualization of mean BS Pre vs. Post (each observation is an individual "forecaster")
hist(recal_50$aos.mean.bs.pre, ylim=c(0,70), xlim=c(0,0.6), breaks=50, col=rgb(1,0,0,1/4),
  main="Mean BS, Before and After Recalibration\n(Avg. Across Each Forecaster's 100 Resamples)",
  xlab="Mean BS")

hist(recal_50$aos.mean.bs.post, ylim=c(0,70), xlim=c(0,0.6), breaks=50, col=rgb(0,0,1,1/4), add=T)

legend(x=0.4,y=45,legend=c("Before Recalibration","After Recalibration"),
col=c(rgb(1,0,0,1/4),rgb(0,0,1,1/4)),
lwd=2, lty=1)

#####
#####

##### Effects of recalibration on prop.samp.mean.bs.impr
#####

##### Q: In what proportion of samples did recalibration reduce mean BS (judgments vs. outcomes)?

## Descriptive Stats (across "forecasters")
mean(recal_50$prop.samp.mean.bs.impr)
#mean = 0.6663 = 66.63%

sd(recal_50$prop.samp.mean.bs.impr)
#sd = 0.2527 = 25.27%

median(recal_50$prop.samp.mean.bs.impr)
#med = 0.72 = 72%

range(recal_50$prop.samp.mean.bs.impr)
#range = [0, 1] = [0%, 100%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$prop.samp.mean.bs.impr, xlim=c(0,1), breaks=50,
  main="Prop. of Samples in which Recalibration Improved Mean BS\n(Prop. Across Each
Forecaster's 100 Resamples)",
  xlab="Proportion of Resamples")

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$prop.samp.mean.bs.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.7268 = 72.68%

wilcox.test(recal_50$prop.samp.mean.bs.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 227540, p < 0.001***

#####
#####

##### Effects of recalibration on aos.diff.median.bs
#####

##### Q: What was the average change in median BS as result of recalibration?
##### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.diff.median.bs)
#mean = 0.0530

sd(recal_50$aos.diff.median.bs)

```

```

#sd = 0.0564

median(recal_50$aos.diff.median.bs)
#med = 0.0344

range(recal_50$aos.diff.median.bs)
#range = [-0.0474, 0.5271]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.diff.median.bs, xlim=c(-0.2,0.6), breaks=50,
  main="Avg. Change in Median BS due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
  xlab="Average Difference")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.diff.median.bs > 0, 1, 0))
#Mass of dist. > 0 = 0.9668 = 96.55%

wilcox.test(recal_50$aos.diff.median.bs)
#Wilcoxon Signed-Rank Test: V = 283350, p < 0.001***

## Visualization of median BS Pre vs. Post (each observation is an individual "forecaster")
hist(recal_50$aos.median.bs.pre, ylim=c(0,150), xlim=c(0,0.6), breaks=50, col=rgb(1,0,0,1/4),
  main="Median BS, Before and After Recalibration\n(Avg. Across Each Forecaster's 100
Resamples)",
  xlab="Median BS")

hist(recal_50$aos.median.bs.post, ylim=c(0,150), xlim=c(0,0.6), breaks=30, col=rgb(0,0,1,1/4),
  add=T)

legend(x=0.4,y=140,legend=c("Before Recalibration","After Recalibration"),
col=c(rgb(1,0,0,1/4),rgb(0,0,1,1/4)),
lwd=2, lty=1)

#####
#### Effects of recalibration on prop.samp.median.bs.impr
#####

##### Q: In what proportion of samples did recalibration reduce median BS (judgments vs. outcomes)?

## Descriptive Stats (across "forecasters")
mean(recal_50$prop.samp.median.bs.impr)
#mean = 0.9633 = 96.33%

sd(recal_50$prop.samp.median.bs.impr)
#sd = 0.1541 = 15.41%

median(recal_50$prop.samp.median.bs.impr)
#med = 1 = 100%

range(recal_50$prop.samp.median.bs.impr)
#range = [0, 1] = [0%, 100%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$prop.samp.median.bs.impr, xlim=c(0,1), breaks=50,
  main="Prop. of Samples in which Recalibration Improved Median BS\n(Prop. Across Each
Forecaster's 100 Resamples)",
  xlab="Proportion of Resamples")

```

```

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$prop.samp.median.bs.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.9655 = 96.55%

wilcox.test(recal_50$prop.samp.median.bs.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 281560, p < 0.001***

#####
##### Effects of recalibration on aos.coh.d_bs.pre.post
#####

##### Q: On average, what was the effect size (Cohen's D) of recalibration on BS?

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.coh.d_bs.pre.post)
#mean = 0.07

sd(recal_50$aos.coh.d_bs.pre.post)
#sd = 0.11

median(recal_50$aos.coh.d_bs.pre.post)
#med = 0.07

range(recal_50$aos.coh.d_bs.pre.post)
#range = [-0.25, 0.53]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.coh.d_bs.pre.post, xlim=c(-0.6,0.6), breaks=50,
     main="Avg. Effect (Cohen's D) of Recalibration on BS\n(Avg. Across Each Forecaster's 100
Resamples)",
     xlab="Cohen's D")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.coh.d_bs.pre.post > 0, 1, 0))
#Mass of dist. > 0 = 0.7414 = 74.14%

wilcox.test(recal_50$aos.coh.d_bs.pre.post)
#Wilcoxon Signed-Rank Test: V = 238460, p < 0.001***

#####
##### Effects of recalibration on aos.mean.diff.bs
#####

##### Q: On average, what was the mean pairwise change in BS, due to recalibration?
##### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.mean.diff.bs)
#mean = 0.0142

sd(recal_50$aos.mean.diff.bs)
#sd = 0.0290

```

```

median(recal_50$aos.mean.diff.bs)
#med = 0.0064

range(recal_50$aos.mean.diff.bs)
#range = [-0.0260, 0.2703]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.mean.diff.bs, xlim=c(-0.1,0.3), breaks=50,
     main="AOS, Avg. Pairwise Difference in BS due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
     xlab="Avg. Difference")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.mean.diff.bs > 0, 1, 0))
#Mass of dist. > 0 = 0.6525 = 65.25%

wilcox.test(recal_50$aos.mean.diff.bs)
#Wilcoxon Signed-Rank Test: V = 222130, p < 0.001***

#####
#####

##### Effects of recalibration on aos.median.diff.bs
#####

##### Q: On average, what was the median pairwise change in BS, due to recalibration?
##### Note: differences calculated as err.pre - err.post --> positive values indicate reduction in
error

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.median.diff.bs)
#mean = 0.0278

sd(recal_50$aos.median.diff.bs)
#sd = 0.0375

median(recal_50$aos.median.diff.bs)
#med = 0.0132

range(recal_50$aos.median.diff.bs)
#range = [-0.0028, 0.2421]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.median.diff.bs, xlim=c(-0.1,0.3), breaks=50,
     main="AOS, Med. Pairwise Difference in BS due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
     xlab="Avg. Difference")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.median.diff.bs > 0, 1, 0))
#Mass of dist. > 0 = 0.9695 = 96.95%

wilcox.test(recal_50$aos.median.diff.bs)
#Wilcoxon Signed-Rank Test: V = 283610, p < 0.001***

#####
#####

```

```

#####
##### Effects of recalibration on aos.diff.rel
#####

##### Q: What was the average effect of recalibration on sample reliability (calibration)?
##### Note: differences calculated as rel.pre - rel.post --> positive values indicate reduction
(improvement) in reliability

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.diff.rel)
#mean = 0.0237

sd(recal_50$aos.diff.rel)
#sd = 0.0356

median(recal_50$aos.diff.rel)
#med = 0.0129

range(recal_50$aos.diff.rel)
#range = [-0.0229, 0.2704]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$aos.diff.rel, xlim=c(-0.1,0.3), breaks=50,
     main="Avg. Change in Sample Reliability due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
     xlab="Avg. Difference")

abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.diff.rel > 0, 1, 0))
#Mass of dist. > 0 = 0.8170 = 81.70%

wilcox.test(recal_50$aos.diff.rel)
#Wilcoxon Signed-Rank Test: V = 260720, p < 0.001***

#####
#####

##### Effects of recalibration on prop.samp.rel.impr
#####

##### Q: In what proportion of samples did recalibration improve reliability (calibration)?

## Descriptive Stats (across "forecasters")
mean(recal_50$prop.samp.rel.impr)
#mean = 0.7654 = 76.54%

sd(recal_50$prop.samp.rel.impr)
#sd = 0.2304 = 23.04%

median(recal_50$prop.samp.rel.impr)
#med = 0.84 = 84%

range(recal_50$prop.samp.rel.impr)
#range = [0.01, 1] = [1.00%, 100%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$prop.samp.rel.impr, xlim=c(0,1), breaks=50,
     main="Prop. of Samples in which Recalibration Improved Reliability\n(Prop. Across Each
Forecaster's 100 Resamples)",
     xlab="Proportion of Resamples")

```

```

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$prop.samp.rel.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.8448 = 84.48%

wilcox.test(recal_50$prop.samp.rel.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 259910, p < 0.001***

#####
##### Effects of recalibration on aos.diff.res
#####

##### Q: What was the average effect of recalibration on sample resolution?
##### Note: differences calculated as res.pre - res.post --> positive values indicate reduction
(decline) in resolution

## Descriptive Stats (across "forecasters")
mean(recal_50$aos.diff.res)
#mean = -0.0001

sd(recal_50$aos.diff.res)
#sd = 0.0003

median(recal_50$aos.diff.res)
#med = 0

range(recal_50$aos.diff.res)
#range = [-1.933275e-03, 2.775558e-19]

## Visualization of Distribution (each observation is an individual "forecaster")
#hist(recal_50$aos.diff.res, xlim=c(-0.1,0.5), breaks=50,
#     main="Avg. Change in Sample Resolution due to Recalibration (Pre - Post)\n(Avg. Across Each
Forecaster's 100 Resamples)",
#     xlab="Avg. Difference")

#abline(v=0, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$aos.diff.res < 0, 1, 0))
#Mass of dist. < 0 = 0.4204 = 42.04%

wilcox.test(recal_50$aos.diff.res)
#Wilcoxon Signed-Rank Test: V = 4.5, p < 0.001*** #NOTE: Effect is not in the expected direction

#####
##### Effects of recalibration on prop.samp.res.impr
#####

##### Q: In what proportion of samples did recalibration improve resolution?

## Descriptive Stats (across "forecasters")
mean(recal_50$prop.samp.res.impr)
#mean = 0.0446 = 4.46%

sd(recal_50$prop.samp.res.impr)
#sd = 0.0862 = 8.62%

```

```

median(recal_50$prop.samp.res.impr)
#med = 0 = 0%

range(recal_50$prop.samp.res.impr)
#range = [0,0.77] = [0%, 77%]

## Visualization of Distribution (each observation is an individual "forecaster")
hist(recal_50$prop.samp.res.impr, xlim=c(0,1), breaks=50,
     main="Prop. of Samples in which Recalibration Improved Resolution\n(Prop. Across Each
Forecaster's 100 Resamples)",
     xlab="Proportion of Resamples")

abline(v=0.5, col="red", lty=2, lwd=2)

## Hypothesis Tests
mean(ifelse(recal_50$prop.samp.res.impr > 0.5, 1, 0))
#Mass of dist. > 0.5 = 0.0040 = 0.40%

wilcox.test(recal_50$prop.samp.res.impr - 0.5)
#Wilcoxon Signed-Rank Test: V = 48, p < 0.001*** #NOTE: Effect is not in the expected direction

#####
#####
#####
```

## APPENDIX B

Appendix B contains PDF facsimiles of the 10 online surveys administered in the March Madness Study. These surveys are presented in the sequence they were administered to participants. The order of the surveys is as follows:

- Consent and additional participation information (Survey 1).
- Play-in game predictions (Survey 2; Field of 68).
- First round predictions (Survey 3; March Madness Round 1; Field of 64).
- Second round predictions A (Survey 4; March Madness Round 2; Field of 32).
- Second round predictions B (Survey 5; March Madness Round 2; Field of 32).
- Second round predictions C (Survey 6; March Madness Round 2; Field of 32).
- Sweet Sixteen predictions (Survey 7; March Madness Round 3; Field of 16).
- Elite Eight predictions (Survey 8; March Madness Round 4; Field of 8).
- Final Four predictions (Survey 9; March Madness Round 6; Field of 4).
- Championship game prediction (Survey 10; March Madness Round 7; Field of 2).

Within each survey, information about each game was presented using the Loop & Merge feature of the Qualtrics online survey platform. This feature allows participants to respond to the same set of questions several times, with each iteration containing unique text piped-in from an associated spreadsheet (and responses from each iteration recorded separately). In the case of the March Madness study, this feature was used to pipe-in information about the games in each survey, and to ask a uniform set of questions about each game. In the online version of each survey, this information was piped-in in

place of the generic variable fields displayed in the survey printouts that follow. In these printouts, Loop & Merge variable fields are indicated by the syntax “\${lm://Field/[n]}”, where [n] corresponds to the integer designation of one of the following fields:

- Field 1: Name of Team A (displayed on the left-hand side of the screen).
- Field 2: Regular season conference, Team A.
- Field 3: Division-1 win/loss record, Team A.
- Field 4: Strength of schedule, Team A (as reported by espn.com).
- Field 5: Name of Team B (displayed on the right-hand side of the screen).
- Field 6: Regular season conference, Team B.
- Field 7: Division-1 win/loss record, Team B.
- Field 8: Strength of schedule, Team B (as reported by espn.com).
- Field 9: Tournament round in which game is being played.
- Field 10: Tournament seeding, Team A.
- Field 11: Tournament seeding, Team B.

For convenience, the Loop & Merge data used in each survey is provided as a stand-alone spreadsheet. These spreadsheets are presented after the relevant survey printout. In each of these spreadsheets, information about games is presented in bracket order and/or play order. In the online version of each survey, all games were presented to participants in a separate, random order. In all cases where surveys were completed, participants were asked to make predictions about all games.

## **Important Information**

### **UPenn March Madness Study: Additional Information**

Congratulations! You are one of the lucky few who have been selected to participate in the UPenn March Madness Study!

Before we can add you to the official participation roster, it is important that we provide you with some additional information. **We ask that you please read this information carefully.**

---

#### **Participation Details:**

If you wish to participate in this study, it is important that you provide predictions for **every game** of the 2017 Men's NCAA basketball tournament, which runs from March 14th to April 3rd. We anticipate that this will require about 3 hours of your time.

However, because we are interested in your predictions about the **actual games** that are played (not your predictions about who will play whom), you will need to provide predictions at several stages throughout the tournament.

To do so, **you will be asked to complete a series of online surveys** that will be emailed to you before each game. These surveys will be sent out as soon as match-ups are determined. In some cases, this may mean that **you will only have a few days to provide your predictions.**

The first two of these surveys will be sent out on Selection Sunday (3/12/17). The first will ask for your predictions about the "first four" play-in games, and must be completed before the first play-in game begins on March 14th. The second will ask for your predictions about the first round of the main tournament, and must be completed before the first first-round game begins on March 16th.

To complete these two surveys, we anticipate that **you will need to set aside 60-90 minutes of your time between Sunday, March 12th and Thursday, March 16th.**

Upon the timely completion of the first two surveys, you will be paid \$10. If you fail to complete these surveys (or if they are submitted late), you will not receive any money for your participation.

Additional information about compensation (up to \$75 more) is provided on the next page.

---

**To acknowledge that you have read this information carefully, please ignore the following math problem and enter the phrase "got it" in the box below:**

$$(3 + 5) * (11 - 4) + 17 = \underline{\hspace{2cm}}$$

## Consent

### CONSENT TO PARTICIPATE IN RESEARCH

#### Purpose

The purpose of this study is to understand how people assign probabilities to uncertain events (e.g. a particular team winning a college basketball game), and to determine what makes some people better than others. In addition, it is hoped that we will be able to use the data gathered in this study to develop mathematical techniques for improving probabilistic predictions. If successful, these techniques will allow decision makers to more accurately assess the risk(s) of future events such as climate change, changes in public policy, and military interventions.

#### Procedures

For the duration of the 2017 NCAA Men's Basketball Championship (i.e., "March Madness"), you will be asked to predict which teams will win each game of the March Madness tournament. Based on these predictions, each participant will also have the opportunity to "wager" on the outcome of a single game and win up to \$15. Along with predicting the winner of each game, you will also be asked to provide numerical probability judgments describing your beliefs that a given team will win (i.e., if you predict that team A will win, we will ask you to provide a probability judgment that reflects your *confidence* that team A will win). When providing these confidence ratings, it is important that your judgments be as accurate as possible, as those who make the best predictions over the course of the tournament will be eligible for up to \$50 in additional prizes! Finally, we will ask all participants to complete several measures related to individual reasoning ability and a brief demographics questionnaire.

**Study time:** Predictions will be gathered before each game of the 2017 March Madness tournament. Total completion time for all surveys is expected to take approximately 3-hours.

#### Benefits

There are no direct benefits to participating in this study other than the financial compensation listed below. However, it is hoped that the results of this study will provide valuable insight into reasoning, judgment, and decision making processes.

### **Risks**

There are minimal foreseeable risks to participating in this study. However, if you are uncomfortable or wish to withdraw for any other reason, you may stop participating at any time.

### **Confidentiality**

All data collected during this experiment will be completely confidential, and will be stored at all times on a password protected computer.

**Retaining research records:** When the research is completed, we may save the electronic files of data for use in future research done by our research team or others. We will retain this study information for up to 5 years after the results of the study are published, to comply with American Psychological Association data-retention rules.

### **Compensation**

Participants will be paid a guaranteed \$10 for completing an initial prediction session in which they fill-out their first round NCAA brackets and complete several questionnaires about individual reasoning (~1 hour). At the conclusion of the tournament, participants who provided predictions for all 63 games will be paid an additional \$10, and will be eligible for up to \$65 in additional compensation. Specifically, all participants who complete the study in full will receive an amount between \$0 and \$15 for one of their "wagers" (selected at random), and the top three performers in the tournament (as measured by the accuracy of their confidence judgments) will receive \$50, \$25, and \$10 in Amazon gift cards, respectively.

### **Rights**

Participation in research is completely voluntary. You have the right to decline to participate or to withdraw at any point in this study without penalty or loss of benefits to which you are otherwise entitled.

### **Questions**

If you have any questions or concerns about this study, you may contact Josh Baker at jbak@sas.upenn.edu. If you have any questions or concerns about your rights and treatment as a research subject, you may contact the Office of Regulatory Affairs at the University of Pennsylvania, at 215.573.2540 or via email at burgess4@upenn.edu.

---

If you agree to participate in this study, please click the "I agree" button below. Otherwise, please click the button labelled "I would NOT like to participate in this study."

- I agree
- I would NOT like to participate in this study

## Opt-Out

You have decided to opt-out of this study.

Thank you for your time!

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

10

## Info and Demogs

To complete the study sign-up process, please provide the following information:

First Name

Last Name

Preferred Email Address

Confirm Email Address

## Opt-In

Congratulations! You have now been added to the official study roster.

You will receive the first two prediction surveys by email on Sunday, 3/12/17.

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

10



## **Introduction**

### **UPenn March Madness Study: Play-In Predictions**

**Start Date:** March 12th, 2017

**End Date:** March 14th, 2017 (prior to the first play-in game of the tournament)

**Number of Predictions in this Survey:** 4

**Expected Completion Time:** 10-15 min.

### **Instructions**

Hello, and welcome to the UPenn March Madness Study!

In today's survey, you will be asked to provide predictions about the outcomes of the four play-in games (i.e., the "First Four") of the 2017 NCAA Division I Men's Basketball Tournament (aka, "March Madness").

For each of these games, we will ask you to provide the following information:

1. A prediction about which team will win.
2. A numerical probability judgment that reflects your confidence that the team you selected will win (more on this in a moment).
3. The minimum price at which you would sell a claim ticket for a wager that pays \$15 if the team you selected wins (and \$0 otherwise).
4. Your choice to either keep the lottery ticket or sell it for the price specified above.

**Before you begin the prediction process, please read the following pages carefully.**

### **Things to Note:**

#### **General:**

- Surveys can be completed at any time before the specified end date, and do not need to be completed in a single sitting.
- When making your predictions, you are welcome to use outside resources! All that we ask is that your answers are your own -- please do not ask any other people to help you with your predictions.

**Payment:**

- Participants will be paid \$10 upon timely completion of this survey ("Play-In Predictions") and the other survey sent out on 3/12/17 ("First Round Predictions").
- Participants will be paid an additional \$10 for providing predictions for all of the remaining games in the tournament.
- Participants who complete this study will also have the opportunity to win up to \$15 more by correctly predicting the outcome of a single game (selected at random).
- The three participants who provide the most accurate probabilistic predictions over the course of the tournament (more on this in a moment) will receive an additional \$50, \$25, and \$10, respectively.
- Participants who do not complete this survey ("Play-In Predictions") and the other survey sent out on 3/12/17 ("First Round Predictions") before the specified end dates (3/14 and 3/16, respectively), **WILL NOT RECEIVE ANY PAYMENT FOR THEIR PARTICIPATION.**

If you have any questions or concerns about today's survey, please contact Josh Baker at [jbak@sas.upenn.edu](mailto:jbak@sas.upenn.edu)

- I acknowledge that I have read and understood the information on this page.

**Additional Resources**

- For any participants who are unfamiliar with the NCAA tournament, additional information about the structure and rules of the tournament can be found [here](#).
- For any participants who would like to see the latest news, information, and schedule for the tournament, all that and more can be found [here](#).
- For any participants who would like to make his or her predictions on paper, a printable version of the NCAA tournament bracket can be found [here](#).

## How to Provide Probability Estimates

In mathematics, probability is typically defined in terms of relative frequencies. When rolling a die, for example, you can estimate the probability of rolling a "4" by making a large number of rolls and counting the number of times a "4" comes up. Although this process isn't perfect, you will eventually find that the number of "4's" you roll is very close to one-sixth of the total number of rolls. Thus, it makes sense to say that the probability of rolling a "4" is one-out-of-six, or 1/6.

For a single basketball game, however, it doesn't make much sense to think about probability in this way. You can't discover the probability of Team A beating Team B by having them play the same game a large number of times. That one game, by definition, only happens once.

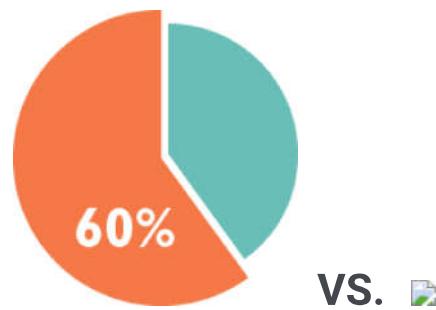
So how can you assign a probability to the outcome of a single game?

One way to proceed is to think about your preferences for various wagers regarding Team A. Suppose, for example, you were given a simple choice between betting on Team A or betting on a spinner (like a roulette wheel) that had a 50% chance of landing on a winning zone, both for the same prize. Which bet would you choose? If you would choose Team A, then your probability for Team A is higher than 50%.

Now, what if the spinner had a 60% chance of winning? At some point, as the probability of the spinner gets higher, you would switch and bet on the spinner rather than Team A. Just before this point, you would not care whether you bet on the spinner or Team A. Because the prizes for the two bets are the same, the fact that you are indifferent between the two suggests that you also believe the probabilities are the same. Thus, you can estimate the probability of Team A winning by thinking about the point where your preferences switch.

**Example:**

Which do you prefer? (1) A bet that pays \$15 if the spinner lands on orange; or (2) a bet that pays \$15 if Team A wins a basketball game?



If you prefer option 1 (the spinner), then the probability of Team A winning must be less than 60%. If you prefer option 2 (the basketball game), then the probability of Team A winning must be greater than 60%.

If the two options are exactly the same, then the probability of Team A winning must be exactly 60%.

You can use this technique to estimate the probability of Team A winning. Slowly raise the probability of an imaginary spinner until betting on the spinner and betting on the basketball game feel exactly the same. Whatever number your imaginary spinner shows at this point reflects your beliefs about the probability of Team A winning, or your confidence that Team A will win.

---

In this study, we will ask you to predict the winner of each NCAA game, and to make probability judgments that reflect your beliefs about the likelihood that your favored team will win.

In most cases, you will probably have some intuition about which team will win, but could imagine being wrong. In these cases, you should provide a probability judgment that is between 50% and 100%. In cases where you are more confident, your probability judgment should be closer to 100%, and in cases where you are less confident, your probability judgments should be closer to 50%.

Ultimately, your accuracy will be determined by how close your judgments come to the true outcome of each game. These outcomes are coded as "1" if the event happens (e.g., Team A wins) and "0" if it does not (e.g., Team A loses). The further you are from the truth, the more your accuracy score will suffer (and the less likely it is that you will win an Amazon gift card). Thus, it is in your best interest to provide the most accurate probability estimates possible.

It's as easy as that! That's everything you need to know to go make predictions!

**Have fun, and good luck!**

## Predictions

### Instructions:

Please review the following information and provide your prediction about who will win.

---

### **Game Information:**

$\$\{lm://Field/1\}$  vs.  $\$\{lm://Field/5\}$   
 $\$\{lm://Field/9\}$

<u><math>\\$\{lm://Field/1\}</math> Information:</u>	<u><math>\\$\{lm://Field/5\}</math> Information:</u>
Regular Season Conference: $\$\{lm://Field/2\}$	Regular Season Conference: $\$\{lm://Field/6\}$

Division 1 Win-Loss Record: \${Im://Field/3} <a href="#">\${Im://Field/1} Information</a> Strength of Schedule*: \${Im://Field/4} Tournament Seeding**: \${Im://Field/10}	Division 1 Win-Loss Record: \${Im://Field/7} <a href="#">\${Im://Field/5} Information</a> Strength of Schedule*: \${Im://Field/8} Tournament Seeding**: \${Im://Field/11}
--	--

\*Strength of Schedule is a numerical ranking provided by espn.com that reflects the difficulty of a team's regular season schedule. Teams with more difficult schedules (i.e., who play more skilled opponents) are given lower numerical rankings.

\*\* Tournament Seeding is a numerical ranking provided by the NCAA that determines who will play whom in the first round of the tournament. Seeding is done independently in each of the four quadrants

of the tournament bracket (corresponding to different geographical regions). Lower (i.e., better) seeding is generally given to teams with better win-loss records and more difficult regular season schedules.

Want to look at the big picture?

[See the whole bracket here](#)

**Which team do you think will win this game?**

- \${Im://Field/1}
- \${Im://Field/5}

**Please provide a probability estimate that reflects your confidence that \${Im://Field/1} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/1} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Please provide a probability estimate that reflects your confidence that \${Im://Field/5} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/5} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/1} wins, and \$0 if \${Im://Field/1} loses.

If you were forced to sell this ticket, what is the minimum amount you would sell it for?

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/5} wins, and \$0 if \${Im://Field/5} loses.

If you were forced to sell this ticket, what is the minimum amount you would sell it for?

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

If someone offered to pay the amount above, would you keep the claim ticket or sell it?

(Please keep in mind: at the end of the tournament, one of your choices will be selected at random and paid-out in real money)

- I would keep the ticket  
 I would sell it

Email

## **That's it for the Play-In Games!**

Thank you for completing this portion of the study. Please don't forget to complete the "First Round Predictions" survey before March 16th!

To ensure that your answers are associated with the right person, please confirm your contact information below.

First Name

Last Name

Powered by Qualtrics

### Loop and Merge Data: Play-In Games (Survey 2)

	<b>Field 1</b>	<b>Field 2</b>	<b>Field 3</b>	<b>Field 4</b>	<b>Field 5</b>	<b>Field 6</b>	<b>Field 7</b>	<b>Field 8</b>	<b>Field 9</b>	<b>Field 10</b>	<b>Field 11</b>
Mount St. Mary's	Northeast	19-15	205	New Orleans	Southland	17-11	292	Play-In Game	16	16	16
N.C. Central	Mid-Eastern	22-8	351	UC Davis	Big West	20-12	322	Play-In Game	16	16	16
Providence	Big East	20-12	49	USC	Pac-12	24-9	76	Play-In Game	11	11	11
Kansas St.	Big 12	20-13	43	Wake Forest	Atlantic Coast	19-13	21	Play-In Game	11	11	11

## Introduction

### UPenn March Madness Study: First Round Predictions

**Start Date:** March 12th, 2017

**End Date:** March 16th, 2017 (prior to the first first-round game of the tournament)

**Number of Predictions in this Survey:** 32

**Expected Completion Time:** 60-90 min.

## Instructions

Hello, and welcome back to the UPenn March Madness Study!

In today's survey, you will be asked to provide predictions about the outcomes of the 32 first-round games of the 2017 NCAA Division I Men's Basketball Tournament (aka, "March Madness").

For each of these games, we will ask you to provide the following information:

1. A prediction about which team will win.
2. A numerical probability judgment that reflects your confidence that the team you selected will win (for more information on probabilistic prediction, please select the appropriate option below).
3. The minimum price at which you would sell a claim ticket for a wager that pays \$15 if the team you selected wins (and \$0 otherwise).
4. Your choice to either keep the lottery ticket or sell it for the price specified above.

After providing these predictions, you will also be asked to complete several tasks related to individual reasoning ability and cognitive style, followed by a brief demographics questionnaire.

**Before we get started, would you like to refresh your memory on any of the following topics?**

(we recommend that you review all of these topics if this is the first UPenn March Madness survey you have completed)

- Study guidelines and payment schedule

- Additional resources related to the NCAA tournament
- Guidelines on how to make probabilistic predictions
- I'm all set. Take me straight to the predictions.

## Study Guidelines and Payment

### Things to Note:

#### General:

- Surveys can be completed at any time before the specified end date, and do not need to be completed in a single sitting.
- When making your predictions, you are welcome to use outside resources! All that we ask is that your answers are your own -- please do not ask any other people to help you with your predictions.

#### Payment:

- Participants will be paid \$10 upon timely completion of this survey ("First Round Predictions") and the other survey sent out on 3/12/17 ("Play-In Predictions").
- Participants will be paid an additional \$10 for providing predictions for all of the remaining games in the tournament.
- Participants who complete this study will also have the opportunity to win up to \$15 more by correctly predicting the outcome of a single game (selected at random).
- The three participants who provide the most accurate probabilistic predictions over the course of the tournament (more on this in a moment) will receive an additional \$50, \$25, and \$10, respectively.
- Participants who do not complete this survey ("First Round Predictions") and the other survey sent out on 3/12/17 ("Play-In Predictions") before the specified end dates (3/16 and 3/14, respectively), **WILL NOT RECEIVE ANY PAYMENT FOR THEIR PARTICIPATION.**

If you have any questions or concerns about today's survey, please contact Josh Baker  
at [jbak@sas.upenn.edu](mailto:jbak@sas.upenn.edu)

- I acknowledge that I have read and understood the information on this page.

## Additional Resources

## Additional Resources

- For any participants who are unfamiliar with the NCAA tournament, additional information about the structure and rules of the tournament can be found [here](#).
- For any participants who would like to see the latest news, information, and schedule for the tournament, all that and more can be found [here](#).
- For any participants who would like to make his or her predictions on paper, a printable version of the NCAA tournament bracket can be found [here](#).

## Prob Tutorial

### How to Provide Probability Estimates

In mathematics, probability is typically defined in terms of relative frequencies. When rolling a die, for example, you can estimate the probability of rolling a "4" by making a large number of rolls and counting the number of times a "4" comes up. Although this process isn't perfect, you will eventually find that the number of "4's" you roll is very close to one-sixth of the total number of rolls. Thus, it makes sense to say that the probability of rolling a "4" is one-out-of-six, or 1/6.

For a single basketball game, however, it doesn't make much sense to think about probability in this way. You can't discover the probability of Team A beating Team B by having them play the same game a large number of times. That one game, by definition, only happens once.

So how can you assign a probability to the outcome of a single game?

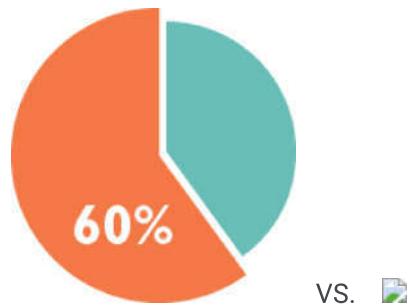
One way to proceed is to think about your preferences for various wagers regarding Team A. Suppose, for example, you were given a simple choice between betting on Team A or betting on a spinner (like a roulette wheel) that had a 50% chance of landing on a winning zone, both for the same prize. Which bet would you choose? If you would choose Team A, then your probability for Team A is higher than 50%.

Now, what if the spinner had a 60% chance of winning? At some point, as the probability of the spinner gets higher, you would switch and bet on the spinner rather than Team A. Just before this point, you would not care whether you bet on the spinner or Team A. Because the prizes for the two bets are the same, the fact that you are indifferent between the two suggests that you also believe the

probabilities are the same. Thus, you can estimate the probability of Team A winning by thinking about the point where your preferences switch.

**Example:**

Which do you prefer? (1) A bet that pays \$15 if the spinner lands on orange; or (2) a bet that pays \$15 if Team A wins a basketball game?



If you prefer option 1 (the spinner), then the probability of Team A winning must be less than 60%. If you prefer option 2 (the basketball game), then the probability of Team A winning must be greater than 60%. If the two options are exactly the same, then the probability of Team A winning must be exactly 60%.

You can use this technique to estimate the probability of Team A winning. Slowly raise the probability of an imaginary spinner until betting on the spinner and betting on the basketball game feel exactly the same. Whatever number your imaginary spinner shows at this point reflects your beliefs about the probability of Team A winning, or your confidence that Team A will win.

---

In this study, we will ask you to predict the winner of each NCAA game, and to make probability judgments that reflect your beliefs about the likelihood that your favored team will win.

In most cases, you will probably have some intuition about which team will win, but could imagine being wrong. In these cases, you should provide a probability judgment that is between 50% and 100%. In cases where you are more confident, your probability judgment should be closer to 100%, and in cases where you are less confident, your probability judgments should be closer to 50%.

Ultimately, your accuracy will be determined by how close your judgments come to the true outcome of each game. These outcomes are coded as "1" if the event happens (e.g., Team A wins) and "0" if it does not (e.g., Team A loses). The further you are from the truth, the more your accuracy score will suffer (and the less likely it is that you will win an Amazon gift card). Thus, it is in your best interest to provide the most accurate probability estimates possible.

It's as easy as that! That's everything you need to know to go make predictions!

**Have fun, and good luck!**

## Predictions

### Instructions:

Please review the following information and provide your prediction about who will win.

---

### Game Information:

$\$Im://Field/1$  vs.  $\$Im://Field/5$   
 $\$Im://Field/9$

$\$Im://Field/1$ Information:	$\$Im://Field/5$ Information:
Regular Season Conference: $\$Im://Field/2$	Regular Season Conference: $\$Im://Field/6$
Division 1 Win-Loss Record: $\$Im://Field/3$	Division 1 Win-Loss Record: $\$Im://Field/7$
Strength of Schedule*: $\$Im://Field/4$	Strength of Schedule* : $\$Im://Field/8$
Tournament Seeding**: $\$Im://Field/10$	Tournament Seeding** : $\$Im://Field/11$

\*Strength of Schedule is a numerical ranking provided by espn.com that reflects the difficulty of a team's regular season schedule. Teams with more difficult schedules (i.e., who play more skilled opponents) are given lower numerical rankings.

\*\* Tournament Seeding is a numerical ranking provided by the NCAA that determines who will play whom in the first round of the tournament. Seeding is done independently in each of the four quadrants of the tournament bracket (corresponding to different geographical regions). Lower (i.e., better) seeding is generally given to teams with better win-loss records and more difficult regular season schedules.

Want to look at the big picture?

[See the whole bracket here](#)

Which team do you think will win this game?

- \${Im://Field/1}
- \${Im://Field/5}

Please provide a probability estimate that reflects your confidence that \${Im://Field/1} will win this game.

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/1} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

Please provide a probability estimate that reflects your confidence that \${Im://Field/5} will win this game.

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/5} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/1} wins, and \$0 if \${Im://Field/1} loses.

If you were forced to sell this ticket, what is the minimum amount you would sell it for?

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/5} wins, and \$0 if \${Im://Field/5} loses.

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**If someone offered to pay the amount above, would you keep the claim ticket or sell it?**

(Please keep in mind: at the end of the tournament, one of your choices will be selected at random and paid-out in real money)

- I would keep the ticket
- I would sell it

### Segue to Indiv. Diffs

#### **That's it for the First Round Games!**

Thank you for completing this portion of the study. Please don't forget to complete the "Play-In Predictions" survey before March 14th!

In the next section of today's survey, you will complete six additional tasks related to individual reasoning ability and cognitive style. **Most of these tasks are quick (2-3 mins.), and you will be able to take breaks in-between each task.**

**If you would like to take a break, now is a good time to do so.**

### CRT

#### **Additional Task \${e://Field/task.num} of 6:**

There will be 18 questions in this section of the study. Please take as much time as you wish to finish these questions. We will be measuring how long you take on each one.

A bat and a ball cost \$1.10 in total. The bat costs a dollar more than the ball. How many cents does the ball cost? (Do not use any decimals or any symbols or letters.)

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

If it takes 5 machines 5 minutes to make 5 widgets, how many minutes would it take 100 machines to make 100 widgets? (Enter the number of minutes)

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how many days would it take for the patch to cover half of the lake? (Enter the number of days)

Have you seen any of the last three questions before?

Yes

No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

All cats are furry. Rabbits are furry. If these two statements are true, can we conclude from them that rabbits are cats?

Yes

No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

All mammals walk. Whales are mammals. If these two statements are true, can we conclude from them that whales walk?

Yes

No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

All fish are swimmers. Some Olympic athletes are swimmers. If these two statements are true, can we conclude from them that some Olympic athletes are fish?

Yes

No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

All flowers have petals. Roses have petals. If these two statements are true, can we conclude from them that roses are flowers?

Yes

No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

All things that have a motor need oil. Automobiles need oil. If these two statements are true, can we conclude from them that automobiles have a motor?

- Yes
- No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

All living things need water. Roses need water. If these two statements are true, can we conclude from them that roses are living things?

- Yes
- No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

All vehicles have wheels. Boats are vehicles. If these two statements are true, can we conclude from them that boats have wheels?

- Yes
- No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

All things that are smoked are good for the health. Cigarettes are smoked. If these two statements are true, can we conclude from them that cigarettes are good for the health?

- Yes  
 No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

All bears are ferocious. Some stuffed animals are bears. If these two statements are true, can we conclude from them that some stuffed animals are ferocious?

- Yes  
 No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

All wives are married. Some women are married. If these two statements are true, can we conclude from them that some women are wives?

- Yes  
 No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

If animals need vitamin Q, can we conclude that oysters need vitamin Q?

- Yes  
 No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds  
Page Submit: 0 seconds  
Click Count: 0 clicks

If oxygen in the air is poisonous to animals, can we conclude that oxygen in the air is poisonous to dogs?

- Yes
- No

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds  
Last Click: 0 seconds  
Page Submit: 0 seconds  
Click Count: 0 clicks

If it takes 2 nurses 2 minutes to measure the blood of 2 patients, how many minutes would it take 200 nurses to measure the blood of 200 patients? (Enter the number of minutes)

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds  
Last Click: 0 seconds  
Page Submit: 0 seconds  
Click Count: 0 clicks

Soup and salad cost \$5.50 in total. The soup costs a dollar more than the salad. How much does the salad cost (in dollars, without a dollar sign)?

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds  
Last Click: 0 seconds  
Page Submit: 0 seconds  
Click Count: 0 clicks

Sally is making sun tea. Every hour, the concentration of the tea doubles. If it takes 6 hours for the tea to be ready, how long would it take for the tea to reach half of the final concentration? (Enter the number of hours)

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

**You're doing great!**

**If you would like to take a break, now is a good time to do so.**

### Berlin Numeracy

#### **Additional Task \${e://Field/task.num} of 6:**

There will be 4 questions in this section of the study. Please answer each question to the best of your ability. Do not use a calculator but feel free to use scratch paper for notes.

Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir?

Please indicate the probability as a decimal between 0 and 1.0

(e.g., 0 = 0% ; 0.5 = 50% ; 1.0 = 100%)

Imagine we are throwing a fair, five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3, or 5)?

Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of 70 throws how many times would the die show the number 6?

In a forest 20% of mushrooms are red, 50% brown and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red?

Please indicate the probability as a decimal between 0 and 1.0  
(e.g., 0 = 0% ; 0.5 = 50% ; 1.0 = 100%)

**You're doing great!**

**If you would like to take a break, now is a good time to do so.**

### **Working Memory**

#### **Additional Task \${e://Field/task.num} of 6:**

In this section of the study, we will test your working memory. You will be presented with 6 items, one at a time, like the example here:

IS  $(2 \times 3) + 7 = 15$ ?

memorize: PHONE

Your task is to answer the math problem (Does  $2 \times 3 + 7 = 15$ ? No) and memorize the indicated word (PHONE). After answering all math questions, you will be asked to recall all of the memorized words in order.

**You will only have 10 seconds to answer each item before the survey advances to the next page.**  
**Please don't worry if you miss some of the questions. This part of the study is intended to be difficult.**

IS  $(5 \times 3) + 4 = 17$ ?

- Yes
- No

memorize: BOOK

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

IS  $(6 \times 2) - 3 = 8$ ?

Yes

No

memorize: HOUSE

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

IS  $(4 \times 4) - 4 = 12$ ?

Yes

No

memorize: JACKET

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

IS  $(3 \times 7) + 6 = 27$ ?

Yes

No

memorize: CAT

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

IS  $(4 \times 8) - 2 = 31$ ?

- Yes
- No

memorize: PEN

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

IS  $(9 \times 2) + 6 = 24$ ?

- Yes
- No

memorize: WATER

**These page timer metrics will not be displayed to the recipient.**

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

Click Count: 0 clicks

Please type the words you were asked to memorize, in order.

- 1
- 2
- 3
- 4
- 5
- 6


You're doing great!

If you would like to take a break, now is a good time to do so.

## AOMT

### Additional Task \${e://Field/task.num} of 6:

Please rate your agreement or disagreement with the following statements.

	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
Intuition is the best guide in making decisions.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
People should take into consideration evidence that goes against their beliefs.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
People should revise their beliefs in response to new information or evidence.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Changing your mind is a sign of weakness.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
People should search actively for reasons why their beliefs might be wrong.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
One should disregard evidence that conflicts with one's established beliefs.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It is more useful to pay attention to those who disagree with us than to pay attention to those who agree.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

persevere in your beliefs even when evidence is brought to bear against them.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	<input type="radio"/>	<input type="radio"/>	Somewhat Agree	<input type="radio"/>	Agree	<input type="radio"/>	Strongly Agree
Allowing oneself to be convinced by an opposing argument is a sign of good character.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>						

You're doing great!

If you would like to take a break, now is a good time to do so.

### Fox Hedgehog

#### Additional Task \${e://Field/task.num} of 6:

In a famous essay, Isaiah Berlin classified thinkers as hedgehogs and foxes: The hedgehog knows one big thing and tries to explain as much as possible using that theory or conceptual framework. The fox knows many small things and is content to improvise explanations on a case-by-case basis.

When it comes to making predictions, would you describe yourself as more of a hedgehog or more of a fox?

- Very Much More Fox-Like
- Somewhat More Fox-Like
- Equally Fox and Hedgehog
- Somewhat More Hedgehog-Like
- Very Much More Hedgehog-Like

You're doing great!

If you would like to take a break, now is a good time to do so.

### Need for Cognition

#### Additional Task \${e://Field/task.num} of 6:

Please rate your agreement or disagreement with the following statements.

The idea of relying on thought to make my

	Strongly Disagree	Disagree	Slightly Disagree	Neither Agree nor Disagree	Slightly Agree	Agree	Strongly Agree
way to the top appeals to me.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I like the responsibility of handling a situation that requires a lot of thinking.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The notion of thinking abstractly is appealing to me.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	Strongly Disagree	Disagree	Slightly Disagree	Neither Agree nor Disagree	Slightly Agree	Agree	Strongly Agree
I only think as hard as I have to.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I find satisfaction in deliberating hard and for long hours.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Learning new ways to think doesn't excite me very much.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I prefer complex to simple problems.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It's enough for me that something gets the job done; I don't care how or why it works.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Thinking is not my idea of fun.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

You're doing great!

If you would like to take a break, now is a good time to do so.

## Demographics

A few last things before you go...

How old are you?

Which option best describes your gender?

Female

Male

Other/Neither

Which option best describes your race/ethnicity?

- |  |  |
|--|--|
| <input type="radio"/> Black/African American | <input type="radio"/> White/Caucasian        |
| <input type="radio"/> Hispanic/Latinx        | <input type="radio"/> Native American        |
| <input type="radio"/> East Asian             | <input type="radio"/> Pacific Islander       |
| <input type="radio"/> South Asian            | <input type="radio"/> Other Indigenous Group |
| <input type="radio"/> Middle Eastern         | <input type="radio"/> Other                  |

Which option best describes your current level of education?

- |  |  |
|--|--|
| <input type="radio"/> Some High School   |  |
| <input type="radio"/> High School, GED, or Equivalent  |  |
| <input type="radio"/> Some undergraduate education (includes current undergraduate students)                       |  |
| <input type="radio"/> Associate's Degree   |  |
| <input type="radio"/> Bachelor's Degree  |  |
| <input type="radio"/> Some post-bachelor's education (includes current graduate students with no graduate degrees) |  |
| <input type="radio"/> Master's Degree  |  |
| <input type="radio"/> PhD or Professional Doctorate (MD, JD, DPT, etc.)  |  |

Are you currently associated with a College or University? If so, which option best describes your affiliation?

- |   |                                     |
|---|-------------------------------------|
| <input type="radio"/> No University Affiliation | <input type="radio"/> Faculty       |
| <input type="radio"/> Undergraduate student     | <input type="radio"/> Staff         |
| <input type="radio"/> Graduate student          | <input type="radio"/> Administrator |

Which College or University are you associated with?

(Please leave blank if you are not affiliated with any College or University)

Please rate your agreement with the following statements:

Neither

	Strongly Disagree	Disagree	Slightly Disagree Slightly Agree	Neither Disagree Agree nor Disagree	Agree nor Disagree	Slightly Agree Slightly Agree	Agree	Agree nor Disagree	Strongly Agree
I have extensive knowledge or expertise in the area of <u>college</u> basketball.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have extensive knowledge of expertise in the area of <u>basketball (in general)</u> .	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have extensive knowledge or expertise in the area of <u>probabilistic prediction</u> .	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have extensive knowledge or expertise in <u>another type of prediction, or prediction in general</u> .	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

To make sure your answers are associated with the right person, please confirm your first and last name below.

First Name

`${m://FirstName}`

Last Name

`${m://LastName}`

### Payment Info

That's it for today's study!

The only thing left is to get you paid.

#### Please read the following information carefully:

By law, the University of Pennsylvania is required to report all payments made to study participants to the Internal Revenue Service (IRS). To facilitate this process, we ask that all participants complete and submit a W9 form (request for taxpayer identification number).

You can find a blank copy of the W9 form [here](#). Once finished, please email a copy of this form to: upenn.march.madness.study@gmail.com.

Without this form, we cannot pay you for your participation.

Please note: all information submitted by participants will be stored on a password protected computer, and will not be shared with anyone except the UPenn business office and the IRS. We will destroy all copies of this information when it is no longer required for the conduct of our research.

**If you have any questions or concerns when completing this form (or, if you are a non-resident alien and do not have a U.S. Social Security Number), please email Josh Baker at [jbak@sas.upenn.edu](mailto:jbak@sas.upenn.edu)**

- I acknowledge that I have read and understood the information on this page

Please specify which payment method you prefer:

- Amazon Gift Card  
 PayPal Transfer (please note: selecting this option may delay your payment by several days)

Please enter the email address to which the Amazon Gift Card should be sent:

Please enter the email address associated with your PayPal account:

(If you do not have a PayPal account, you can create one [here](#) for free)

Powered by Qualtrics

### Loop and Merge Data: First Round (Survey 3)

	<b>Field 1</b>	<b>Field 2</b>	<b>Field 3</b>	<b>Field 4</b>	<b>Field 5</b>	<b>Field 6</b>	<b>Field 7</b>	<b>Field 8</b>	<b>Field 9</b>	<b>Field 10</b>	<b>Field 11</b>
Villanova	Big East	31-3	37 Mt. St. Mary's / N. Orleans	play-in	play-in	Atlantic Coast	22-10	75 First Round	1	16	9
Wisconsin	Big Ten	25-8	83 VA Tech	Colonial	27-5	145 First Round	5	12			
Virginia	Atlantic Coast	22-10	4 UNC Wilmington	Southern	25-7	209 First Round	4	13			
Florida	Southeastern	24-8	8 ETSU	play-in	play-in	Western Athletic	25-5	291 First Round	6	11	
SMU	AAC	29-4	93 Providence / USC	Big East	19-12	59 First Round	7	10			
Baylor	Big 12	24-7	5 New Mexico St.	Sun Belt	20-14	235 First Round	2	15			
So. Carolina	Southeastern	21-10	47 Marquette	Summit League	16-16	152 First Round	1	16			
Duke	Atlantic Coast	27-8	9 Troy	Southeastern	19-15	1 First Round	8	9			
Gonzaga	West Coast	32-1	102 So. Dakota St.	Ivy League	21-6	165 First Round	5	12			
Northwestern	Big Ten	23-11	64 Vanderbilt	Patriot	26-8	213 First Round	4	13			
Notre Dame	Atlantic Coast	25-9	34 Princeton	Big East	21-13	16 First Round	6	11			
W. Virginia	Big 12	26-8	72 Bucknell	ASUN	23-7	269 First Round	3	14			
Maryland	Big Ten	23-8	54 Xavier	Atlantic 10	26-7	56 First Round	7	10			
Florida St.	Atlantic Coast	25-8	20 FGCU	Summit League	19-9	315 First Round	2	15			
St. Mary's	West Coast	28-4	78 VCU	play-in	play-in	Big Ten	19-14	10 First Round	1	16	
Arizona	Pac-12	30-4	24 N. Dakota	Mountain West	28-6	144 First Round	5	12			
Kansas	Big 12	28-4	35 N.C. Central / UC Davis	America East	28-5	222 First Round	4	13			
Miami (Fla.)	Atlantic Coast	21-11	41 Michigan St.	Atlantic 10	23-9	51 First Round	6	11			
Iowa St.	Big 12	23-10	28 Nevada	Metro Atlantic	22-12	174 First Round	3	14			
Purdue	Big Ten	25-7	62 Vermont	Big 12	19-12	15 First Round	7	10			
Creighton	Big East	24-9	50 Rhode Island	OVC	18-14	244 First Round	2	15			
Oregon	Pac-12	28-5	44 Iona	Southwestern	23-11	316 First Round	1	16			
Michigan	Big Ten	23-11	26 Oklahoma St.	Big East	21-11	52 First Round	8	9			
Louisville	Atlantic Coast	24-8	2 Jacksonville St.	Conference USA	29-4	167 First Round	5	12			
N. Carolina	Atlantic Coast	26-7	14 Texas Southern	Big South	24-6	267 First Round	4	13			
Arkansas	Southeastern	25-8	58 Seton Hall	play-in	play-in	Mid-American	6	11			
Minnesota	Big Ten	24-9	17 Middle Tenn.	Missouri Valley	29-4	224 First Round	3	14			
Butler	Big East	23-8	13 Winthrop	Horizon	22-10	225 First Round	2	15			
Cincinnati	AAC	29-4	66 Kans. St. / Wake Forest								
UCLA	Pac-12	29-4	106 Kent St.								
Dayton	Atlantic 10	23-7	71 Wichita St.								
Kentucky	Southeastern	28-5	23 N. Kentucky								

## Title Page

# UPenn March Madness Study: Second Round Predictions: Survey 1 (Thu --> Sat)

**Start Date:** March 16th, 2017

**End Date:** March 18th, 2017 (prior to the first Saturday game)

**Number of Predictions in this Survey:** 4

**Expected Completion Time:** 10-15 min.

## General Instructions

### Instructions

Hello, and welcome back to the UPenn March Madness Study!

In today's survey, you will be asked to provide predictions about four of the second-round games of the 2017 NCAA Division I Men's Basketball Tournament (aka, "March Madness").

For each of these games, we will ask you to provide the following information:

1. A prediction about which team will win.
2. A numerical probability judgment that reflects your confidence that the team you selected will win (for more information on probabilistic prediction, please select the appropriate option below).
3. The minimum price at which you would sell a claim ticket for a wager that pays \$15 if the team you selected wins (and \$0 otherwise).
4. Your choice to either keep the lottery ticket or sell it for the price specified above.

**Before we get started, would you like to refresh your memory on any of the following topics?**

- Study guidelines and payment schedule
- Additional resources related to the NCAA tournament
- Guidelines on how to make probabilistic predictions

I'm all set. Take me straight to the predictions.

## Things to Note

### Things to Note:

#### General:

- Surveys can be completed at any time before the specified end date, and do not need to be completed in a single sitting.
- When making your predictions, you are welcome to use outside resources! All that we ask is that your answers are your own -- please do not ask any other people to help you with your predictions.

#### Payment:

- Participants will be paid \$10 upon timely completion of the First Round Predictions survey and the Play-In Predictions survey (already completed).
- Participants will be paid an additional \$10 for providing predictions for all of the remaining games in the tournament.
- Participants who complete this study will also have the opportunity to win up to \$15 more by correctly predicting the outcome of a single game (selected at random).
- The three participants who provide the most accurate probabilistic predictions over the course of the tournament (more on this in a moment) will receive an additional \$50, \$25, and \$10, respectively.

If you have any questions or concerns about today's survey, please contact Josh Baker at [jbak@sas.upenn.edu](mailto:jbak@sas.upenn.edu)

I acknowledge that I have read and understood the information on this page.

## Additional Resources

### Additional Resources

- For any participants who are unfamiliar with the NCAA tournament, additional information about the structure and rules of the tournament can be found [here](#).
- For any participants who would like to see the latest news, information, and schedule for the tournament, all that and more can be found [here](#).
- For any participants who would like to make his or her predictions on paper, a printable version of the NCAA tournament bracket can be found [here](#).

## **Probability Inst.**

### **How to Provide Probability Estimates**

In mathematics, probability is typically defined in terms of relative frequencies. When rolling a die, for example, you can estimate the probability of rolling a "4" by making a large number of rolls and counting the number of times a "4" comes up. Although this process isn't perfect, you will eventually find that the number of "4's" you roll is very close to one-sixth of the total number of rolls. Thus, it makes sense to say that the probability of rolling a "4" is one-out-of-six, or 1/6.

For a single basketball game, however, it doesn't make much sense to think about probability in this way. You can't discover the probability of Team A beating Team B by having them play the same game a large number of times. That one game, by definition, only happens once.

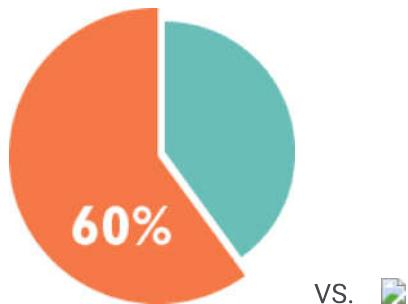
So how can you assign a probability to the outcome of a single game?

One way to proceed is to think about your preferences for various wagers regarding Team A. Suppose, for example, you were given a simple choice between betting on Team A or betting on a spinner (like a roulette wheel) that had a 50% chance of landing on a winning zone, both for the same prize. Which bet would you choose? If you would choose Team A, then your probability for Team A is higher than 50%.

Now, what if the spinner had a 60% chance of winning? At some point, as the probability of the spinner gets higher, you would switch and bet on the spinner rather than Team A. Just before this point, you would not care whether you bet on the spinner or Team A. Because the prizes for the two bets are the same, the fact that you are indifferent between the two suggests that you also believe the probabilities are the same. Thus, you can estimate the probability of Team A winning by thinking about the point where your preferences switch.

**Example:**

Which do you prefer? (1) A bet that pays \$15 if the spinner lands on orange; or (2) a bet that pays \$15 if Team A wins a basketball game?



If you prefer option 1 (the spinner), then the probability of Team A winning must be less than 60%. If you prefer option 2 (the basketball game), then the probability of Team A winning must be greater than 60%. If the two options are exactly the same, then the probability of Team A winning must be exactly 60%.

You can use this technique to estimate the probability of Team A winning. Slowly raise the probability of an imaginary spinner until betting on the spinner and betting on the basketball game feel exactly the same. Whatever number your imaginary spinner shows at this point reflects your beliefs about the probability of Team A winning, or your confidence that Team A will win.

---

In this study, we will ask you to predict the winner of each NCAA game, and to make probability judgments that reflect your beliefs about the likelihood that your favored team will win.

In most cases, you will probably have some intuition about which team will win, but could imagine being wrong. In these cases, you should provide a probability judgment that is between 50% and 100%. In cases where you are more confident, your probability judgment should be closer to 100%, and in cases where you are less confident, your probability judgments should be closer to 50%.

Ultimately, your accuracy will be determined by how close your judgments come to the true outcome of each game. These outcomes are coded as "1" if the event happens (e.g., Team A wins) and "0" if it does not (e.g., Team A loses). The further you are from the truth, the more your accuracy score will suffer (and the less likely it is that you will win an Amazon gift card). Thus, it is in your best interest to provide the most accurate probability estimates possible.

It's as easy as that! That's everything you need to know to go make predictions!

**Have fun, and good luck!**

## Predictions

### Instructions:

Please review the following information and provide your prediction about who will win.

---

### **Game Information:**

\${Im://Field/1} vs. \${Im://Field/5}  
 \${Im://Field/9}

<b><u> \${Im://Field/1} Information:</u></b>	<b><u> \${Im://Field/5} Information:</u></b>
Regular Season Conference: \${Im://Field/2} Division 1 Win-Loss Record: \${Im://Field/3} Strength of Schedule*: \${Im://Field/4} Tournament Seeding**: \${Im://Field/10}	Regular Season Conference: \${Im://Field/6} Division 1 Win-Loss Record: \${Im://Field/7} Strength of Schedule* : \${Im://Field/8} Tournament Seeding** : \${Im://Field/11}

\*Strength of Schedule is a numerical ranking provided by espn.com that reflects the difficulty of a team's regular season schedule. Teams with more difficult schedules (i.e., who play more skilled opponents) are given lower numerical rankings.

\*\* Tournament Seeding is a numerical ranking provided by the NCAA that determines who will play whom in the first round of the tournament. Seeding is done independently in each of the four quadrants of the tournament bracket (corresponding to different geographical regions). Lower (i.e., better) seeding is generally given to teams with better win-loss records and more difficult regular season schedules.

Want to look at the big picture?

[See the whole bracket here](#)

**Which team do you think will win this game?**

- \${Im://Field/1}
- \${Im://Field/5}

**Please provide a probability estimate that reflects your confidence that \${Im://Field/1} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/1} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Please provide a probability estimate that reflects your confidence that \${Im://Field/5} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/5} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/1} wins, and \$0 if \${Im://Field/1} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/5} wins, and \$0 if \${Im://Field/5} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**If someone offered to pay the amount above, would you keep the claim ticket or sell it?**

(Please keep in mind: at the end of the tournament, one of your choices will be selected at random and paid-out in real money)

- I would keep the ticket
- I would sell it

### Email

**That's it for today's survey!**

Thank you for completing this portion of the study. Please don't forget the two other second-round surveys, coming your way in the next 24-hours!

To ensure that your answers are associated with the right person, please confirm your contact information below.

First Name

\${m://FirstName}

Last Name

\${m://LastName}

Powered by Qualtrics

### Loop and Merge Data: Second Round A (Survey 4)

	<b>Field 1</b>	<b>Field 2</b>	<b>Field 3</b>	<b>Field 4</b>	<b>Field 5</b>	<b>Field 6</b>	<b>Field 7</b>	<b>Field 8</b>	<b>Field 9</b>	<b>Field 10</b>	<b>Field 11</b>
Middle Tenn.	Conference USA	29-4	167	Butler	Big East	23-8	13	Second Round	12	4	
Northwestern	Big Ten	22-11	64	Gonzaga	West Coast	32-1	102	Second Round	8	1	
Virginia	Atlantic Coast	22-10	4	Florida	Southeastern	24-8	8	Second Round	5	4	
Notre Dame	Atlantic Coast	25-9	34	W. Virginia	Big 12	26-8	72	Second Round	5	4	

## Title Page

# UPenn March Madness Study: Second Round Predictions: Survey 2 (Fri --> Sat)

**Start Date:** March 17th, 2017

**End Date:** March 18th, 2017 (prior to the first Saturday game)

**Number of Predictions in this Survey:** 4

**Expected Completion Time:** 10-15 min.

## General Instructions

### Instructions

Hello, and welcome back to the UPenn March Madness Study!

In today's survey, you will be asked to provide predictions about four of the second-round games of the 2017 NCAA Division I Men's Basketball Tournament (aka, "March Madness").

For each of these games, we will ask you to provide the following information:

1. A prediction about which team will win.
2. A numerical probability judgment that reflects your confidence that the team you selected will win (for more information on probabilistic prediction, please select the appropriate option below).
3. The minimum price at which you would sell a claim ticket for a wager that pays \$15 if the team you selected wins (and \$0 otherwise).
4. Your choice to either keep the lottery ticket or sell it for the price specified above.

**Before we get started, would you like to refresh your memory on any of the following topics?**

- Study guidelines and payment schedule
- Additional resources related to the NCAA tournament
- Guidelines on how to make probabilistic predictions

I'm all set. Take me straight to the predictions.

## Things to Note

### Things to Note:

#### General:

- Surveys can be completed at any time before the specified end date, and do not need to be completed in a single sitting.
- When making your predictions, you are welcome to use outside resources! All that we ask is that your answers are your own – please do not ask any other people to help you with your predictions.

#### Payment:

- Participants will be paid \$10 upon timely completion of the First Round Predictions survey and the Play-In Predictions survey (already completed).
- Participants will be paid an additional \$10 for providing predictions for all of the remaining games in the tournament.
- Participants who complete this study will also have the opportunity to win up to \$15 more by correctly predicting the outcome of a single game (selected at random).
- The three participants who provide the most accurate probabilistic predictions over the course of the tournament (more on this in a moment) will receive an additional \$50, \$25, and \$10, respectively.

If you have any questions or concerns about today's survey, please contact Josh Baker  
at [jbak@sas.upenn.edu](mailto:jbak@sas.upenn.edu)

I acknowledge that I have read and understood the information on this page.

## Additional Resources

### Additional Resources

- For any participants who are unfamiliar with the NCAA tournament, additional information about the structure and rules of the tournament can be found [here](#).
- For any participants who would like to see the latest news, information, and schedule for the tournament, all that and more can be found [here](#).
- For any participants who would like to make his or her predictions on paper, a printable version of the NCAA tournament bracket can be found [here](#).

## Probability Inst.

### How to Provide Probability Estimates

In mathematics, probability is typically defined in terms of relative frequencies. When rolling a die, for example, you can estimate the probability of rolling a "4" by making a large number of rolls and counting the number of times a "4" comes up. Although this process isn't perfect, you will eventually find that the number of "4's" you roll is very close to one-sixth of the total number of rolls. Thus, it makes sense to say that the probability of rolling a "4" is one-out-of-six, or 1/6.

For a single basketball game, however, it doesn't make much sense to think about probability in this way. You can't discover the probability of Team A beating Team B by having them play the same game a large number of times. That one game, by definition, only happens once.

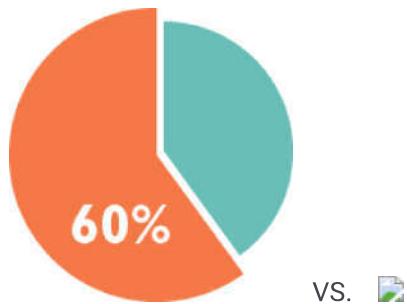
So how can you assign a probability to the outcome of a single game?

One way to proceed is to think about your preferences for various wagers regarding Team A. Suppose, for example, you were given a simple choice between betting on Team A or betting on a spinner (like a roulette wheel) that had a 50% chance of landing on a winning zone, both for the same prize. Which bet would you choose? If you would choose Team A, then your probability for Team A is higher than 50%.

Now, what if the spinner had a 60% chance of winning? At some point, as the probability of the spinner gets higher, you would switch and bet on the spinner rather than Team A. Just before this point, you would not care whether you bet on the spinner or Team A. Because the prizes for the two bets are the same, the fact that you are indifferent between the two suggests that you also believe the probabilities are the same. Thus, you can estimate the probability of Team A winning by thinking about the point where your preferences switch.

**Example:**

Which do you prefer? (1) A bet that pays \$15 if the spinner lands on orange; or (2) a bet that pays \$15 if Team A wins a basketball game?



If you prefer option 1 (the spinner), then the probability of Team A winning must be less than 60%. If you prefer option 2 (the basketball game), then the probability of Team A winning must be greater than 60%. If the two options are exactly the same, then the probability of Team A winning must be exactly 60%.

You can use this technique to estimate the probability of Team A winning. Slowly raise the probability of an imaginary spinner until betting on the spinner and betting on the basketball game feel exactly the same. Whatever number your imaginary spinner shows at this point reflects your beliefs about the probability of Team A winning, or your confidence that Team A will win.

---

In this study, we will ask you to predict the winner of each NCAA game, and to make probability judgments that reflect your beliefs about the likelihood that your favored team will win.

In most cases, you will probably have some intuition about which team will win, but could imagine being wrong. In these cases, you should provide a probability judgment that is between 50% and 100%. In cases where you are more confident, your probability judgment should be closer to 100%, and in cases where you are less confident, your probability judgments should be closer to 50%.

Ultimately, your accuracy will be determined by how close your judgments come to the true outcome of each game. These outcomes are coded as "1" if the event happens (e.g., Team A wins) and "0" if it does not (e.g., Team A loses). The further you are from the truth, the more your accuracy score will suffer (and the less likely it is that you will win an Amazon gift card). Thus, it is in your best interest to provide the most accurate probability estimates possible.

It's as easy as that! That's everything you need to know to go make predictions!

**Have fun, and good luck!**

## Predictions

### Instructions:

Please review the following information and provide your prediction about who will win.

---

### Game Information:

`${Im://Field/1} vs. ${Im://Field/5}`  
 `${Im://Field/9}`

<code> \${Im://Field/1} Information:</code>	<code> \${Im://Field/5} Information:</code>
Regular Season Conference: <code> \${Im://Field/2}</code> Division 1 Win-Loss Record: <code> \${Im://Field/3}</code> Strength of Schedule*: <code> \${Im://Field/4}</code> Tournament Seeding**: <code> \${Im://Field/10}</code>	Regular Season Conference: <code> \${Im://Field/6}</code> Division 1 Win-Loss Record: <code> \${Im://Field/7}</code> Strength of Schedule* : <code> \${Im://Field/8}</code> Tournament Seeding** : <code> \${Im://Field/11}</code>

\*Strength of Schedule is a numerical ranking provided by espn.com that reflects the difficulty of a team's regular season schedule. Teams with more difficult schedules (i.e., who play more skilled opponents) are given lower numerical rankings.

\*\* Tournament Seeding is a numerical ranking provided by the NCAA that determines who will play whom in the first round of the tournament. Seeding is done independently in each of the four quadrants of the tournament bracket (corresponding to different geographical regions). Lower (i.e., better) seeding is generally given to teams with better win-loss records and more difficult regular season schedules.

Want to look at the big picture?

[See the whole bracket here](#)

**Which team do you think will win this game?**

- `${Im://Field/1}`
- `${Im://Field/5}`

**Please provide a probability estimate that reflects your confidence that \${Im://Field/1} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/1} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Please provide a probability estimate that reflects your confidence that \${Im://Field/5} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/5} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/1} wins, and \$0 if \${Im://Field/1} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/5} wins, and \$0 if \${Im://Field/5} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**If someone offered to pay the amount above, would you keep the claim ticket or sell it?**

(Please keep in mind: at the end of the tournament, one of your choices will be selected at random and paid-out in real money)

- I would keep the ticket
- I would sell it

**Email**

**That's it for today's survey!**

Thank you for completing this portion of the study. **Please don't forget to complete the remainder of the second-round surveys! They're all quick, and they're all due soon!!**

**Survey 1:**

4 other predictions for Saturday's games (already sent out), due Saturday (3/18) at noon.

**Survey 2:**

This is the survey you just completed!

**Survey 3:**

8 predictions for Sunday's games (will be sent out ~midnight Friday), due Sunday (3/19) at noon.

To ensure that your answers are associated with the right person, please confirm your contact information below.

First Name

\${m://FirstName}

Last Name

\${m://LastName}

Powered by Qualtrics

### Loop and Merge Data: Second Round B (Survey 5)

	<b>Field 1</b>	<b>Field 2</b>	<b>Field 3</b>	<b>Field 4</b>	<b>Field 5</b>	<b>Field 6</b>	<b>Field 7</b>	<b>Field 8</b>	<b>Field 9</b>	<b>Field 10</b>	<b>Field 11</b>
Wisconsin	Big Ten	25-8	83	Villanova	Big East	31-3	37	Second Round	8	1	
Xavier	Big East	21-13	16	Florida St.	Atlantic Coast	25-8	20	Second Round	11	3	
St. Mary's	West Coast	28-4	78	Arizona	Pac-12	30-4	24	Second Round	7	2	
Iowa St.	Big 12	23-10	28	Purdue	Big Ten	25-7	62	Second Round	5	4	

## Title Page

# UPenn March Madness Study: Second Round Predictions: Survey 3 (Thu & Fri --> Sun)

**Start Date:** March 17th, 2017

**End Date:** March 19th, 2017 (prior to the first Sunday game)

**Number of Predictions in this Survey:** 8

**Expected Completion Time:** 20-30 min.

## General Instructions

### Instructions

Hello, and welcome back to the UPenn March Madness Study!

In today's survey, you will be asked to provide predictions about eight of the second-round games of the 2017 NCAA Division I Men's Basketball Tournament (aka, "March Madness").

For each of these games, we will ask you to provide the following information:

1. A prediction about which team will win.
2. A numerical probability judgment that reflects your confidence that the team you selected will win (for more information on probabilistic prediction, please select the appropriate option below).
3. The minimum price at which you would sell a claim ticket for a wager that pays \$15 if the team you selected wins (and \$0 otherwise).
4. Your choice to either keep the claim ticket or sell it for the price specified above.

**Before we get started, would you like to refresh your memory on any of the following topics?**

- Study guidelines and payment schedule
- Additional resources related to the NCAA tournament
- Guidelines on how to make probabilistic predictions

I'm all set. Take me straight to the predictions.

## Things to Note

### Things to Note:

#### General:

- Surveys can be completed at any time before the specified end date, and do not need to be completed in a single sitting.
- When making your predictions, you are welcome to use outside resources! All that we ask is that your answers are your own – please do not ask any other people to help you with your predictions.

#### Payment:

- Participants will be paid \$10 upon timely completion of the First Round Predictions survey and the Play-In Predictions survey (already completed).
- Participants will be paid an additional \$10 for providing predictions for all of the remaining games in the tournament.
- Participants who complete this study will also have the opportunity to win up to \$15 more by correctly predicting the outcome of a single game (selected at random).
- The three participants who provide the most accurate probabilistic predictions over the course of the tournament (more on this in a moment) will receive an additional \$50, \$25, and \$10, respectively.

If you have any questions or concerns about today's survey, please contact Josh Baker  
at [jbak@sas.upenn.edu](mailto:jbak@sas.upenn.edu)

I acknowledge that I have read and understood the information on this page.

## Additional Resources

### Additional Resources

- For any participants who are unfamiliar with the NCAA tournament, additional information about the structure and rules of the tournament can be found [here](#).
- For any participants who would like to see the latest news, information, and schedule for the tournament, all that and more can be found [here](#).
- For any participants who would like to make his or her predictions on paper, a printable version of the NCAA tournament bracket can be found [here](#).

## Probability Inst.

### How to Provide Probability Estimates

In mathematics, probability is typically defined in terms of relative frequencies. When rolling a die, for example, you can estimate the probability of rolling a "4" by making a large number of rolls and counting the number of times a "4" comes up. Although this process isn't perfect, you will eventually find that the number of "4's" you roll is very close to one-sixth of the total number of rolls. Thus, it makes sense to say that the probability of rolling a "4" is one-out-of-six, or 1/6.

For a single basketball game, however, it doesn't make much sense to think about probability in this way. You can't discover the probability of Team A beating Team B by having them play the same game a large number of times. That one game, by definition, only happens once.

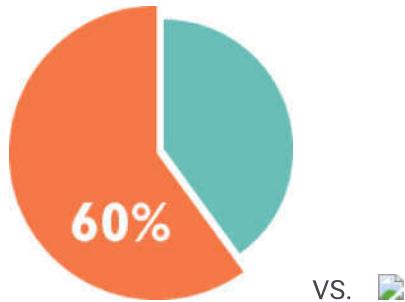
So how can you assign a probability to the outcome of a single game?

One way to proceed is to think about your preferences for various wagers regarding Team A. Suppose, for example, you were given a simple choice between betting on Team A or betting on a spinner (like a roulette wheel) that had a 50% chance of landing on a winning zone, both for the same prize. Which bet would you choose? If you would choose Team A, then your probability for Team A is higher than 50%.

Now, what if the spinner had a 60% chance of winning? At some point, as the probability of the spinner gets higher, you would switch and bet on the spinner rather than Team A. Just before this point, you would not care whether you bet on the spinner or Team A. Because the prizes for the two bets are the same, the fact that you are indifferent between the two suggests that you also believe the probabilities are the same. Thus, you can estimate the probability of Team A winning by thinking about the point where your preferences switch.

**Example:**

Which do you prefer? (1) A bet that pays \$15 if the spinner lands on orange; or (2) a bet that pays \$15 if Team A wins a basketball game?



If you prefer option 1 (the spinner), then the probability of Team A winning must be less than 60%. If you prefer option 2 (the basketball game), then the probability of Team A winning must be greater than 60%. If the two options are exactly the same, then the probability of Team A winning must be exactly 60%.

You can use this technique to estimate the probability of Team A winning. Slowly raise the probability of an imaginary spinner until betting on the spinner and betting on the basketball game feel exactly the same. Whatever number your imaginary spinner shows at this point reflects your beliefs about the probability of Team A winning, or your confidence that Team A will win.

---

In this study, we will ask you to predict the winner of each NCAA game, and to make probability judgments that reflect your beliefs about the likelihood that your favored team will win.

In most cases, you will probably have some intuition about which team will win, but could imagine being wrong. In these cases, you should provide a probability judgment that is between 50% and 100%. In cases where you are more confident, your probability judgment should be closer to 100%, and in cases where you are less confident, your probability judgments should be closer to 50%.

Ultimately, your accuracy will be determined by how close your judgments come to the true outcome of each game. These outcomes are coded as "1" if the event happens (e.g., Team A wins) and "0" if it does not (e.g., Team A loses). The further you are from the truth, the more your accuracy score will suffer (and the less likely it is that you will win an Amazon gift card). Thus, it is in your best interest to provide the most accurate probability estimates possible.

It's as easy as that! That's everything you need to know to go make predictions!

**Have fun, and good luck!**

## Predictions

### Instructions:

Please review the following information and provide your prediction about who will win.

---

### Game Information:

`${Im://Field/1} vs. ${Im://Field/5}`  
 `${Im://Field/9}`

<code> \${Im://Field/1} Information:</code>	<code> \${Im://Field/5} Information:</code>
Regular Season Conference: <code> \${Im://Field/2}</code> Division 1 Win-Loss Record: <code> \${Im://Field/3}</code> Strength of Schedule*: <code> \${Im://Field/4}</code> Tournament Seeding**: <code> \${Im://Field/10}</code>	Regular Season Conference: <code> \${Im://Field/6}</code> Division 1 Win-Loss Record: <code> \${Im://Field/7}</code> Strength of Schedule* : <code> \${Im://Field/8}</code> Tournament Seeding** : <code> \${Im://Field/11}</code>

\*Strength of Schedule is a numerical ranking provided by espn.com that reflects the difficulty of a team's regular season schedule. Teams with more difficult schedules (i.e., who play more skilled opponents) are given lower numerical rankings.

\*\* Tournament Seeding is a numerical ranking provided by the NCAA that determines who will play whom in the first round of the tournament. Seeding is done independently in each of the four quadrants of the tournament bracket (corresponding to different geographical regions). Lower (i.e., better) seeding is generally given to teams with better win-loss records and more difficult regular season schedules.

Want to look at the big picture?

[See the whole bracket here](#)

**Which team do you think will win this game?**

- `${Im://Field/1}`
- `${Im://Field/5}`

**Please provide a probability estimate that reflects your confidence that \${Im://Field/1} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/1} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Please provide a probability estimate that reflects your confidence that \${Im://Field/5} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/5} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/1} wins, and \$0 if \${Im://Field/1} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/5} wins, and \$0 if \${Im://Field/5} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**If someone offered to pay the amount above, would you keep the claim ticket or sell it?**

(Please keep in mind: at the end of the tournament, one of your choices will be selected at random and paid-out in real money)

- I would keep the ticket
- I would sell it

### Email

**That's it for today's survey!**

Thank you for completing this portion of the study. **Please don't forget to complete the remainder of the second-round surveys! They're all quick, and they're all due soon!!**

**Survey 1:**

4 predictions for Saturday's games (already sent out), due Saturday (3/18) at noon.  
Thursday first round --> Saturday second round

**Survey 2:**

4 additional predictions for Saturday's games (already sent out), due Saturday (3/18) at noon.  
Friday first round --> Saturday second round

**Survey 3:**

This is the survey you just completed!  
Thursday & Friday first round --> Sunday second round

To ensure that your answers are associated with the right person, please confirm your contact information below.

First Name

\${m://FirstName}

Last Name

\${m://LastName}

### Loop and Merge Data: Second Round C (Survey 6)

	<b>Field 1</b>	<b>Field 2</b>	<b>Field 3</b>	<b>Field 4</b>	<b>Field 5</b>	<b>Field 6</b>	<b>Field 7</b>	<b>Field 8</b>	<b>Field 9</b>	<b>Field 10</b>	<b>Field 11</b>
Rhode Island	Atlantic 10	23-9	51 Oregon	Pac-12	28-5	44	Second Round	11			3
Michigan	Big Ten	23-11	26 Louisville	Atlantic Coast	24-8	2	Second Round	7			2
Michigan St.	Big Ten	19-14	10 Kansas	Big 12	28-4	35	Second Round	9			1
USC	Pac-12	24-9	76 Baylor	Big 12	24-7	5	Second Round	11			3
Arkansas	Southeastern	25-8	58 N. Carolina	Atlantic Coast	26-7	14	Second Round	8			1
Duke	Atlantic Coast	27-8	9 So. Carolina	Southeastern	21-10	47	Second Round	2			7
Kentucky	Southeastern	28-5	23 Wichita St.	Missouri Valley	29-4	186	Second Round	2			10
Cincinnati	AAC	29-4	66 UCLA	Pac-12	29-4	106	Second Round	6			3

## Title Page

# UPenn March Madness Study: Sweet Sixteen Predictions

**Start Date:** March 20th, 2017

**End Date:** March 23rd, 2017, (prior to the first sweet-sixteen game at 7pm)

**Number of Predictions in this Survey:** 8

**Expected Completion Time:** 20-30 min.

## General Instructions

### Instructions

Hello, and welcome back to the UPenn March Madness Study!

In today's survey, you will be asked to provide predictions about the eight "Sweet Sixteen" games of the 2017 NCAA Division I Men's Basketball Tournament (aka, "March Madness").

For each of these games, we will ask you to provide the following information:

1. A prediction about which team will win.
2. A numerical probability judgment that reflects your confidence that the team you selected will win (for more information on probabilistic prediction, please select the appropriate option below).
3. The minimum price at which you would sell a claim ticket for a wager that pays \$15 if the team you selected wins (and \$0 otherwise).
4. Your choice to either keep the claim ticket or sell it for the price specified above.

**Before we get started, would you like to refresh your memory on any of the following topics?**

- Study guidelines and payment schedule
- Additional resources related to the NCAA tournament
- Guidelines on how to make probabilistic predictions

I'm all set. Take me straight to the predictions.

## Things to Note

### Things to Note:

#### General:

- Surveys can be completed at any time before the specified end date, and do not need to be completed in a single sitting.
- When making your predictions, you are welcome to use outside resources! All that we ask is that your answers are your own -- please do not ask any other people to help you with your predictions.

#### Payment:

- Participants will be paid \$10 upon timely completion of the First Round Predictions survey and the Play-In Predictions survey (already completed).
- Participants will be paid an additional \$10 for providing predictions for all of the remaining games in the tournament.
- Participants who complete this study will also have the opportunity to win up to \$15 more by correctly predicting the outcome of a single game (selected at random).
- The three participants who provide the most accurate probabilistic predictions over the course of the tournament (more on this in a moment) will receive an additional \$50, \$25, and \$10, respectively.

If you have any questions or concerns about today's survey, please contact Josh Baker  
at [jbak@sas.upenn.edu](mailto:jbak@sas.upenn.edu)

I acknowledge that I have read and understood the information on this page.

## Additional Resources

### Additional Resources

- For any participants who are unfamiliar with the NCAA tournament, additional information about the structure and rules of the tournament can be found [here](#).
- For any participants who would like to see the latest news, information, and schedule for the tournament, all that and more can be found [here](#).
- For any participants who would like to make his or her predictions on paper, a printable version of the NCAA tournament bracket can be found [here](#).

## **Probability Inst.**

### **How to Provide Probability Estimates**

In mathematics, probability is typically defined in terms of relative frequencies. When rolling a die, for example, you can estimate the probability of rolling a "4" by making a large number of rolls and counting the number of times a "4" comes up. Although this process isn't perfect, you will eventually find that the number of "4's" you roll is very close to one-sixth of the total number of rolls. Thus, it makes sense to say that the probability of rolling a "4" is one-out-of-six, or 1/6.

For a single basketball game, however, it doesn't make much sense to think about probability in this way. You can't discover the probability of Team A beating Team B by having them play the same game a large number of times. That one game, by definition, only happens once.

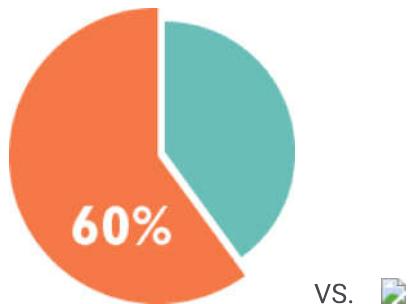
So how can you assign a probability to the outcome of a single game?

One way to proceed is to think about your preferences for various wagers regarding Team A. Suppose, for example, you were given a simple choice between betting on Team A or betting on a spinner (like a roulette wheel) that had a 50% chance of landing on a winning zone, both for the same prize. Which bet would you choose? If you would choose Team A, then your probability for Team A is higher than 50%.

Now, what if the spinner had a 60% chance of winning? At some point, as the probability of the spinner gets higher, you would switch and bet on the spinner rather than Team A. Just before this point, you would not care whether you bet on the spinner or Team A. Because the prizes for the two bets are the same, the fact that you are indifferent between the two suggests that you also believe the probabilities are the same. Thus, you can estimate the probability of Team A winning by thinking about the point where your preferences switch.

**Example:**

Which do you prefer? (1) A bet that pays \$15 if the spinner lands on orange; or (2) a bet that pays \$15 if Team A wins a basketball game?



If you prefer option 1 (the spinner), then the probability of Team A winning must be less than 60%. If you prefer option 2 (the basketball game), then the probability of Team A winning must be greater than 60%. If the two options are exactly the same, then the probability of Team A winning must be exactly 60%.

You can use this technique to estimate the probability of Team A winning. Slowly raise the probability of an imaginary spinner until betting on the spinner and betting on the basketball game feel exactly the same. Whatever number your imaginary spinner shows at this point reflects your beliefs about the probability of Team A winning, or your confidence that Team A will win.

---

In this study, we will ask you to predict the winner of each NCAA game, and to make probability judgments that reflect your beliefs about the likelihood that your favored team will win.

In most cases, you will probably have some intuition about which team will win, but could imagine being wrong. In these cases, you should provide a probability judgment that is between 50% and 100%. In cases where you are more confident, your probability judgment should be closer to 100%, and in cases where you are less confident, your probability judgments should be closer to 50%.

Ultimately, your accuracy will be determined by how close your judgments come to the true outcome of each game. These outcomes are coded as "1" if the event happens (e.g., Team A wins) and "0" if it does not (e.g., Team A loses). The further you are from the truth, the more your accuracy score will suffer (and the less likely it is that you will win an Amazon gift card). Thus, it is in your best interest to provide the most accurate probability estimates possible.

It's as easy as that! That's everything you need to know to go make predictions!

**Have fun, and good luck!**

## Predictions

### Instructions:

Please review the following information and provide your prediction about who will win.

---

### **Game Information:**

\${Im://Field/1} vs. \${Im://Field/5}  
 \${Im://Field/9}

<b><u> \${Im://Field/1} Information:</u></b>	<b><u> \${Im://Field/5} Information:</u></b>
Regular Season Conference: \${Im://Field/2} Division 1 Win-Loss Record: \${Im://Field/3} Strength of Schedule*: \${Im://Field/4} Tournament Seeding**: \${Im://Field/10}	Regular Season Conference: \${Im://Field/6} Division 1 Win-Loss Record: \${Im://Field/7} Strength of Schedule* : \${Im://Field/8} Tournament Seeding** : \${Im://Field/11}

\*Strength of Schedule is a numerical ranking provided by espn.com that reflects the difficulty of a team's regular season schedule. Teams with more difficult schedules (i.e., who play more skilled opponents) are given lower numerical rankings.

\*\* Tournament Seeding is a numerical ranking provided by the NCAA that determines who will play whom in the first round of the tournament. Seeding is done independently in each of the four quadrants of the tournament bracket (corresponding to different geographical regions). Lower (i.e., better) seeding is generally given to teams with better win-loss records and more difficult regular season schedules.

Want to look at the big picture?

[See the whole bracket here](#)

**Which team do you think will win this game?**

- \${Im://Field/1}
- \${Im://Field/5}

**Please provide a probability estimate that reflects your confidence that \${Im://Field/1} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/1} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Please provide a probability estimate that reflects your confidence that \${Im://Field/5} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/5} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/1} wins, and \$0 if \${Im://Field/1} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/5} wins, and \$0 if \${Im://Field/5} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

If someone offered to pay the amount above, would you keep the claim ticket or sell it?

(Please keep in mind: at the end of the tournament, one of your choices will be selected at random and paid-out in real money)

- I would keep the ticket
- I would sell it

### Email

**That's it for today's survey!**

Thank you for completing this portion of the study.

**The next round of predictions (for games among the "elite eight") will be sent out at ~midnight on Friday (3/24), and will be due before the first game the following day (likely in the evening). This survey will comprise 4 predictions in total.**

To ensure that your answers are associated with the right person, please confirm your contact information below.

First Name

\${m://FirstName}

Last Name

\${m://LastName}

Powered by Qualtrics

### Loop and Merge Data: Sweet Sixteen (Survey 7)

	<b>Field 1</b>	<b>Field 2</b>	<b>Field 3</b>	<b>Field 4</b>	<b>Field 5</b>	<b>Field 6</b>	<b>Field 7</b>	<b>Field 8</b>	<b>Field 9</b>	<b>Field 10</b>	<b>Field 11</b>
Butler	Big East	23-8	13 N. Carolina	Atlantic Coast	26-7	14 Sweet Sixteen				4	1
UCLA	Pac-12	29-4	106 Kentucky	Southeastern	28-5	23 Sweet Sixteen				3	2
So. Carolina	Southeastern	21-10	47 Baylor	Big 12	24-7	5 Sweet Sixteen				7	3
Wisconsin	Big Ten	25-8	83 Florida	Southeastern	24-8	8 Sweet Sixteen				8	4
Michigan	Big Ten	23-11	26 Oregon	Pac-12	28-5	44 Sweet Sixteen				7	3
Purdue	Big Ten	25-7	62 Kansas	Big 12	28-4	35 Sweet Sixteen				4	1
W. Virginia	Big 12	26-8	72 Gonzaga	West Coast	32-1	102 Sweet Sixteen				4	1
Xavier	Big East	21-13	16 Arizona	Pac-12	30-4	24 Sweet Sixteen				11	2

## Title Page

# UPenn March Madness Study: Elite Eight Predictions

**Start Date:** March 24th, 2017

**End Date:** March 25th, 2017, (Saturday, prior to the first elite eight game at 6pm)

**Number of Predictions in this Survey:** 4

**Expected Completion Time:** 10-15 min.

## General Instructions

### Instructions

Hello, and welcome back to the UPenn March Madness Study!

In today's survey, you will be asked to provide predictions about the four "Elite Eight" games of the 2017 NCAA Division I Men's Basketball Tournament (aka, "March Madness").

For each of these games, we will ask you to provide the following information:

1. A prediction about which team will win.
2. A numerical probability judgment that reflects your confidence that the team you selected will win (for more information on probabilistic prediction, please select the appropriate option below).
3. The minimum price at which you would sell a claim ticket for a wager that pays \$15 if the team you selected wins (and \$0 otherwise).
4. Your choice to either keep the claim ticket or sell it for the price specified above.

**Before we get started, would you like to refresh your memory on any of the following topics?**

- Study guidelines and payment schedule
- Additional resources related to the NCAA tournament
- Guidelines on how to make probabilistic predictions

I'm all set. Take me straight to the predictions.

## Things to Note

### Things to Note:

#### General:

- Surveys can be completed at any time before the specified end date, and do not need to be completed in a single sitting.
- When making your predictions, you are welcome to use outside resources! All that we ask is that your answers are your own -- please do not ask any other people to help you with your predictions.

#### Payment:

- Participants will be paid \$10 upon timely completion of the First Round Predictions survey and the Play-In Predictions survey (already completed).
- Participants will be paid an additional \$10 for providing predictions for all of the remaining games in the tournament.
- Participants who complete this study will also have the opportunity to win up to \$15 more by correctly predicting the outcome of a single game (selected at random).
- The three participants who provide the most accurate probabilistic predictions over the course of the tournament (more on this in a moment) will receive an additional \$50, \$25, and \$10, respectively.

If you have any questions or concerns about today's survey, please contact Josh Baker  
at [jbak@sas.upenn.edu](mailto:jbak@sas.upenn.edu)

I acknowledge that I have read and understood the information on this page.

## Additional Resources

### Additional Resources

- For any participants who are unfamiliar with the NCAA tournament, additional information about the structure and rules of the tournament can be found [here](#).
- For any participants who would like to see the latest news, information, and schedule for the tournament, all that and more can be found [here](#).
- For any participants who would like to make his or her predictions on paper, a printable version of the NCAA tournament bracket can be found [here](#).

## **Probability Inst.**

### **How to Provide Probability Estimates**

In mathematics, probability is typically defined in terms of relative frequencies. When rolling a die, for example, you can estimate the probability of rolling a "4" by making a large number of rolls and counting the number of times a "4" comes up. Although this process isn't perfect, you will eventually find that the number of "4's" you roll is very close to one-sixth of the total number of rolls. Thus, it makes sense to say that the probability of rolling a "4" is one-out-of-six, or 1/6.

For a single basketball game, however, it doesn't make much sense to think about probability in this way. You can't discover the probability of Team A beating Team B by having them play the same game a large number of times. That one game, by definition, only happens once.

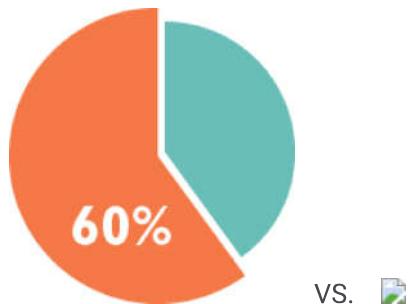
So how can you assign a probability to the outcome of a single game?

One way to proceed is to think about your preferences for various wagers regarding Team A. Suppose, for example, you were given a simple choice between betting on Team A or betting on a spinner (like a roulette wheel) that had a 50% chance of landing on a winning zone, both for the same prize. Which bet would you choose? If you would choose Team A, then your probability for Team A is higher than 50%.

Now, what if the spinner had a 60% chance of winning? At some point, as the probability of the spinner gets higher, you would switch and bet on the spinner rather than Team A. Just before this point, you would not care whether you bet on the spinner or Team A. Because the prizes for the two bets are the same, the fact that you are indifferent between the two suggests that you also believe the probabilities are the same. Thus, you can estimate the probability of Team A winning by thinking about the point where your preferences switch.

**Example:**

Which do you prefer? (1) A bet that pays \$15 if the spinner lands on orange; or (2) a bet that pays \$15 if Team A wins a basketball game?



If you prefer option 1 (the spinner), then the probability of Team A winning must be less than 60%. If you prefer option 2 (the basketball game), then the probability of Team A winning must be greater than 60%. If the two options are exactly the same, then the probability of Team A winning must be exactly 60%.

You can use this technique to estimate the probability of Team A winning. Slowly raise the probability of an imaginary spinner until betting on the spinner and betting on the basketball game feel exactly the same. Whatever number your imaginary spinner shows at this point reflects your beliefs about the probability of Team A winning, or your confidence that Team A will win.

---

In this study, we will ask you to predict the winner of each NCAA game, and to make probability judgments that reflect your beliefs about the likelihood that your favored team will win.

In most cases, you will probably have some intuition about which team will win, but could imagine being wrong. In these cases, you should provide a probability judgment that is between 50% and 100%. In cases where you are more confident, your probability judgment should be closer to 100%, and in cases where you are less confident, your probability judgments should be closer to 50%.

Ultimately, your accuracy will be determined by how close your judgments come to the true outcome of each game. These outcomes are coded as "1" if the event happens (e.g., Team A wins) and "0" if it does not (e.g., Team A loses). The further you are from the truth, the more your accuracy score will suffer (and the less likely it is that you will win an Amazon gift card). Thus, it is in your best interest to provide the most accurate probability estimates possible.

It's as easy as that! That's everything you need to know to go make predictions!

**Have fun, and good luck!**

## Predictions

### Instructions:

Please review the following information and provide your prediction about who will win.

---

### **Game Information:**

\${Im://Field/1} vs. \${Im://Field/5}  
 \${Im://Field/9}

<b><u> \${Im://Field/1} Information:</u></b>	<b><u> \${Im://Field/5} Information:</u></b>
Regular Season Conference: \${Im://Field/2} Division 1 Win-Loss Record: \${Im://Field/3} Strength of Schedule*: \${Im://Field/4} Tournament Seeding**: \${Im://Field/10}	Regular Season Conference: \${Im://Field/6} Division 1 Win-Loss Record: \${Im://Field/7} Strength of Schedule* : \${Im://Field/8} Tournament Seeding** : \${Im://Field/11}

\*Strength of Schedule is a numerical ranking provided by espn.com that reflects the difficulty of a team's regular season schedule. Teams with more difficult schedules (i.e., who play more skilled opponents) are given lower numerical rankings.

\*\* Tournament Seeding is a numerical ranking provided by the NCAA that determines who will play whom in the first round of the tournament. Seeding is done independently in each of the four quadrants of the tournament bracket (corresponding to different geographical regions). Lower (i.e., better) seeding is generally given to teams with better win-loss records and more difficult regular season schedules.

Want to look at the big picture?

[See the whole bracket here](#)

**Which team do you think will win this game?**

- \${Im://Field/1}
- \${Im://Field/5}

**Please provide a probability estimate that reflects your confidence that \${Im://Field/1} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/1} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Please provide a probability estimate that reflects your confidence that \${Im://Field/5} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/5} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/1} wins, and \$0 if \${Im://Field/1} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/5} wins, and \$0 if \${Im://Field/5} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

If someone offered to pay the amount above, would you keep the claim ticket or sell it?

(Please keep in mind: at the end of the tournament, one of your choices will be selected at random and paid-out in real money)

- I would keep the ticket
- I would sell it

Email

**That's it for today's survey!**

Thank you for completing this portion of the study.

**The next round of predictions (for games among the "final four") will be sent out on Monday (3/27), and will be due before the first final four game on April 1st (likely in the evening). This survey will comprise 2 predictions in total.**

To ensure that your answers are associated with the right person, please confirm your contact information below.

First Name

Last Name

Powered by Qualtrics

### Loop and Merge Data: Elite Eight (Survey 8)

	<b>Field 1</b>	<b>Field 2</b>	<b>Field 3</b>	<b>Field 4</b>	<b>Field 5</b>	<b>Field 6</b>	<b>Field 7</b>	<b>Field 8</b>	<b>Field 9</b>	<b>Field 10</b>	<b>Field 11</b>
Kentucky	Southeastern	28-5	23	N. Carolina	Atlantic Coast	26-7	14	Elite Eight	2	1	
Xavier	Big East	21-13	16	Gonzaga	West Coast	32-1	102	Elite Eight	11	1	
Oregon	Pac-12	28-5	44	Kansas	Big 12	28-4	35	Elite Eight	3	1	
So. Carolina	Southeastern	21-10	47	Florida	Southeastern	24-8	8	Elite Eight	7	4	

## Title Page

# UPenn March Madness Study: Final Four Predictions

**Start Date:** March 27th, 2017

**End Date:** April 1st, 2017, (prior to the first final four game at 6pm)

**Number of Predictions in this Survey:** 2

**Expected Completion Time:** 5-10 min.

## General Instructions

### Instructions

Hello, and welcome back to the UPenn March Madness Study!

In today's survey, you will be asked to provide predictions about the two "Final Four" games of the 2017 NCAA Division I Men's Basketball Tournament (aka, "March Madness").

For each of these games, we will ask you to provide the following information:

1. A prediction about which team will win.
2. A numerical probability judgment that reflects your confidence that the team you selected will win (for more information on probabilistic prediction, please select the appropriate option below).
3. The minimum price at which you would sell a claim ticket for a wager that pays \$15 if the team you selected wins (and \$0 otherwise).
4. Your choice to either keep the claim ticket or sell it for the price specified above.

**Before we get started, would you like to refresh your memory on any of the following topics?**

- Study guidelines and payment schedule
- Additional resources related to the NCAA tournament
- Guidelines on how to make probabilistic predictions

I'm all set. Take me straight to the predictions.

## Things to Note

### Things to Note:

#### General:

- Surveys can be completed at any time before the specified end date, and do not need to be completed in a single sitting.
- When making your predictions, you are welcome to use outside resources! All that we ask is that your answers are your own -- please do not ask any other people to help you with your predictions.

#### Payment:

- Participants will be paid \$10 upon timely completion of the First Round Predictions survey and the Play-In Predictions survey (already completed).
- Participants will be paid an additional \$10 for providing predictions for all of the remaining games in the tournament.
- Participants who complete this study will also have the opportunity to win up to \$15 more by correctly predicting the outcome of a single game (selected at random).
- The three participants who provide the most accurate probabilistic predictions over the course of the tournament (more on this in a moment) will receive an additional \$50, \$25, and \$10, respectively.

If you have any questions or concerns about today's survey, please contact Josh Baker  
at [jbak@sas.upenn.edu](mailto:jbak@sas.upenn.edu)

I acknowledge that I have read and understood the information on this page.

## Additional Resources

### Additional Resources

- For any participants who are unfamiliar with the NCAA tournament, additional information about the structure and rules of the tournament can be found [here](#).
- For any participants who would like to see the latest news, information, and schedule for the tournament, all that and more can be found [here](#).
- For any participants who would like to make his or her predictions on paper, a printable version of the NCAA tournament bracket can be found [here](#).

## **Probability Inst.**

### **How to Provide Probability Estimates**

In mathematics, probability is typically defined in terms of relative frequencies. When rolling a die, for example, you can estimate the probability of rolling a "4" by making a large number of rolls and counting the number of times a "4" comes up. Although this process isn't perfect, you will eventually find that the number of "4's" you roll is very close to one-sixth of the total number of rolls. Thus, it makes sense to say that the probability of rolling a "4" is one-out-of-six, or 1/6.

For a single basketball game, however, it doesn't make much sense to think about probability in this way. You can't discover the probability of Team A beating Team B by having them play the same game a large number of times. That one game, by definition, only happens once.

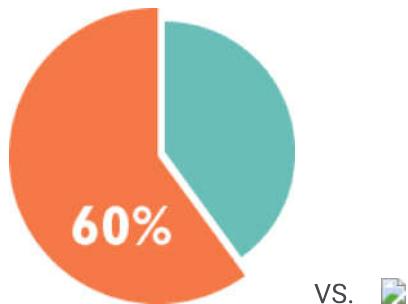
So how can you assign a probability to the outcome of a single game?

One way to proceed is to think about your preferences for various wagers regarding Team A. Suppose, for example, you were given a simple choice between betting on Team A or betting on a spinner (like a roulette wheel) that had a 50% chance of landing on a winning zone, both for the same prize. Which bet would you choose? If you would choose Team A, then your probability for Team A is higher than 50%.

Now, what if the spinner had a 60% chance of winning? At some point, as the probability of the spinner gets higher, you would switch and bet on the spinner rather than Team A. Just before this point, you would not care whether you bet on the spinner or Team A. Because the prizes for the two bets are the same, the fact that you are indifferent between the two suggests that you also believe the probabilities are the same. Thus, you can estimate the probability of Team A winning by thinking about the point where your preferences switch.

**Example:**

Which do you prefer? (1) A bet that pays \$15 if the spinner lands on orange; or (2) a bet that pays \$15 if Team A wins a basketball game?



If you prefer option 1 (the spinner), then the probability of Team A winning must be less than 60%. If you prefer option 2 (the basketball game), then the probability of Team A winning must be greater than 60%. If the two options are exactly the same, then the probability of Team A winning must be exactly 60%.

You can use this technique to estimate the probability of Team A winning. Slowly raise the probability of an imaginary spinner until betting on the spinner and betting on the basketball game feel exactly the same. Whatever number your imaginary spinner shows at this point reflects your beliefs about the probability of Team A winning, or your confidence that Team A will win.

---

In this study, we will ask you to predict the winner of each NCAA game, and to make probability judgments that reflect your beliefs about the likelihood that your favored team will win.

In most cases, you will probably have some intuition about which team will win, but could imagine being wrong. In these cases, you should provide a probability judgment that is between 50% and 100%. In cases where you are more confident, your probability judgment should be closer to 100%, and in cases where you are less confident, your probability judgments should be closer to 50%.

Ultimately, your accuracy will be determined by how close your judgments come to the true outcome of each game. These outcomes are coded as "1" if the event happens (e.g., Team A wins) and "0" if it does not (e.g., Team A loses). The further you are from the truth, the more your accuracy score will suffer (and the less likely it is that you will win an Amazon gift card). Thus, it is in your best interest to provide the most accurate probability estimates possible.

It's as easy as that! That's everything you need to know to go make predictions!

**Have fun, and good luck!**

## Predictions

### Instructions:

Please review the following information and provide your prediction about who will win.

---

### **Game Information:**

\${Im://Field/1} vs. \${Im://Field/5}  
 \${Im://Field/9}

<b><u> \${Im://Field/1} Information:</u></b>	<b><u> \${Im://Field/5} Information:</u></b>
Regular Season Conference: \${Im://Field/2} Division 1 Win-Loss Record: \${Im://Field/3} Strength of Schedule*: \${Im://Field/4} Tournament Seeding**: \${Im://Field/10}	Regular Season Conference: \${Im://Field/6} Division 1 Win-Loss Record: \${Im://Field/7} Strength of Schedule* : \${Im://Field/8} Tournament Seeding** : \${Im://Field/11}

\*Strength of Schedule is a numerical ranking provided by espn.com that reflects the difficulty of a team's regular season schedule. Teams with more difficult schedules (i.e., who play more skilled opponents) are given lower numerical rankings.

\*\* Tournament Seeding is a numerical ranking provided by the NCAA that determines who will play whom in the first round of the tournament. Seeding is done independently in each of the four quadrants of the tournament bracket (corresponding to different geographical regions). Lower (i.e., better) seeding is generally given to teams with better win-loss records and more difficult regular season schedules.

Want to look at the big picture?

[See the whole bracket here](#)

**Which team do you think will win this game?**

- \${Im://Field/1}
- \${Im://Field/5}

**Please provide a probability estimate that reflects your confidence that \${Im://Field/1} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/1} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Please provide a probability estimate that reflects your confidence that \${Im://Field/5} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/5} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/1} wins, and \$0 if \${Im://Field/1} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/5} wins, and \$0 if \${Im://Field/5} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

If someone offered to pay the amount above, would you keep the claim ticket or sell it?

(Please keep in mind: at the end of the tournament, one of your choices will be selected at random and paid-out in real money)

- I would keep the ticket
- I would sell it

Email

**That's it for today's survey!**

Thank you for completing this portion of the study.

**The final survey (for the championship game) will be sent out on April 1st, and will be due before the championship game on April 3rd (likely in the evening). This survey will comprise 1 prediction in total.**

To ensure that your answers are associated with the right person, please confirm your contact information below.

First Name

\${m://FirstName}

Last Name

\${m://LastName}

Powered by Qualtrics

### **Loop and Merge Data: Final Four (Survey 9)**

	<b>Field 1</b>	<b>Field 2</b>	<b>Field 3</b>	<b>Field 4</b>	<b>Field 5</b>	<b>Field 6</b>	<b>Field 7</b>	<b>Field 8</b>	<b>Field 9</b>	<b>Field 10</b>	<b>Field 11</b>
Oregon	Pac-12	28-5	44	N. Carolina	Atlantic Coast	26-7	14	Final Four	2	2	3
So. Carolina	Southeastern	21-10	47	Gonzaga	West Coast	32-1	102	Final Four	7	7	1

## Title Page

# UPenn March Madness Study: Championship Predictions

**Start Date:** April 2nd, 2017

**End Date:** April 3rd, 2017, (prior to the championship game at 9:20pm)

**Number of Predictions in this Survey:** 1

**Expected Completion Time:** 3-5 min.

## General Instructions

### Instructions

Hello, and welcome back to the UPenn March Madness Study!

In today's survey, you will be asked to provide your prediction about the Championship game of the 2017 NCAA Division I Men's Basketball Tournament (aka, "March Madness").

For this **final** game, we will ask you to provide the following information:

1. A prediction about which team will win.
2. A numerical probability judgment that reflects your confidence that the team you selected will win  
(for more information on probabilistic prediction, please select the appropriate option below).
3. The minimum price at which you would sell a claim ticket for a wager that pays \$15 if the team you selected wins (and \$0 otherwise).
4. Your choice to either keep the claim ticket or sell it for the price specified above.

**Before we get started, would you like to refresh your memory on any of the following topics?**

- Study guidelines and payment schedule
- Additional resources related to the NCAA tournament
- Guidelines on how to make probabilistic predictions

I'm all set. Take me straight to the predictions.

## Things to Note

### Things to Note:

#### General:

- Surveys can be completed at any time before the specified end date, and do not need to be completed in a single sitting.
- When making your predictions, you are welcome to use outside resources! All that we ask is that your answers are your own – please do not ask any other people to help you with your predictions.

#### Payment:

- Participants will be paid \$10 upon timely completion of the First Round Predictions survey and the Play-In Predictions survey (already completed).
- Participants will be paid an additional \$10 for providing predictions for all of the remaining games in the tournament.
- Participants who complete this study will also have the opportunity to win up to \$15 more by correctly predicting the outcome of a single game (selected at random).
- The three participants who provide the most accurate probabilistic predictions over the course of the tournament (more on this in a moment) will receive an additional \$50, \$25, and \$10, respectively.

If you have any questions or concerns about today's survey, please contact Josh Baker  
at [jbak@sas.upenn.edu](mailto:jbak@sas.upenn.edu)

I acknowledge that I have read and understood the information on this page.

## Additional Resources

### Additional Resources

- For any participants who are unfamiliar with the NCAA tournament, additional information about the structure and rules of the tournament can be found [here](#).
- For any participants who would like to see the latest news, information, and schedule for the tournament, all that and more can be found [here](#).
- For any participants who would like to make his or her predictions on paper, a printable version of the NCAA tournament bracket can be found [here](#).

## Probability Inst.

### How to Provide Probability Estimates

In mathematics, probability is typically defined in terms of relative frequencies. When rolling a die, for example, you can estimate the probability of rolling a "4" by making a large number of rolls and counting the number of times a "4" comes up. Although this process isn't perfect, you will eventually find that the number of "4's" you roll is very close to one-sixth of the total number of rolls. Thus, it makes sense to say that the probability of rolling a "4" is one-out-of-six, or 1/6.

For a single basketball game, however, it doesn't make much sense to think about probability in this way. You can't discover the probability of Team A beating Team B by having them play the same game a large number of times. That one game, by definition, only happens once.

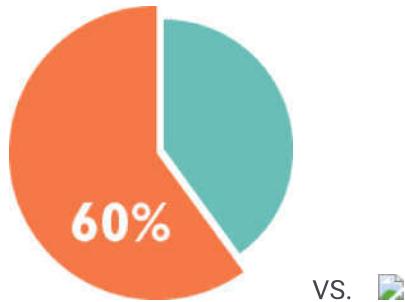
So how can you assign a probability to the outcome of a single game?

One way to proceed is to think about your preferences for various wagers regarding Team A. Suppose, for example, you were given a simple choice between betting on Team A or betting on a spinner (like a roulette wheel) that had a 50% chance of landing on a winning zone, both for the same prize. Which bet would you choose? If you would choose Team A, then your probability for Team A is higher than 50%.

Now, what if the spinner had a 60% chance of winning? At some point, as the probability of the spinner gets higher, you would switch and bet on the spinner rather than Team A. Just before this point, you would not care whether you bet on the spinner or Team A. Because the prizes for the two bets are the same, the fact that you are indifferent between the two suggests that you also believe the probabilities are the same. Thus, you can estimate the probability of Team A winning by thinking about the point where your preferences switch.

**Example:**

Which do you prefer? (1) A bet that pays \$15 if the spinner lands on orange; or (2) a bet that pays \$15 if Team A wins a basketball game?



If you prefer option 1 (the spinner), then the probability of Team A winning must be less than 60%. If you prefer option 2 (the basketball game), then the probability of Team A winning must be greater than 60%. If the two options are exactly the same, then the probability of Team A winning must be exactly 60%.

You can use this technique to estimate the probability of Team A winning. Slowly raise the probability of an imaginary spinner until betting on the spinner and betting on the basketball game feel exactly the same. Whatever number your imaginary spinner shows at this point reflects your beliefs about the probability of Team A winning, or your confidence that Team A will win.

---

In this study, we will ask you to predict the winner of each NCAA game, and to make probability judgments that reflect your beliefs about the likelihood that your favored team will win.

In most cases, you will probably have some intuition about which team will win, but could imagine being wrong. In these cases, you should provide a probability judgment that is between 50% and 100%. In cases where you are more confident, your probability judgment should be closer to 100%, and in cases where you are less confident, your probability judgments should be closer to 50%.

Ultimately, your accuracy will be determined by how close your judgments come to the true outcome of each game. These outcomes are coded as "1" if the event happens (e.g., Team A wins) and "0" if it does not (e.g., Team A loses). The further you are from the truth, the more your accuracy score will suffer (and the less likely it is that you will win an Amazon gift card). Thus, it is in your best interest to provide the most accurate probability estimates possible.

It's as easy as that! That's everything you need to know to go make predictions!

**Have fun, and good luck!**

## Predictions

### Instructions:

Please review the following information and provide your prediction about who will win.

---

### Game Information:

`${Im://Field/1} vs. ${Im://Field/5}`  
 `${Im://Field/9}`

<code> \${Im://Field/1} Information:</code>	<code> \${Im://Field/5} Information:</code>
Regular Season Conference: <code> \${Im://Field/2}</code> Division 1 Win-Loss Record: <code> \${Im://Field/3}</code> Strength of Schedule*: <code> \${Im://Field/4}</code> Tournament Seeding**: <code> \${Im://Field/10}</code>	Regular Season Conference: <code> \${Im://Field/6}</code> Division 1 Win-Loss Record: <code> \${Im://Field/7}</code> Strength of Schedule* : <code> \${Im://Field/8}</code> Tournament Seeding** : <code> \${Im://Field/11}</code>

\*Strength of Schedule is a numerical ranking provided by espn.com that reflects the difficulty of a team's regular season schedule. Teams with more difficult schedules (i.e., who play more skilled opponents) are given lower numerical rankings.

\*\* Tournament Seeding is a numerical ranking provided by the NCAA that determines who will play whom in the first round of the tournament. Seeding is done independently in each of the four quadrants of the tournament bracket (corresponding to different geographical regions). Lower (i.e., better) seeding is generally given to teams with better win-loss records and more difficult regular season schedules.

Want to look at the big picture?

[See the whole bracket here](#)

**Which team do you think will win this game?**

- `${Im://Field/1}`
- `${Im://Field/5}`

**Please provide a probability estimate that reflects your confidence that \${Im://Field/1} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/1} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Please provide a probability estimate that reflects your confidence that \${Im://Field/5} will win this game.**

**Note:**

- Higher numbers indicate higher confidence that \${Im://Field/5} will win.
- 50% = even chance, or a random guess.
- 100% = absolute certainty.

(Please enter a number between 50 and 100, excluding the percentage sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/1} wins, and \$0 if \${Im://Field/1} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

**Imagine you have a claim ticket for a wager that will pay \$15 if \${Im://Field/5} wins, and \$0 if \${Im://Field/5} loses.**

**If you were forced to sell this ticket, what is the minimum amount you would sell it for?**

(Please enter a dollar amount between \$0 and \$15, excluding the dollar sign)

If someone offered to pay the amount above, would you keep the claim ticket or sell it?

(Please keep in mind: at the end of the tournament, one of your choices will be selected at random and paid-out in real money)

- I would keep the ticket
- I would sell it

Email

**That's it for the entire study!**

Thanks so much for your participation!

**Final payments and a summary of your performance will be sent out as soon as we can tally up everyone's scores -- likely no later than next weekend (4/8/17 - 4/9/17).**

**As a reminder, here's what you stand to win:**

1. **An additional \$10 if you completed at least 80% of the predictions in this study.**
2. **Up to \$15 more for one of your "choices" (about keeping or selling the claim ticket for each game), selected at random.**
3. **The three participants who made the most accurate probabilistic predictions will win an additional \$50, \$25, and \$10, respectively!**

To ensure that your answers are associated with the right person, please confirm your contact information below.

First Name

\${m://FirstName}

Last Name

\${m://LastName}

Powered by Qualtrics

### **Loop and Merge Data: Championship Game (Survey 10)**

<b>Field 1</b>	<b>Field 2</b>	<b>Field 3</b>	<b>Field 4</b>	<b>Field 5</b>	<b>Field 6</b>	<b>Field 7</b>	<b>Field 8</b>	<b>Field 9</b>	<b>Field 10</b>	<b>Field 11</b>
Gonzaga	West Coast	32-1	102	N. Carolina	Atlantic Coast	26-7	14	Championship	1	1

## **APPENDIX C**

Appendix C contains PDF facsimiles of all materials administered in the Philadelphia air-temperature study. These materials are presented in the sequence they were administered to participants. Due to the way these materials were coded, the appearance of items is likely to have varied slightly across web-browsers. The present facsimiles were printed from the Google Chrome browser using the Foxit PDF printing application. The only substantive difference between these materials and those seen by participants is the inclusion of the subheading “(baker2)” on the first page of the survey. This subheading was not included in the version of the survey administered to participants.

# Temperature predictions (baker2)

This is about predicting tomorrow's temperature from today's, with probabilities. We give you a day's temperature, and your task is to give the probability that tomorrow's temperature will higher, or lower, and whether it will be at least 5 degrees (F) higher or lower.

The temperatures are the daily high temperatures recorded at Philadelphia International Airport (southwest of the city, near Wilmington Delaware). We are using only the months of January, the coldest month, and July, the hottest. The average daily high temperatures are 40.5 degrees (F) for January, with a range from 13 to 68, and 88.6 degrees for July with a range from 72 to 103.

Answer in percent probability. You may use decimals, such as 99.5 or 52.2, but no letters and no % sign. Try to be precise, as we are interested in differences among conditions as well as overall

Item 1 out of 40:

The 5th lowest temperature in January 2008 was 35 degrees F.

What is the probability (in %) that the next day's temperature was lower than 35?

What is the probability that the next day's temperature was 40 or higher?

What is the probability that the next day's temperature was 30 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 2 out of 40:

The 5th highest temperature in January 2008 was 50 degrees F.

What is the probability (in %) that the next day's temperature was lower than 50?

What is the probability that the next day's temperature was 55 or higher?

What is the probability that the next day's temperature was 45 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 3 out of 40:

The 5th lowest temperature in July 2008 was 83 degrees F.

What is the probability (in %) that the next day's temperature was higher than 83?

What is the probability that the next day's temperature was 88 or higher?

What is the probability that the next day's temperature was 78 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 4 out of 40:

The 5th highest temperature in July 2008 was 92 degrees F.

What is the probability (in %) that the next day's temperature was higher than 92?

What is the probability that the next day's temperature was 97 or higher?

What is the probability that the next day's temperature was 87 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 5 out of 40:

The 5th lowest temperature in January 2009 was 29 degrees F.

What is the probability (in %) that the next day's temperature was lower than 29?

What is the probability that the next day's temperature was 34 or higher?

What is the probability that the next day's temperature was 24 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 6 out of 40:

The 5th highest temperature in January 2009 was 43 degrees F.

What is the probability (in %) that the next day's temperature was lower than 43?

What is the probability that the next day's temperature was 48 or higher?

What is the probability that the next day's temperature was 38 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 7 out of 40:

The 5th lowest temperature in July 2009 was 81 degrees F.

What is the probability (in %) that the next day's temperature was higher than 81?

What is the probability that the next day's temperature was 86 or higher?

What is the probability that the next day's temperature was 76 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 8 out of 40:

The 5th highest temperature in July 2009 was 89 degrees F.

What is the probability (in %) that the next day's temperature was higher than 89?

What is the probability that the next day's temperature was 94 or higher?

What is the probability that the next day's temperature was 84 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 9 out of 40:

The 5th lowest temperature in January 2010 was 30 degrees F.

What is the probability (in %) that the next day's temperature was lower than 30?

What is the probability that the next day's temperature was 35 or higher?

What is the probability that the next day's temperature was 25 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 10 out of 40:

The 5th highest temperature in January 2010 was 53 degrees F.

What is the probability (in %) that the next day's temperature was lower than 53?

What is the probability that the next day's temperature was 58 or higher?

What is the probability that the next day's temperature was 48 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 11 out of 40:

The 5th lowest temperature in July 2010 was 85 degrees F.

What is the probability (in %) that the next day's temperature was higher than 85?

What is the probability that the next day's temperature was 90 or higher?

What is the probability that the next day's temperature was 80 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 12 out of 40:

The 5th highest temperature in July 2010 was 97 degrees F.

What is the probability (in %) that the next day's temperature was higher than 97?

What is the probability that the next day's temperature was 102 or higher?

What is the probability that the next day's temperature was 92 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 13 out of 40:

The 5th lowest temperature in January 2011 was 30 degrees F.

What is the probability (in %) that the next day's temperature was lower than 30?

What is the probability that the next day's temperature was 35 or higher?

What is the probability that the next day's temperature was 25 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 14 out of 40:

The 5th highest temperature in January 2011 was 39 degrees F.

What is the probability (in %) that the next day's temperature was lower than 39?

What is the probability that the next day's temperature was 44 or higher?

What is the probability that the next day's temperature was 34 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 15 out of 40:

The 5th lowest temperature in July 2011 was 88 degrees F.

What is the probability (in %) that the next day's temperature was higher than 88?

What is the probability that the next day's temperature was 93 or higher?

What is the probability that the next day's temperature was 83 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 16 out of 40:

The 5th highest temperature in July 2011 was 96 degrees F.

What is the probability (in %) that the next day's temperature was higher than 96?

What is the probability that the next day's temperature was 101 or higher?

What is the probability that the next day's temperature was 91 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.  
July average is 88.6, range is 72 to 103.

Please write any comments on this page here (up to 255 characters):

---

Item 17 out of 40:

The 5th lowest temperature in January 2012 was 34 degrees F.

What is the probability (in %) that the next day's temperature was lower than 34?

What is the probability that the next day's temperature was 39 or higher?

What is the probability that the next day's temperature was 29 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

Please write any comments on this page here (up to 255 characters):

Item 18 out of 40:

The 5th highest temperature in January 2012 was 54 degrees F.

What is the probability (in %) that the next day's temperature was lower than 54?

What is the probability that the next day's temperature was 59 or higher?

What is the probability that the next day's temperature was 49 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 19 out of 40:

The 5th lowest temperature in July 2012 was 85 degrees F.

What is the probability (in %) that the next day's temperature was higher than 85?

What is the probability that the next day's temperature was 90 or higher?

What is the probability that the next day's temperature was 80 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 20 out of 40:

The 5th highest temperature in July 2012 was 97 degrees F.

What is the probability (in %) that the next day's temperature was higher than 97?

What is the probability that the next day's temperature was 102 or higher?

What is the probability that the next day's temperature was 92 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 21 out of 40:

The 5th lowest temperature in January 2013 was 29 degrees F.

What is the probability (in %) that the next day's temperature was lower than 29?

What is the probability that the next day's temperature was 34 or higher?

What is the probability that the next day's temperature was 24 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 22 out of 40:

The 5th highest temperature in January 2013 was 53 degrees F.

What is the probability (in %) that the next day's temperature was lower than 53?

What is the probability that the next day's temperature was 58 or higher?

What is the probability that the next day's temperature was 48 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 23 out of 40:

The 5th lowest temperature in July 2013 was 82 degrees F.

What is the probability (in %) that the next day's temperature was higher than 82?

What is the probability that the next day's temperature was 87 or higher?

What is the probability that the next day's temperature was 77 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 24 out of 40:

The 5th highest temperature in July 2013 was 94 degrees F.

What is the probability (in %) that the next day's temperature was higher than 94?

What is the probability that the next day's temperature was 99 or higher?

What is the probability that the next day's temperature was 89 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 25 out of 40:

The 5th lowest temperature in January 2014 was 21 degrees F.

What is the probability (in %) that the next day's temperature was lower than 21?

What is the probability that the next day's temperature was 26 or higher?

What is the probability that the next day's temperature was 16 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 26 out of 40:

The 5th highest temperature in January 2014 was 52 degrees F.

What is the probability (in %) that the next day's temperature was lower than 52?

What is the probability that the next day's temperature was 57 or higher?

What is the probability that the next day's temperature was 47 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 27 out of 40:

The 5th lowest temperature in July 2014 was 81 degrees F.

What is the probability (in %) that the next day's temperature was higher than 81?

What is the probability that the next day's temperature was 86 or higher?

What is the probability that the next day's temperature was 76 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 28 out of 40:

The 5th highest temperature in July 2014 was 93 degrees F.

What is the probability (in %) that the next day's temperature was higher than 93?

What is the probability that the next day's temperature was 98 or higher?

What is the probability that the next day's temperature was 88 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 29 out of 40:

The 5th lowest temperature in January 2015 was 32 degrees F.

What is the probability (in %) that the next day's temperature was lower than 32?

What is the probability that the next day's temperature was 37 or higher?

What is the probability that the next day's temperature was 27 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 30 out of 40:

The 5th highest temperature in January 2015 was 45 degrees F.

What is the probability (in %) that the next day's temperature was lower than 45?

What is the probability that the next day's temperature was 50 or higher?

What is the probability that the next day's temperature was 40 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 31 out of 40:

The 5th lowest temperature in July 2015 was 83 degrees F.

What is the probability (in %) that the next day's temperature was higher than 83?

What is the probability that the next day's temperature was 88 or higher?

What is the probability that the next day's temperature was 78 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 32 out of 40:

The 5th highest temperature in July 2015 was 91 degrees F.

What is the probability (in %) that the next day's temperature was higher than 91?

What is the probability that the next day's temperature was 96 or higher?

What is the probability that the next day's temperature was 86 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 33 out of 40:

The 5th lowest temperature in January 2016 was 31 degrees F.

What is the probability (in %) that the next day's temperature was lower than 31?

What is the probability that the next day's temperature was 36 or higher?

What is the probability that the next day's temperature was 26 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 34 out of 40:

The 5th highest temperature in January 2016 was 50 degrees F.

What is the probability (in %) that the next day's temperature was lower than 50?

What is the probability that the next day's temperature was 55 or higher?

What is the probability that the next day's temperature was 45 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 35 out of 40:

The 5th lowest temperature in July 2016 was 84 degrees F.

What is the probability (in %) that the next day's temperature was higher than 84?

What is the probability that the next day's temperature was 89 or higher?

What is the probability that the next day's temperature was 79 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 36 out of 40:

The 5th highest temperature in July 2016 was 95 degrees F.

What is the probability (in %) that the next day's temperature was higher than 95?

What is the probability that the next day's temperature was 100 or higher?

What is the probability that the next day's temperature was 90 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 37 out of 40:

The 5th lowest temperature in January 2017 was 35 degrees F.

What is the probability (in %) that the next day's temperature was lower than 35?

What is the probability that the next day's temperature was 40 or higher?

What is the probability that the next day's temperature was 30 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 38 out of 40:

The 5th highest temperature in January 2017 was 54 degrees F.

What is the probability (in %) that the next day's temperature was lower than 54?

What is the probability that the next day's temperature was 59 or higher?

What is the probability that the next day's temperature was 49 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 39 out of 40:

The 5th lowest temperature in July 2017 was 82 degrees F.

What is the probability (in %) that the next day's temperature was higher than 82?

What is the probability that the next day's temperature was 87 or higher?

What is the probability that the next day's temperature was 77 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 40 out of 40:

The 5th highest temperature in July 2017 was 93 degrees F.

What is the probability (in %) that the next day's temperature was higher than 93?

What is the probability that the next day's temperature was 98 or higher?

What is the probability that the next day's temperature was 88 or lower?

---

Reminders:

January average is 40.5, range is 13 to 68.

July average is 88.6, range is 72 to 103.

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

Item 41 out of 40:

### Final questions

How much experience have you had thinking about temperatures in temperate climates (like Philadelphia)?

- I am a weather buff. I'm always looking at temperatures and forecasts.
  - I pay attention to temperatures, and I am familiar with temperate climates.
  - I don't have much relevant interest or experience.
  - This is all totally new to me.
- 

### Questions about thinking

Allowing oneself to be convinced by a solid opposing argument is a sign of good character.

- Completely agree  1  2  3  4  
 5 Completely disagree

People should take into consideration evidence that goes against conclusions they favor.

- Completely agree  1  2  3  4  
 5 Completely disagree

Being undecided or unsure is the result of muddled thinking.

- Completely agree  1  2  3  4  
 5 Completely disagree

People should revise their conclusions in response to relevant new information.

- Completely agree  1  2  3  4  
 5 Completely disagree

Changing your mind is a sign of weakness.

- Completely agree  1  2  3  4  
 5 Completely disagree

People should search actively for reasons why they might be wrong.

- Completely agree  1  2  3  4  
 5 Completely disagree

It is OK to ignore evidence against your established beliefs.

- Completely agree  1  2  3  4  
 5 Completely disagree

It is important to be loyal to your beliefs even when evidence is brought to bear against them.

- Completely agree  1  2  3  4  
 5 Completely disagree

When we are faced with a new question, the first answer that occurs to us is usually best.

Completely agree  1  2  3  4  
 5 Completely disagree

Good thinking leads to uncertainty when there are good arguments on both sides.

Completely agree  1  2  3  4  
 5 Completely disagree

When faced with a new question, we should consider more than one possible answer before reaching a conclusion.

Completely agree  1  2  3  4  
 5 Completely disagree

[Click here to go on.](#)

Please write any comments on this page here (up to 255 characters):

---

Please write any additional comments here (up to 255 characters):

---

If you disconnected from the Internet, you must reconnect before you click: