The Impact of Different Forms of Decision-Aids on User Best Assessments

Marc-André Desrosiers, FCAS, MBA

Abstract: In a world where information can be gathered, analyzed, interpreted and diffused much faster than for prior generations, we inquire about optimal schemes for the presentation of predictive models. We asked subjects to make predictions of quarterback ratings by presenting them with different information sets, some information sets including a predictive model prediction (with some different ways, some direct, some indirect, of presenting this information). We investigate (1) actual and optimal model usage by model users, (2) preferences of model users over ways to present information to them, and (3) inter-personal consistency of predictions. We find that (i) subjects are over-confident in their predictions, (ii) that this over-confidence is reduced when the subjects are presented with the predictive model, (iii) that subjects prefer an indirect presentation of the predictive model, where the model is presented as a deviation from a base statistic that is perceived to be relevant and credible, (iv) that their preferences are aligned with their own informal predictive model, and (v) that inter-personal consistency of predictions is fostered by the indirect presentation of the model (as a proposed deviation from a base statistic that is perceived to be credible and relevant).

Keywords: Clinical versus actuarial controversy, clinical synthesis, behavioral economics, information processing, business engineering, over-confidence, anchoring effects.

1. INTRODUCTION

When peering through a book like *Macrowkinomics* (Tapscott and Williams 2012), it becomes apparent that the ease with which information can be accumulated, processed, interpreted and used has greatly increased with the advent of new information technologies. While the authors may not focus on this, we believe that the exact nature of information processing by individuals, acting on their own or as part of a group, needs to be understood to be able to guide or better use the new possibilities arising out of increased access to (processed) information.

From a business perspective, in many fields, predictive models have long been distributed to users. For example, for insurance pricing purposes, rating manuals have been extent at least since the beginning of the 20th century. With the advent of the powerful central computers and databases, businesses could begin to gather massive amounts of information from their activities, process this information with statistical technology that was becoming more and more powerful, and harness the results in tools that could be deployed to field operatives. That trend has continued to the point that, with today’s technology, it is relatively easy to present field operatives with "informational dashboards" to assist them in their decision-making.

However, more information is not always better information and, even if all the pieces of available information were always relevant and pertinent, in the end, human beings still need to comprehend it, process it, interpret it and, ultimately, take decisions in reaction to it. Therefore, it is
worthwhile to understand the influence of the presentation of information has on human decision-making.

To investigate this issue, we decided to ask subjects to make predictions with access to different information sets: we have asked American undergraduate business students to make predictions of quarterback ratings. As such, we have decided to focus on the cognitive aspect of the influence of information sets; more specifically, the cognitive influence (in human decision-making) of the different ways a predictive model can be communicated.

Because of known behavioral effects such as over-confidence and anchoring, issues of actual and optimal model usage are especially interesting. How much do the subjects use the model when it is presented to them? What variables influence that usage? Can the subjects bring information that can help them beat the model? How much should the subjects be using the model? What variables influence how much the subjects should be using the model? Are the subjects bringing in this information in an optimal way? Moreover, is the across subjects consistency of predictions affected by the way the information is presented, keeping in mind that consistency places an upper bound on reliability of predictions?

Also, because information technology is not evolving in a vacuum, it is also worthwhile to ask about the preferences of the subjects over different ways information is presented to them. Do the subjects perceive differently information sets, in terms of the credibility and the relevance of the information they are provided with? How does this perception match up with the informal models subjects can use to make predictions? Does this preference line up with their own confidence in their predictions?
We find the following.

1. The subjects display massive over-confidence in their own predictive abilities, as optimal model usage should be about 70% (for the population of subjects as a whole) while actual model usage is bounded by 30% (again, for the population of subjects as a whole) in our experimental design, where the subjects that were attracted to the experiment tended to generally perceive themselves to be knowledgeable about the subject-matter.

2. The subjects can bring valuable information to the predictive models, as predictive accuracy could be increased by adding human inputs. However, the subjects do not bring in this information optimally as their actual predictive accuracy when they are presented with the model is less than optimal.

3. Presenting the subjects with the predictive model as the only piece of extra information most favors model usage: inducing about 30% model usage. Presenting the model indirectly, as a proposed deviation from a selected contextually relevant historical mean reduces model usage to about 15%. We interpret this extra use of the model when it is presented alone to be the result of an anchoring effect.

4. The only variable that significantly affects actual model usage is the Judging-Perceiving MBTI personality dimension where Judging individuals actually use the model more. The only variable that affects optimal model usage is the self-reported familiarity with football (that is, expertise in the field of interest): subjects more knowledgeable about the subject-matter need the model less.

5. Across subjects prediction consistency is fostered by presenting the predictive model indirectly as a proposed deviation from a base statistic that is perceived to be relevant and credible.

6. The subjects perceive more favorably base statistics that match better with their own informal predictive models.

7. When they are presented with the predictive model, the subjects prefer it to be presented indirectly, as a proposed deviation from a favorably perceived base statistic.

8. The perception of credibility and relevance of the provided information lines up well with the self-reported confidence the subjects have in their own predictions.
1.1. Research Context

The present research can be conceived to exist at the confluence of three major research programs: (1) the actuarial versus clinical controversy, (2) the research on information processing arising out of behavioral economics, and (3) the business engineering applied research stream. We will discuss each in turn.

1.1.1. Clinical versus actuarial controversy

In psychology and medicine, the clinical versus actuarial controversy relates to the extent to which a clinician can do and actually does better (or worse) than an available model based on predictive modeling. In "Man versus models of man: A rationale, plus some evidence, for a method of improving on clinical inferences" (Goldberg 1970), when examining how clinical psychologists, physicians, and other professionals that are typically called on to combine cues to arrive at some diagnostic or prognostic decision, it was found that, for the diagnostic task, models of the human decision-making were generally more valid than the humans themselves, even when models were developed on a relatively small set of cases and then humans and model competed on a completely new set. Such was the case because mathematical representations of such clinical judges can often be constructed to capture critical elements of their judgment strategies. Keeping in mind that judgment consistency sets an upper bound on judgment reliability, models of human decision-making, that are free from inconsistencies, outperforming actual human decision-making is indicative that judgment consistency is empirically more important than the ability of humans to incorporate information in more complex ways or using qualitative information that cannot be incorporated in the model of human decision-making. In "Effect of input from a mechanical model on clinical judgment" (Peterson and Pitz 1986), when exploring clinical synthesis, it was found that performance of subjects improved when the model was provided, but subjects still did less well than the model.

Peterson and Pitz further researched the topic. They found that it is worthwhile to make a distinction between the belief that a prediction is correct (i.e. confidence) and the ability of a subject making a prediction to imagine scenarios in which a prediction is not realized (i.e. uncertainty). They find that confidence increases as more information is provided to the subjects but that it was decreased when the difficulty of the task was increased. On the other hand, uncertainty increased as subjects were provided with more information. (Peterson and Pitz, Confidence, Uncertainty, and the Use of Information 1988, 85) They also found tendencies of over-confidence, with over-confidence a decreasing function of quantity of information provided. (Peterson and Pitz, Effects of Amount of Information on Predictions of Uncertain Quantities 1986, 229)

---

1 Clinical synthesis is giving the decision makers predictions from a model as input but let him or her make the final judgment.
Finally, in "A comparison of the predictive accuracy of loan officers and their linear-additive models" (Zimmer 1981), a clinical versus actuarial-type study was conducted on loan officers and found materially the same effects as when an health care practitioner population is examined.

An important insight that emerges from the actuarial versus clinical literature is that, because of the "broken leg" problem, where a user of the model may have a material insight into the problem at hand that cannot be easily or at all integrated into the model prediction, it is often sensible, for business purposes, to allow users of decision-aids to make the final selection.

1.1.2. Behavioral economics: over-confidence and anchoring

In economics, following the work of Simon (A Behavioral Model of Rational Choice 1955), significant efforts have been dedicated to refining theoretical models of preferences and human information processing. A seminal work in this area was "Judgment under Uncertainty: Heuristics and Biases" (Tversky and Kahnemen 1974) where the authors explore three families of heuristics that tend to induce individuals to misestimate probabilities:

(1) the representativeness heuristic, exemplified by the following biases:
   (a) insensitivity to prior probability of outcome,
   (b) insensitivity to sample size,
   (c) misconception of chance, e.g., gambler’s fallacy,
   (d) insensitivity to predictability,
   (e) illusion of validity, and
   (f) misconception of regression [towards the mean];
(2) the availability heuristic, exemplified by the following biases:
   (a) biases due to the retrievability of instances,
   (b) biases of imaginability, and
   (c) illusory correlation; and
(3) the adjustment and anchoring heuristic, exemplified by the following biases:
   (a) insufficient adjustment,
   (b) biases in the evaluation of conjunctive and disjunctive events, and
   (c) anchoring in the assessment of subjective probability distribution.

Thus, two factors that can influence actual and optimal model usage are (1) subject over-confidence due to subjects using a heuristic scheme to evaluate the relevant probabilities where the heuristic scheme leads to estimated probabilities that can be quite different from those that would be obtained under an optimal scheme (like Bayes’ theorem) and (2) anchoring which is particularly relevant for our purposes since it predicts that we can influence the selections made by the mere presentation of a numerical piece of information to the subjects. In "Framing, Probability Distortions, and Insurance Decisions" (Johnson, et al. 1993), potential business consequences of these information processing effects were explored. Among the findings of the study, it was found that sophisticated subjects can be subject to the same imperfect information processing bias as subjects in the general population are subject to.
1.1.3. Business engineering

Finally, there is an array of applied business research that examines how (imperfect) information processing effects influence business outcomes and whether or not business practices can be adapted to take into account these behavioral effects to help businesses attain their objectives. An example of applied research that examines reliance on statistical models is "Decision Aid Reliance: A Longitudinal Field Study Involving Professional Buy-Side Financial Analysts" (Hunton, Arnold and Reck 2010): this study examines the discretionary decision-aid reliance behavior of professional buy-side financial analysts. The researchers find that "the most interesting finding from theoretical and practice perspectives is that increased task ability, as determined by objective historical task performance, was associated with increased reliance on the DA." (Hunton, Arnold and Reck 2010, 1019) Another example of applied business research is "Decision making in an organizational setting: Cognitive and organizational influences on risk assessment in commercial lending" (McNamara and Bromiley 1997), a field study of decision-making, where it was found that "organizational effects appear to dominate cognitive ones" (McNamara and Bromiley 1997, 1083). Finally, in "Bridging the marketing theory-practice gap with marketing engineering" (Lilien, et al. 2002), the authors "provide several illustrations of the successful application of the marketing engineering concept" (Lilien, et al. 2002, 111), that are based on marketing management support systems that enrich decision-making.

This literature focuses on the application of pure knowledge about human behavior to solve business problems/issues. This involves applied research, but note also that other pure research avenues are often opened. As noted above, common applications relate to banking, insurance, finance, marketing, accounting, etc. What is generally found is that there is a gap between pure knowledge and the necessary optimal design that requires field testing. As such, this often brings up the issue of external validity that allows one to bridge from results obtained in one context to another, presumably similar, context.

1.2. Outline

The remaining will go as follows. Section 2 will be dedicated to the background and methods. Section 2.1 will contain a description of the experiment. Section 2.2 will relate specifically to the construction of the decision-aid used in the experiment. Section 3 will present and discuss the results. In section 3.1, we will present a model-of-man that takes into account the treatment the subjects received. In section 3.2, actual model compliance will be examined; in section 3.2.1, the effect of presenting a proposed deviation will be presented; in section 3.2.2, the actual (implicit) deviation will be explored as a function of the proposed (implicit) deviation; in section 3.2.3, drivers of actual model compliance will be sought; in section 3.3, optimal model compliance will be explored; in section 3.3.1, drivers of optimal model compliance will be sought; in section 3.4, the
reported perceptions of credibility and relevance of the presented information will be explored; in section 3.5, the self-reported confidence of subjects regarding their prediction will be explored; in section 3.6, the interpersonal agreement of subject predictions will be presented; in section 3.7, the net compensation outcomes will be presented and discussed; and, in section 3.8, external validity considerations will be discussed. Section 4 will conclude.

2. BACKGROUND AND METHODS

In this section, we will describe the exact nature of the experiment and the associated predictive model-building activities that needed to take place before the experiment could go live.

2.1. Experiment Description

Targeting undergraduate business students from University of Wisconsin-Madison, an internet survey asking subjects to make predictions about the quarterback rating results for the coming week of National Football League activity was distributed for three weeks in a row (NFL weeks 14, 15, and 16 of the 2012-2013 season). The target subjects were reached through in-class presentations of the experiment and targeted e-mail distribution lists. In particular, for week 16, subjects that had participated in weeks 14 and 15 were solicited to participate again.

<table>
<thead>
<tr>
<th>Week</th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>82</td>
<td></td>
<td>82</td>
</tr>
<tr>
<td>15</td>
<td>60</td>
<td>12</td>
<td>72</td>
</tr>
<tr>
<td>16</td>
<td>44</td>
<td>51</td>
<td>95</td>
</tr>
<tr>
<td>Total</td>
<td>186</td>
<td>63</td>
<td>249</td>
</tr>
</tbody>
</table>

Table 1: Number of subjects by week

2.1.1. Task description

The subjects were asked to make predictions about the quarterback ratings for the expected starters for the Sunday and Monday night games of the upcoming week of NFL games. The subject were then presented with some high level and general information concerning quarterback ratings: they were referred to the appropriate Wikipedia entry describing quarterback/passer rating and they were told about the possible values for the statistic, what a quarterback needed to accomplish to obtain a perfect score, the approximate league-wise average for the score, about an intra-week measure of league-wise quarterback rating dispersion, and about an across-weeks measure of intra-individual measure of quarterback rating dispersion.

Note that no predictions were asked for the Thursday night game: only predictions relating to the Sunday and Monday night games were asked.
2.1.2. Compensation description

They were then told what the compensation scheme would be. Subjects could gain from 0 to 15 USD, or up to 50¢ per prediction. They were told that their total compensation would be a sum of their by quarterback prediction compensations. Their by quarterback prediction compensations were computed using the following function.

![Graph 1: Per Quarterback Prediction Compensation Function]

The compensation function was selected to incentivize subjects to get every prediction as right as possible. We believe that subject risk aversion, cautiousness, financial cautiousness, etc. would not play a material role in affecting subject predictions: we will come back to this when examining drivers of model compliance.

2.1.3. Treatments descriptions

Table 2 provides a high level description of the treatments. Table 3 presents for which weeks each treatment was ran. Table 4 presents the verbose that was presented to the subjects under each treatment. Note that for each week, the 30 predictions that needed to be made were separated in two batches of 15 predictions, which were the same 15 predictions under all the treatments of the week. The subjects randomly received a treatment for each batch, independently of the other batch of prediction as well as independently of other subjects.

After each prediction batch, the subjects were also asked to reflect back on the predictions they just made. They were asked, in order, (1) how confident they felt about their predictions, (2) how confident they felt about the credibility and relevance of the information they were provided with for their predictions, and (3) to voluntarily discuss, in a free-form field, the strategies that supported their predictions.
The Impact of Different Forms of Decision-Aids on User Best Assessments

<table>
<thead>
<tr>
<th>Provided Information</th>
<th>A·</th>
<th>B·</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>·1 Overall Average</td>
<td>Individual Average</td>
<td>(only differ in nature of the average provided)</td>
<td></td>
</tr>
<tr>
<td>·2 Predictive Model Only</td>
<td>Predictive Model Only</td>
<td>(same)</td>
<td></td>
</tr>
<tr>
<td>·3 Overall Average plus Predictive Model Presented as a Proposed Deviation</td>
<td>Individual Average plus Predictive Model Presented as a Proposed Deviation</td>
<td>(see line ·1)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Descriptions of treatments
The possible treatments are A1, A2/B2, A3, B1, and B3.

<table>
<thead>
<tr>
<th>Possible Treatments</th>
<th>Week 14</th>
<th>Week 15</th>
<th>Week 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>YES</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>A2/B2</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>A3</td>
<td>YES</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>YES</td>
<td></td>
<td>YES</td>
</tr>
<tr>
<td>B3</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

Table 3: Time of the treatments

Note that the subjects were not described the predictive model beyond the details provided in Table 4. This was voluntary on our part as we wanted to make sure to reproduce a setting that is common in the deployment of predictive models: the predictive model is constructed using potentially complex statistical methodologies that are generally not described in details to the users of the predictive models. It is not uncommon that the users are only (1) told that the predictive model is the ‘best’ available model and (2) shown what variables enter the predictive model.

2.1.4. Subject population description

Demographic information was asked to the subjects in two batches: a batch of questions was asked to the subjects prior to their making their predictions and a second batch of questions was asked after the subjects completed their predictions.

The first batch of questions can be divided in three parts: (1) general demographic questions, (2) football trivia, and (3) self-assessed familiarity with football and fantasy sports. Under the heading of general demographics, questions were asked about: gender, age, level of study, study area, Grade Point Average, and mathematical abilities.
For each of the listed quarterback below, please enter your forecast for their quarterback rating in the coming game. You will find in brackets, first, their own team, second, the team they are playing against.

### A1
- **You will also find the overall average of the quarterback rating for the starting quarterbacks (OAAvg):**
  - Player (ForTeam, @AgainstTeam) OAAvg: aa.a

### A2/B2
- **You will also find the prediction from a statistical model (Stat. Model):**
  - The statistical model takes into account the following factors:
    - the overall quarterback rating average for the starters
    - the individual's season-to-date quarterback rating
    - the average season-to-date allowed quarterback rating for the opposing team
    - the consistency of the quarterback rating for the individual
  - Player (ForTeam, @AgainstTeam) Stat. Model: bb.b

### A3
- You will also find:
  1. **the overall average of the quarterback rating for the starting quarterbacks (OAAvg), and**
  2. **the prediction from a statistical model (Stat. Model Dev.) expressed as a deviation from the overall average. That is, Stat. Model = OAAvg + Stat.**

### B1
- **You will also find the individual average of the quarterback rating for the starting quarterbacks (IndAvg):**
  - Player (ForTeam, @AgainstTeam) IndAvg: dd.d

### B3
- **You will also find:**
  1. **the season-to-date quarterback rating for the individual (IndAvg), and**
  2. **the prediction from a statistical model (Stat. Model Dev.) expressed as a deviation from the season-to-date quarterback rating for the individual.**

In the table above, one can find the way the information was presented to subjects for each of the treatments described in Table 2.

The football trivia section was originally designed as an entertainment section: that is, it was intended to be fun for the subjects. The questions asked were:
1. "Which team won the last Super Bowl (played in February 2012)?",
2. "Who is a quarterback for the Green Bay Packers?",
3. "Which of these players is a defensive end who, in the 2011-2012 season, was a member of the Super Bowl winning team, went to the Pro Bowl, and lead his team for the number of sacks in the season?", and,
4. "Which of these players has posted the most games with a perfect passer rating?".

Given that UW-Madison is in Wisconsin, it was expected that most subjects would accurately identify Aaron Rodgers as the quarterback for the Green Bay Packers and this expectation was met. Given the publicity and level of public attention surrounding the Super Bowl, it was expected that most subjects would have known that the New York Giants had won the last played Super Bowl at the time of the survey. More subjects picked the wrong answer, but no subject picked a team that had not played in the Super Bowl game. Questions 3 and 4 were of a higher level of difficulty for the subjects and the accuracy of their answers followed accordingly.

While the intent behind asking questions about football was to make the survey more fun for the subjects, it may have reduced the participation of subjects whose level of familiarity with football may have been average or less than average. This provided us with a subject population biased towards ‘football experts’ (as contrasted with the general American population), but it also potentially had the side-effect of making the subject population more homogeneous.

Under the heading of self-assessed familiarity with football and fantasy sports, questions were asked about:

- the self-reported familiarity of the subjects with fantasy sports,
- the self-reported familiarity of the subjects with football,
- the average number of days per week that the subjects watch or read sports news, and,
- whether the subjects has a fantasy football team and, if so, whether the team was doing well or not.
The second batch of questions can be divided in two parts: (1) a simplified Myers-Briggs personality assessment, and (2) other demographic questions. The Myers-Briggs questions covered the following four personality dimensions (in parenthesis, the approximate proportion of the population that reported the dimension):

- the Judging (65-70%) versus Perceiving dimension,
- the Thinking (65-70%) versus Feeling dimension,
- the Sensing (55-60%) versus Intuiting dimension, and,
- the Introversion (60-65%) versus Extroversion dimension.

The other questions covered cautiousness, financial cautiousness, locus of control, and other questions about the statistics and sports statistics.

With all this demographic information available, it was easy for us to verify that the randomizing scheme integrated with our web-survey was adequate, as the generated sample appeared (materially and statistically) well-balanced on all dimensions.

2.2. Preparation of the Predictive Model of Quarterback Rating

The process to prepare predictive model quarterback rating predictions for the upcoming week of NFL activity went as follows. First, using a fantasy football website, the expected starting quarterbacks were retrieved. Because of ease of access of information and to ensure consistency of information processing with sources of information that the subjects could access on their own, some pieces of information were also retrieved from the said website: the individual season-to-date quarterback rating and detailed fantasy football experts’ predictions. The expert predictions were retrieved but ultimately not used in predictive modeling, because we wanted to ensure that the subjects could bring forward valuable qualitative information not reflected in the model, and not just information that would have arisen between the time the experts entered their predictions and the time when the subjects made their predictions. Still, the detailed expert predictions were translated into a predicted quarterback rating by first averaging the details of the predictions across the (three) experts and then converting to a predicted quarterback rating using the appropriate formula. Second, using football statistics websites, the remaining necessary information was retrieved:

- the overall quarterback rating average for the starters,
- the individual’s season-to-date quarterback rating,
- the average season-to-date allowed quarterback rating for the opposing team,
- the individual’s prior week quarterback rating result,
- the individual’s two week prior quarterback rating result,
- the consistency of the quarterback rating for the individual,

3 http://fantasynews.cbssports.com/fantasyfootball/playerindex/POS_QB.
4 Note that the order of averaging and applying the quarterback rating formula does not affect materially the predicted quarterback rating.
• the own team win-loss record, and,
• the opposing team win-loss record.

Each week, using the latest available information, the predictive model was calibrated by maximizing the compensation function across all (known) weeks (played at that time), using a weighting scheme that put more weight on the more recent weeks.\(^6\)

It is worthwhile to note that, from a predictive modeling point of view, both the overall quarterback rating average for the starters and the individual’s season-to-date quarterback rating are basically equally powerful one-variable predictors: they both would have yielded an average compensation of 16.1¢ per prediction had they been used for all predictions from weeks 6 to 17 (358 predictions). This is greatly due to the significant regression-to-the-mean that happens both at the league-wise level and at the intra-individual level. Note also that the predictive model would have generated a higher average compensation (on retrodicted quarterback ratings) than either (1) using only the overall mean, (2) using only the individual mean, (3) using only the predictions from the panel of experts on football, or (4) using a weighted average of the overall mean and the predictions of the panel of experts. The following table details these elements.

---

\(^6\) Please note that there was a significant revision to the predictive model between week 15 and 16: thus, the coefficients of the predictive model changed materially between the two weeks.
The Impact of Different Forms of Decision-Aids on User Best Assessments

Table 5: Coefficients of predictive models for weeks 14, 15 and 16, with and without predictions of panel of experts as an input.

<table>
<thead>
<tr>
<th>Week</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>St. Error</td>
<td>Estimate</td>
</tr>
<tr>
<td>Including Panel of Experts Predictions?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Starters Season-to-date</td>
<td>-0.42</td>
<td>0.43</td>
<td>0.19</td>
</tr>
<tr>
<td>Individual Season-to-date</td>
<td>-0.19</td>
<td>0.30</td>
<td>-0.65</td>
</tr>
<tr>
<td>Allowed QB Rating Opposing Team</td>
<td>1.11</td>
<td>0.26</td>
<td>0.16</td>
</tr>
<tr>
<td>Prior Week Result</td>
<td>0.07</td>
<td>0.12</td>
<td>-0.07</td>
</tr>
<tr>
<td>Prior Week Missing Indicator</td>
<td>6.97</td>
<td>5.72</td>
<td>-0.20</td>
</tr>
<tr>
<td>Two Prior Week Result</td>
<td>0.13</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>Two Prior Week Missing Indicator</td>
<td>12.01</td>
<td>6.41</td>
<td>1.02</td>
</tr>
<tr>
<td>Opposing Team Record</td>
<td>4.18</td>
<td>4.61</td>
<td>1.17</td>
</tr>
<tr>
<td>Opposing Team Record</td>
<td>29.10</td>
<td>5.93</td>
<td>6.86</td>
</tr>
<tr>
<td>Panel of Experts Predictions</td>
<td>1.11</td>
<td>0.24</td>
<td>1.10</td>
</tr>
<tr>
<td>Random-Effects-Like Parameter</td>
<td>8.17</td>
<td>5.63</td>
<td>0.00</td>
</tr>
<tr>
<td>Average Compensation on Retrodictions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Across 8 Weeks</td>
<td>0.176</td>
<td>0.014</td>
<td>0.183</td>
</tr>
<tr>
<td>Weighted Average of Panel Predictions and Overall Average</td>
<td>0.167</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>Starters Season-to-date</td>
<td>0.162</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>Panel of Experts Predictions</td>
<td>0.156</td>
<td>0.015</td>
<td></td>
</tr>
</tbody>
</table>

The above table contains the coefficients used in the predictive models for weeks 14, 15 and 16 with associated standard errors estimated using block bootstrap (that is, a bootstrap procedure where complete weeks are re-sampled). Note that the coefficients of the predictive models are generally not precisely estimated. Note that there was a significant model change between weeks 15 and 16. Nonetheless, the predictive model fares better on the average compensation (over many weeks) for retrodicted quarterback ratings than just using the predictions of the panel of experts, just using the average season-to-date for all starters, or using a weighted average of the predictions of the panel of experts and the overall average of the starting quarterbacks. Note also that, when the predictions of the panel of experts are included in the model, the average compensation of optimized retrodictions is higher than when that information is not included in the predictive models. We interpret this as evidence that human input can be significant in improving predictive performance.

In Table 5 above, note that when the predictions of the panel of experts is added to the predictive model, the average compensation (optimized) for the retrodicted quarterback ratings can be improved from what was attainable without the human input. We interpret this as evidence that human inputs can materially improve predictive performance. Furthermore, this improvement in
predictive performance seems to be substantially related to the random-effects-like\(^7\) parameter becoming unnecessary when the prediction of the panel of experts is added as a predictive variable.

\[\text{3. RESULTS AND DISCUSSION}\]

We can now turn our attention to the analysis of the results. First, we will quickly describe minor data manipulations that were done in order to ensure the interpretability of the data. Records where the quarterbacks the panel of experts predicted would play (that is, the quarterbacks for whom model predictions were built) but did not actually play were removed from the analysis and treated as ‘not available’. Cases were the subjects were effectively predicting that the quarterbacks for whom a prediction was sought would not actually play were also removed and treated as ‘not available’: a prediction threshold of 12.5 was used to accomplish this, as this constituted a natural cutoff point in the data.

\(^7\) Random-effects are meant to reflect fundamentally the same effects as those reflected in greatest accuracy credibility theory. Examples of predictive applications of random-effects-like predictive modeling applications can be found in (Fundamentals of Individual Risk Rating, Part I 1992). Random-effects-like formulas are meant to allow the predictive model to reflect individual differences, but only to the extent they are credible (or predictive).
3.1 Model-of-Man

Our first step in the analysis of the data was the construction of predictive models of subject predictions (that is, models-of-man).

![Table]

<table>
<thead>
<tr>
<th>Predictive Model Coefficients</th>
<th>Model-of-Man (without Model)</th>
<th>Model-of-Man (with Model)</th>
<th>Model-of-Man (with Model and Cognitive Dissonance)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
<td><strong>Week 14</strong></td>
<td><strong>Week 15</strong></td>
<td><strong>Week 16</strong></td>
</tr>
<tr>
<td><strong>Estimate</strong></td>
<td><strong>Std. Error</strong></td>
<td><strong>Estimate</strong></td>
<td><strong>Std. Error</strong></td>
</tr>
<tr>
<td>Intercept</td>
<td>-35.56</td>
<td>-14.22</td>
<td>79.08</td>
</tr>
<tr>
<td>Individual Season-to-date</td>
<td>-0.19</td>
<td>-0.21</td>
<td>0.75</td>
</tr>
<tr>
<td>Model</td>
<td>0.31</td>
<td>0.07</td>
<td>0.38</td>
</tr>
<tr>
<td>Opposing Team Record</td>
<td>4.18</td>
<td>3.91</td>
<td>13.18</td>
</tr>
<tr>
<td>Opposing Team Record A2/B2</td>
<td>29.10</td>
<td>28.46</td>
<td>3.25</td>
</tr>
<tr>
<td>Individual Season-to-date A2/B2</td>
<td>-0.25</td>
<td>0.07</td>
<td>0.21</td>
</tr>
<tr>
<td>Individual Season-to-date A3</td>
<td>-0.08</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>Individual Season-to-date B3</td>
<td>0.07</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>Model A2/B2</td>
<td>0.24</td>
<td>0.08</td>
<td>0.22</td>
</tr>
<tr>
<td>Model A3</td>
<td>0.08</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>Model B1</td>
<td>0.09</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>Model B3</td>
<td>0.06</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td>Allowed QB Rating Opposing Team</td>
<td>1.11</td>
<td>0.85</td>
<td>0.58</td>
</tr>
<tr>
<td>Prior Week Result</td>
<td>0.07</td>
<td>0.09</td>
<td>0.10</td>
</tr>
<tr>
<td>Prior Week Missing Indicator</td>
<td>6.97</td>
<td>7.61</td>
<td>1.26</td>
</tr>
<tr>
<td>Two Prior Week Result</td>
<td>0.13</td>
<td>0.17</td>
<td>0.22</td>
</tr>
<tr>
<td>Two Prior Week Missing Indicator</td>
<td>12.01</td>
<td>14.80</td>
<td>8.52</td>
</tr>
<tr>
<td>Individual Season-to-date A2/B2 Large Cognitive Dissonance</td>
<td>0.09</td>
<td>0.02</td>
<td>***</td>
</tr>
<tr>
<td>Individual Season-to-date A3/B2 Large Cognitive Dissonance</td>
<td>-0.05</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Individual Season-to-date B1/B2 Large Cognitive Dissonance</td>
<td>-0.01</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Individual Season-to-date B1/B2 Large Cognitive Dissonance</td>
<td>-0.07</td>
<td>0.04</td>
<td>*</td>
</tr>
<tr>
<td>Random-Effects-Like Parameter</td>
<td>8.17</td>
<td>7.06</td>
<td>11.43</td>
</tr>
<tr>
<td>Average Weighted Compensation (Retrodict)</td>
<td>0.176</td>
<td>0.181</td>
<td>0.177</td>
</tr>
</tbody>
</table>

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 6: Model-of-man with the predictive models attached.

The table above contains both the predictive models that were used to provide the subjects with guidance in treatments A2/B2, A3 and B3 and models-of-man, both including and excluding the model as a factor. Note that, while the predictive models vary from week to week, the weighted average compensation they would have generated on past predictions remain approximately constant. The factor "Random-Effects-Like Parameter" is effectively a ballast term to allow the model to appropriately recognize the average residuals (of a model that does not take into account performance consistency) for a given individual. When the predictive model is incorporated in the model-of-man, only the individual mean and the model prediction seem to interact with treatment A2/B2. The coefficients suggest that, for treatment A2/B2, the subjects were using less of the individual mean and substituting it for the model. This is consistent with an anchoring effect. The model-of-man that does not incorporate the model is suggestive of why the subjects would substitute from the individual mean under A2/B2: the individual mean is the most important component of their internal model. This is consistent with an analysis of median deviation from the presented relevant mean for A1 and B1: the median deviation is about 10 for A1 (which presents only the overall mean) and only about 5 for B1 (which presents only the individual mean).

Also, note that when the subjects do not agree with the model (this is indicated by the ‘Large Cognitive Dissonance’ variable which is an indicator that the unaided prediction and the model prediction are more than their median distance apart from each other), they tend to rely more on the individual average, but the effect is attenuated when the subjects are explicitly presented with the model.

At a first glance, it is interesting to note that the subjects were effectively using a model that was dissimilar to the optimized predictive model. For example, the mental model of the subjects did not appear to incorporate well regression-to-the-mean effects, as evidenced by the negligible coefficient
attached to the intercept for the model-of-man that does not include the model as a predictive variable. This was the case even though the subjects were explicitly told that regression-to-the-mean was a significant feature of the data. Also, the subjects seemed to believe that the individual mean was a powerful predictor of future performance, as evidenced by both the 0.53 highly significant coefficient attached to the ‘Individual Season-to-date’ variable in the model-of-man without the model as a variable as well as the fact that the median absolute deviation (from the selected mean) from the treatment were the subjects only received the individual mean was about half of the median absolute deviation under the treatment received only the overall mean.

When the ‘model’ variable is added to the model-of-man, a clear effect of information substitution appears: when presented only with the model, the subjects materially and significantly substitute the model prediction for the individual mean, as evidenced by +0.24 highly significant coefficient and -0.25 highly significant coefficient attached to the ‘Model:A2/B2 ‘ and ‘Individual Season-to-date:A2/B2’, respectively. Thus, we can speculate the following mental process could be occurring in the subjects. When they are presented with the model as a proposed deviation (that is, in treatments either A3 or B3), the subjects examine the information that the predictive model is providing them with and incorporate it in the best assessment predictions. The coefficient of 0.31 attached to the ‘Model’ variable in the model-of-man suggests that the subjects have a somewhat low valuation of the predictive model predictions. But, when they are presented only with the model as an extra piece of information (that is, in treatments A2/B2), then the subjects get anchored to the predictive model prediction and use the model more; that is to say, the data seems to suggest that the extra use of the predictive model by the subjects could well come from an anchoring effect where the mental model of the subjects gives extra weight to the model prediction because it is a (unique) number that has been floated to them just before they have to make their predictions. That the subjects substitute from the ‘Individual Season-to-date’ variable seems natural given its importance in their uninfluenced mental model.
Furthermore, we are led to wonder about the following. What happens to subject predictions when their own uninfluenced predictions would be ‘far away’ from the predictive model predictions; that is, what do subjects do when the predictive model appears to them as particularly less relevant? In the ‘Model-of-Man (with Model and Cognitive Dissonance)’, we added a ‘Large Cognitive Dissonance’ indicator variable: the indicator was set to 1 when the distance between the uninfluenced subject (average) prediction and the predictive model prediction was more than the median distance between the two predictions. From the coefficients of the ‘Individual Season-to-date’ crossed with the ‘Large Cognitive Dissonance’ crossed with the treatments variables, we see that the ‘Individual Season-to-date’ variable received a little bit more weight when the subjects agreed less with the predictive model. This leads us to suppose that the subjects gave increased weight to their own internal model when they agreed less with the predictive model prediction. This can be interpreted as a cognitive dissonance effect where the subjects ignore the model when they (particularly) do not agree with it.

3.2. Actual Model Compliance

We can now turn our attention towards quantifying model compliance under different treatments.

3.2.1. Effect of presenting a proposed deviation

From Graphs 2 and 3, we see that model compliance is greatest under treatments A2/B2 (at about 30% model compliance). The model compliance is reduced by about half when the subjects are presented with the model as a proposed deviation: there is about 15% model compliance under treatments A3 and B3. This is consistent with the effects found in the model-of-man analysis where (1) the subjects were found not to be utilizing the predictive model very much and (2) the subjects gave extra weight to the predictive model when that was all they were presented with.
Graph 2: Model compliance, expressed as a relationship between the actual average prediction for a given treatment against the average prediction when only the overall mean is presented, both axes normalized by the model prediction, under treatments A1, A2 and A3 for weeks 14 and 15.

A slope of 1 indicates no model usage: which necessarily happens for A1. A slope of 0 would indicate complete model compliance. The line corresponding to treatment A2 has a slope of approximately 0.7: thus, implying a model usage of about 30%. The line corresponding to treatment A3 has a slope of about 0.85: thus, implying a model compliance of about 15%. If the fitted line for treatment B3 had been added to the graph above, it would materially and statistically overlap the fitted line for treatment A3.

Graph 3: Model compliance, expressed as a relationship between the actual average prediction for a given treatment against the average prediction when only the overall mean is presented, both axes normalized by the model prediction, under treatments B1, B2 and B3 for week 16.

Similarly to Graph 2, the line corresponding to treatment B2 has a slope of approximately 0.7: thus, implying a model usage of about 30%; and, the line corresponding to treatment B3 has a slope of about 0.85: thus, implying a model compliance of about 15%.
3.2.2. Actual (implicit) deviation as a function of the proposed (implicit) deviation

Graphs 4 and 5 present the actual deviation as a function of the proposed deviation. Note that only under treatments A3 and B3 is the model presented explicitly as a proposed deviation. For models A1, B1 and A2/B2, this proposed deviation is implicit as the subjects either were not informed of the predictive model prediction (A1/B1) or were not presented the predictive model predictions in this way (A2/B2). A slope of 1 with a $R^2$ of 1 in Graphs 4 and 5 would indicate complete model compliance. As was found in the previous section, the subjects tended to use more of the proposed (implicit) deviation under treatments A2/B2, less so under models A3 and B3, and even less so under treatments A1 and B1. What is also apparent in Graphs 4 and 5 is that the proportion of the subjects actual (implicit) deviation that can be explained by the proposed (implicit) deviation increases from treatments A1/B1 to A3/B3 to A2/B2. This is consistent with our working hypothesis that the model ‘takes up more mental place’ when the subjects are only presented the predictive model as supplementary information.

It is interesting to note that one may measure the causal impact on subject predictions of changing the provided information from that of an initial information set to another information set by using the experimental methodology laid out above. Of particular interest to the experimenter was the causal impact of changing the base from which the model is presented as a proposed deviation; that is, the causal impact of changing the information set from that of A3 (overall average with model as proposed deviation) to B3 (individual average with model as proposed deviation). With the data obtained under the experiment, it is difficult to conclude the following definitely (because of the high leverage points in Graphs 4 and 5), but, visually, it appears that the actual deviation as a function of the proposed deviation is not overly affected by the change of basis from which the model is presented as a deviation. This may suggest that the subjects could be selecting an average level of the proposed deviation that they will use and that this level may be a function of the model and information set only and not of the basis from which the model is presented as a deviation. For business purposes, identifying if that effect was actually present would be relevant if one was attempting to forecast the impact of a change in information set on the ultimate usage of the model by users. In a marketing or sales setting, this could arise if users were presented with an array of prices (e.g. the cost at target profit, with the walk-away price, the suggested price of a first offer, the predicted valuation of the best alternative offer available to the customer, etc. presented as deviations).
The Impact of Different Forms of Decision-Aids on User Best Assessments

Graph 4: The actual (implicit) deviation from the model expressed as a function of the proposed (implicit) deviation, for treatments A1, A2, and A3, for weeks 14 and 15.

Note that the deviations from the model are implicit for treatments A1 and A2, as the subjects were not explicitly presented a proposed deviation. A slope of 0 means no agreement with the model, while a slope of 1 indicates complete agreement with the model. Similarly to what was found in Graph 2, the fitted line for treatment A3 approximately bisects the angle formed by the lines for A1 and A2. Note also that the proportion of the subjects’ predictions that can be explained by the (implicit) proposed deviation going from treatment A1 to A3 to A2 increases.

Graph 5: The actual (implicit) deviation from the model expressed as a function of the proposed (implicit) deviation, for treatments B1, B2, and B3, for week 16.

Note that the deviations from the model are implicit for treatments B1 and B2, as the subjects were not explicitly presented a proposed deviation. Similarly to what was found in Graph 3, the fitted line for treatment B3 approximately bisects the angle formed by the lines for B1 and B2. Note that the slope for treatment B1 is closer to 1 than the slope for treatment A1. The change in slope obtained by going from B1 to B3 is numerically similar to that obtained from going from A1 to A3; similarly, the change in slope obtained by going from B3 to B2 is numerically similar from that obtained by going from A3 to A2. As in Graph 4, note also that the proportion of the subjects’ predictions that can be explained by the (implicit) proposed deviation going from treatment B1 to B3 to B2 increases.

3.2.3. Driver of actual model compliance

We examined if we could identify demographic characteristics of the subjects that could help us predict which subjects would follow the model more or less. As mentioned earlier, the prudence or
risk aversion of the subjects could have potentially affected the response of the subjects such that the exact nature of the compensation function (as opposed to its essence which was to encourage subjects to provide us with their best assessment) could have influenced the outcome of the experiment.

The only demographic dimension that seemed to influence the behavior of the subjects was the indicator for the Judging inclination of our (simplified) Myers-Briggs test. For that dimension, subjects that self-reported the Judging inclination tended to make greater use the model: note that the effect was not consistently observed as can be seen in Table 7.

<table>
<thead>
<tr>
<th>Actual Deviation as a function of Proposed Deviation</th>
<th>Weeks 14 &amp; 15</th>
<th></th>
<th>Weeks 16</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.06</td>
<td></td>
<td>0.33</td>
<td>0.04</td>
</tr>
<tr>
<td>Indicator for Judging Indination</td>
<td>0.19</td>
<td>0.07</td>
<td>0.01</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 7: Actual deviation expressed a function of the proposed deviation for those treatments with explicitly provided proposed deviations: that is, weeks 14 and 15 A3 and B3 as well as B3 for week 16. For A3 in weeks 14 and 15 and B3 for week 16, the proportion of the proposed deviation actually used by the subject is materially and statistically higher. This could be explained by individuals with a judging inclination seeing "the need for most rules" and liking "to make & stick with plans" (PersonalityType.com/LLC n.d.).

A working hypothesis to explain that effect is that subjects self-reporting the Judging inclination reported, among other things, to see the need for most rules and prefer to stick with plans. It is then no great stretch to see model compliance as either thoughtful rule abiding (presuming a neutral or positive perception of the predictive model) or as a form of disciplining of the subject's own predictions.

3.3. Optimal Model Compliance

Given known over-confidence behavioral effects, we examined whether the subjects, that were somewhat using the model, were doing so optimally or whether the subjects were ignoring valuable insights from the model and over-weighting their own mental model.
For subjects that were not explicitly presented with the predictive model (that is, subjects that experienced the A1 and B1 treatments), we formed the convex combination of the subjects’ predictions and the predictive model predictions that generated the most favorable compensation for the subjects. To obtain an estimated distribution for the level of optimal model compliance, we then repeated the exercise with bootstrapped samples. This distribution is presented below.

The identified optimal model compliance is about 70%. This suggests massive over-confidence of the subjects in their own differential predictive abilities as actual model usage is much lower than 70% when the subjects are presented with the model (30% for A2/B2 and 15% for A3/B3).

Note that our estimation procedure presumed that subjects not explicitly presented with the model were effectively not influenced by the predictive model. While this effective assumption cannot be practically improved upon, it is less than perfect as the overall mean and the individual mean were variables that were part of the predictive model and thus correlated with the model.

### 3.3.1. Drivers of optimal model compliance

Contrasted with the drivers of actual model compliance are the drivers of optimal model compliance. Here, we are attempting to identify which subjects should have used the model more or less.
Here again, only one demographic dimension appears important: the subject self-reported familiarity with football. Subjects that self-reported they knew football the most were also the ones that should be using the model the least. However, even subjects that needed the model the least were still over-confident in their abilities: their model usage is bounded above by 30% and their optimal model usage was about 50%.

<table>
<thead>
<tr>
<th>Bootstrapped Optimal Model Compliance</th>
<th>Self-Reported Familiarity with Football</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familiarity Level:</td>
<td>1</td>
</tr>
<tr>
<td>Min.</td>
<td>0.00</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.28</td>
</tr>
<tr>
<td>Median</td>
<td>0.69</td>
</tr>
<tr>
<td>Mean</td>
<td>0.53</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>0.79</td>
</tr>
<tr>
<td>Max.</td>
<td>1.00</td>
</tr>
<tr>
<td>Number of Subjects</td>
<td>94</td>
</tr>
<tr>
<td>Prop. of Subjects</td>
<td>64%</td>
</tr>
</tbody>
</table>

Welch Two Sample t-test of whether the mean of the optimal model usage is the same for 1 (very familiar) as for the other self-reported levels of familiarity with football.

Test statistic: -8.1172
Degrees of Freedom: 1179.003
p-value: 1.191 x 10^-15

Table 8: Optimal model compliance as a function of the self-reported level of familiarity with football.

The distribution of the estimated optimal compliance levels were obtained using (1) bootstrapped samples of predictions of subject receiving the A1 and B1 treatments, and (2) optimizing the average prediction compensation by taking a weighted average of the subject prediction and the model prediction. Visually, level 1 (very familiar), levels 2 to 4, and level 5 should be binned together. Because the number of subjects reporting level 5 (very unfamiliar) is small, levels 2 to 5 are binned together. Note that the subject count only covers treatments A1 and B1, but counts them once per time within a week where they get treatment: keep in mind that, for each week, each subject received two treatments. Using the proposed binning, the mean optimal model compliance for level 1 is materially and statistically different from that of the other levels.

Note that subjects that repeated the experiment also needed the model less than other subjects, but it is also the case that subjects that repeated the experiment also had greater self-reported familiarity with football. This is a possibly natural association as subjects familiar with football should also be subjects that found the experiment more fun and thus worthy of repetition.
3.4. Perceptions of Credibility and Relevance

As mentioned in the ‘Treatments descriptions’ section, the subjects were asked about their perceptions of the information sets. Subjects’ reported perceptions of the provided tools was thought to be relevant because, in a business setting, if employees feel that they do not have access to the necessary tools to do their work well, the irritation that the employees feel towards the employer may become so dramatic as to cause significantly lower levels of employee engagement and thus lead to decreased productivity.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1//B1</td>
<td>11%</td>
<td>24%</td>
<td>25%</td>
<td>23%</td>
<td>16%</td>
<td>17%</td>
<td>17%</td>
<td>41%</td>
<td>28%</td>
<td>12%</td>
</tr>
<tr>
<td>A2/B2</td>
<td>5%</td>
<td>38%</td>
<td>37%</td>
<td>17%</td>
<td>3%</td>
<td>4%</td>
<td>4%</td>
<td>35%</td>
<td>42%</td>
<td>18%</td>
</tr>
<tr>
<td>A3</td>
<td>9%</td>
<td>25%</td>
<td>35%</td>
<td>25%</td>
<td>7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>9%</td>
<td>51%</td>
<td>27%</td>
<td>11%</td>
<td>2%</td>
<td>17%</td>
<td>38%</td>
<td>36%</td>
<td>9%</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CDF (from C5 to C1)</th>
<th>Weeks 14 and 15 - C1</th>
<th>Weeks 14 and 15 - C2</th>
<th>Weeks 14 and 15 - C3</th>
<th>Weeks 14 and 15 - C4</th>
<th>Weeks 14 and 15 - C5</th>
<th>Week 16 - C1</th>
<th>Week 16 - C2</th>
<th>Week 16 - C3</th>
<th>Week 16 - C4</th>
<th>Week 16 - C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1//B1</td>
<td>100%</td>
<td>89%</td>
<td>65%</td>
<td>39%</td>
<td>16%</td>
<td>100%</td>
<td>100%</td>
<td>83%</td>
<td>42%</td>
<td>14%</td>
</tr>
<tr>
<td>A2/B2</td>
<td>100%</td>
<td>95%</td>
<td>57%</td>
<td>20%</td>
<td>3%</td>
<td>100%</td>
<td>100%</td>
<td>96%</td>
<td>62%</td>
<td>20%</td>
</tr>
<tr>
<td>A3</td>
<td>100%</td>
<td>91%</td>
<td>67%</td>
<td>32%</td>
<td>7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>100%</td>
<td>91%</td>
<td>40%</td>
<td>13%</td>
<td>2%</td>
<td>100%</td>
<td>100%</td>
<td>83%</td>
<td>45%</td>
<td>9%</td>
</tr>
</tbody>
</table>

Table 9: Subject reported perception of credibility and relevance of provided information for A1, A2, A3 and B3 for weeks 14 and 15 and for B1, B2 and B3 for week 16. The table at the top is an empirical probability mass function over all possible responses from C5 (not at all confident in credibility and relevance) to C1 (highly confident in credibility and relevance). The bottom table is an empirical cumulative distribution function starting at C5.

Note that, if the provided information was unequivocally perceived as being more credible and relevant by the subjects than another alternative, then the empirical cumulative distribution function of credibility and relevance would be first-order stochastically dominated by that of the alternative. We see that providing only the model (A2) generates less negative perceptions of credibility and relevance than either of only providing the overall mean (A1) or providing the overall mean and the model as a proposed deviation (A3). Note, however, how providing the individual mean (as in B1 and B3) decreases the perceptions of negative credibility and relevance of information. This suggests the following simplified rule for predicting the perception of credibility and relevance: receiving the overall mean is less preferred than receiving the model only which is less preferred than receiving the individual mean. When a selected mean is provided, receiving the model as supplementary information does not appear to affect materially the perception of credibility and relevance.

The results from Table 9 suggest that subjects preferred least the information set under A1; that is, being only provided with the overall mean. Note that the subjects had a much less negative perception of the information set under B1 (only the individual mean). This alignment of subject preferences and internal subject model suggests that subjects prefer being provided with a model with which they agree even if, as is the case here, the model is actually not more predictive than other models. Recall that the overall mean and the individual mean are actually equally valuable individual pieces of information when presented alone and that this is largely due to the large regression to the mean at the inter-individual and intra-individual levels.8

---

8 It is worthwhile to note that both the overall average and the individual average represent valid perspectives on the predictive problem at hand, as is evidenced by the fact that they are both as powerful one-way predictors of quarterback
When the model is presented, the subjects appear to prefer it when it is presented as a proposed deviation from the individual mean but the subjects appear to prefer it to be presented directly if the alternative is to receive the model as a proposed deviation from the overall mean. This, then, suggests that the base (from which the model is presented as a deviation of) is an important driver of subject preferences (over information sets).

Recalling that subjects appeared to revert back to their internal model when they experienced cognitive dissonance with the predictive model, this preference for a base in which they believe may be rationalized by the ease with which subjects can revert back to their own internal model when they do experience cognitive dissonance.

Considering only week 16, it does appear that the subjects do (marginally) prefer also having the model (as opposed to not having it) when they are presented with the individual mean.

3.5. Self-Reported Confidence

The other side of subject preference is the induced self-reported confidence of the subjects after they make their predictions.

The findings relating to self-reported confidence (in predictions) line up pretty well with findings concerning the perceptions of credibility and relevance of the provided information sets. As a general rule, when the information set was thought to be more relevant and credible, the self-reported confidence in predictions also improved.

As can be seen in Table 10, the slight exception to the general rule can be found for week 16 in the comparison of treatments B1 and B3 were the B1 subjects reported less negative self-confidence than the B3 subjects but where the B3 information set was perceived less negatively than the B1 information set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1/B1</td>
<td>3%</td>
<td>33%</td>
<td>46%</td>
<td>11%</td>
<td>8%</td>
<td>3%</td>
<td>51%</td>
<td>35%</td>
<td>12%</td>
<td>0%</td>
</tr>
<tr>
<td>A2/B2</td>
<td>5%</td>
<td>39%</td>
<td>39%</td>
<td>15%</td>
<td>3%</td>
<td>5%</td>
<td>40%</td>
<td>36%</td>
<td>16%</td>
<td>2%</td>
</tr>
<tr>
<td>A3</td>
<td>0%</td>
<td>40%</td>
<td>43%</td>
<td>17%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>B3</td>
<td>1%</td>
<td>44%</td>
<td>41%</td>
<td>13%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CDF (from C5 to C1)</th>
<th>Weeks 14 and 15 - C1</th>
<th>Weeks 14 and 15 - C2</th>
<th>Weeks 14 and 15 - C3</th>
<th>Weeks 14 and 15 - C4</th>
<th>Weeks 14 and 15 - C5</th>
<th>Week 16 - C1</th>
<th>Week 16 - C2</th>
<th>Week 16 - C3</th>
<th>Week 16 - C4</th>
<th>Week 16 - C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1/B1</td>
<td>100%</td>
<td>98%</td>
<td>63%</td>
<td>19%</td>
<td>8%</td>
<td>100%</td>
<td>97%</td>
<td>46%</td>
<td>12%</td>
<td>0%</td>
</tr>
<tr>
<td>A2/B2</td>
<td>100%</td>
<td>99%</td>
<td>56%</td>
<td>17%</td>
<td>3%</td>
<td>100%</td>
<td>95%</td>
<td>55%</td>
<td>18%</td>
<td>2%</td>
</tr>
<tr>
<td>A3</td>
<td>100%</td>
<td>100%</td>
<td>60%</td>
<td>17%</td>
<td>0%</td>
<td>100%</td>
<td>94%</td>
<td>55%</td>
<td>15%</td>
<td>0%</td>
</tr>
<tr>
<td>B3</td>
<td>100%</td>
<td>99%</td>
<td>55%</td>
<td>13%</td>
<td>0%</td>
<td>100%</td>
<td>94%</td>
<td>55%</td>
<td>15%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 10: Subject self-reported perception of confidence in their predictions for A1, A2, A3 and B3 for weeks 14 and 15 and for B1, B2 and B3 for week 16. The table at the top is an empirical probability mass function over all possible ratings. As such, this experiment does not explore the effect of presenting a biased or voluntarily distorted statistic as a guide for prediction or as a base from which a model is presented as a proposed deviation.
responses from C5 (not confident at all) to C1 (very confident). The bottom table is an empirical cumulative distribution function starting at C5.

Similarly to Table 9, a treatment that induces greater self-reported confidence than another treatment would be stochastically dominated by the other treatment. The data above can be approximately summarized as follows: receiving the overall mean induces more negative self-reported confidence than receiving the overall mean and the model as a proposed deviation or just the model, which induces more negative self-reported confidence than receiving the individual mean and the model as a proposed deviation, with only receiving the individual mean inducing least negative self-reported confidence in the predictions.

3.6. Interpersonal Agreement

Before moving on to the net effect of treatment on the average compensation to the subjects, we wanted to examine the effect of the treatments on inter-personal agreement. This question is interesting because it has been found in the actuarial versus clinical debate that increased prediction consistency is an important factor in increased prediction accuracy, as consistency sets an upper bound on reliability.

As can be seen in Table 11, the inter-personal prediction consistency is roughly similar across treatments. However, subjects given the B3 treatments seem to have made more consistent predictions. The results of Table 11A confirm this diagnostic as the linear model of across subject standard deviation and variance of predictions for a given quarterback/week has a highly significantly negative average treatment effect for the B3 treatment.

Having seen in the "Actual Model Compliance" section that subjects given the B3 treatment use the predictive model about half as much as subjects given the A2/B2 treatments, we can now wonder if the increased consistency of predictions in the B3 treatment can empirically counter-balance the decreased model usage and allow the average compensation of the B3 treatment to be similar to that of the A2/B2 treatment.
The Impact of Different Forms of Decision-Aids on User Best Assessments

<table>
<thead>
<tr>
<th>Across Subjects</th>
<th>Standard Deviation of Predictions</th>
<th>Weeks 14 and 15</th>
<th>Week 16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>St. Dev.</td>
<td>Average</td>
</tr>
<tr>
<td>A1//B1</td>
<td>11.9</td>
<td>2.7</td>
<td>9.7</td>
</tr>
<tr>
<td>A2/B2</td>
<td>10.5</td>
<td>3.0</td>
<td>10.4</td>
</tr>
<tr>
<td>A3</td>
<td>11.3</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>9.6</td>
<td>2.4</td>
<td>10.4</td>
</tr>
</tbody>
</table>

Table 11: The average and standard deviation of the standard deviation of across subjects standard deviation of prediction for A1, A2, A3 and B3 for weeks 14 and 15 and for B1, B2 and B3 for week 16.

The inter-personal agreement of predictions is roughly similar across weeks and treatments. However, a potentially interesting difference is between B3 and A1 for weeks 14 and 15: B3 subjects seem to have been more consistent than A1 subjects.

<table>
<thead>
<tr>
<th>Standard Deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment #</td>
<td>Estimate</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Base Level</td>
<td>11.15</td>
</tr>
<tr>
<td>A2/B2</td>
<td>-0.68</td>
</tr>
<tr>
<td>A3</td>
<td>0.12</td>
</tr>
<tr>
<td>B3</td>
<td>-1.27</td>
</tr>
</tbody>
</table>

Table 11A: A linear model of across subject standard deviation and variance across quarterback/weeks prediction sets, by treatment.

The linear model of across subject standard deviation (and variance) for quarterback/week predictions is consistent with the results found in Table 11: treatment B3 induces increased consistency of predictions across subjects. Increased consistency of predictions can be a driving force of prediction accuracy, as was empirically found in the actuarial versus clinical debate.

3.7. Net Compensation Outcomes

We are now ready to analyze the net compensation outcomes by treatment. Because there are only three weeks of data, some caution is necessary in the interpretation of the results. In particular, there were only 90 quarterback/weeks in the sample; therefore, the performance comparison between the subjects and the predictive model must be examined with prudence. Moreover, even across treatments comparisons must be examined with care as the particular observed outcome differential was obtained under non-experimentally designed variations in information sets: that is, the observed difference in outcomes may reflect a particular mix of information sets induced by chance. In particular, the information sets were not designed to induce the optimal variation in variables that could have been of interest such as in generating a wide and balanced variety of (1) proposed deviation, (2) difference between the uninfluenced (average) subject prediction and predictive model predictions, (3) overall mean, (4) individual mean, etc.

<table>
<thead>
<tr>
<th>Compensation (per prediction) ($)</th>
<th>Weeks 14 and 15</th>
<th>Week 16</th>
<th>All Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1//B1</td>
<td>0.1770</td>
<td>1.15</td>
<td>0.0051</td>
</tr>
<tr>
<td>A2/B2</td>
<td>0.1826</td>
<td>1.080</td>
<td>0.0052</td>
</tr>
<tr>
<td>A3</td>
<td>0.1803</td>
<td>0.992</td>
<td>0.0053</td>
</tr>
<tr>
<td>B3</td>
<td>0.1921</td>
<td>1.160</td>
<td>0.0056</td>
</tr>
<tr>
<td>Total</td>
<td>0.1832</td>
<td>4.369</td>
<td>0.0026</td>
</tr>
<tr>
<td>Total excl. A1//B1</td>
<td>0.1853</td>
<td>3.232</td>
<td>0.0030</td>
</tr>
<tr>
<td>Model</td>
<td>0.1700</td>
<td>0.1820</td>
<td>0.1746</td>
</tr>
</tbody>
</table>
Table 12: The average compensation per prediction (with the associated number of predictions and standard deviation of compensation) for A1, A2, A3 and B3 for weeks 14 and 15 and for B1, B2 and B3 for week 16; the table includes sub-totals for those treatments where the subjects saw the model and the compensation that would have been obtained under the model.

There is a significant difference in compensation between A1 and B3 for weeks 14 and 15. The sources for the difference could be (1) increased prediction consistency, (2) increased perception of credibility and relevance of information, (3) increased quality of information provided. Note, however, that other differences do not appear as significant. This suggests that subjects needed to be provided with a reference that they believed in more and a model prediction to significantly increase their compensation performance. Keep in mind that the compensation scheme was constructed to incentivize subjects to attempt to be as correct as possible with regards their predictions.

When the subjects had access to the model, they performed better than without access to the model. In weeks 14 and 16, they did not (on average) beat the model; but they did (on average) beat the model in week 15.

From Table 12, we can see that treatments A2/B2 and B3 generated the same average compensation across the three weeks. Treatment A3 comes next. The clear loser in terms in generating favorable compensation for the subjects was treatment A1 with average compensation significantly and materially lower than the average compensation for treatment B3 in weeks 14 and 15. Compared to the compensation that would have been generated if the subjects had perfect model compliance, the subjects fared better than the predictive model over the course of the three weeks: even those subjects that were not presented with the model. Again, some caution is necessary here since the predictive model would have beat the subjects (on average) under all treatments for weeks 14 and 16, but the predictive model performed poorly in week 15. The evidence does suggest that many of the model predictions were useful to the subjects (even for week 15) as subjects that had access to the model fared better than those that did not have access to it.

It is also worthwhile to re-examine the results obtained in the ‘Optimal Model Compliance’ section: what would have been the average compensation of the subjects had they been using the model optimally? Table 13 answers this question. We can see that, under optimal model usage, the subjects would have generated an average compensation (0.1858) 2% higher than what they generated on average under treatments A2/B2 and B3 (0.1822). While a 2% improvement may not appear large at a first glance, if one thinks about the effect on Return on Equity of a 2% increase in the net income ratio, then this improvement appears much more substantial.

| Average Compensation per Prediction at Best Model Usage ($) |
|-----------------------------|-------------|
| Mean                        | 0.1858      |
| St. Dev.                    | 0.0036      |
| 1st Q                       | 0.1834      |
| Median                      | 0.1857      |
| 3rd Q                       | 0.1880      |

Table 13: Estimated features of the distribution of compensation of subjects if they used the model optimally.
This table represents the compensation the subjects would have attained by following the optimal weighted average of their unaided assessment and the model. Notice that this is statistically and materially (being a 2% performance improvement over the average compensation obtained under A2/B2 and B3) significant.

The results from Table 13 beg the question of what would have been the optimal (average) compensation for subjects that were already influenced by the predictive model.

<table>
<thead>
<tr>
<th>(Further) Model Use</th>
<th>A1//B1</th>
<th>A2/B2</th>
<th>A3</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0.0%</td>
<td>3.9%</td>
<td>2.9%</td>
<td>3.9%</td>
</tr>
<tr>
<td>25%</td>
<td>2.9%</td>
<td>4.3%</td>
<td>4.3%</td>
<td>4.6%</td>
</tr>
<tr>
<td>50%</td>
<td>5.2%</td>
<td>4.7%</td>
<td>5.3%</td>
<td>5.0%</td>
</tr>
<tr>
<td>75%</td>
<td>5.8%</td>
<td>4.7%</td>
<td>5.7%</td>
<td>4.9%</td>
</tr>
<tr>
<td>100%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>-3.0%</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

Table 14: Estimated gains from further model use by treatment.

The results shown in the above table can be interpreted in the following way. At no supplementary model use, the best outcomes are achieved under A2/B2 and B3, consistent with Table 12. At full model use, we obtain that there would be a gain compared to the unaided prediction under every treatment but A3, because of the poor performance of the model in week 15, as noted in the comments of Table 12. Consistent with the result of Table 13, in the table above, the best performance for treatments A1//B1 is achieved at 75% model for a 5.8% performance improvement. At best model usage, however, the optimal attainable performance is highest under the A1//B1 treatments. This may occur because the subject predictions correlate more with the model predictions when the subjects are presented with the model. Compare this with the engineering problem of identifying the best mix of instruments and weights assigned to instruments to generate the ‘best’ measurement: one would pick instruments whose measurement errors would be as little positively correlated as possible. In this case, it appears that the subjects are ‘destroying’ some of the statistical signal that they would be picking if they did not see the model when they are explicitly presented with the model.

Table 14 suggests that subjects that were already influenced by the predictive model could not possibly achieve a better (average) compensation by using a weighted average of the predictions under the treatment and the predictive model predictions. One working hypothesis one might have had is that the optimal attainable (average) compensation should not be a function of the treatment. However, think of the following analog problem. Suppose we are attempting to take a measure of an empirical reality. First, suppose that we have only two (unbiased) instruments, each with their own level of precision. If measurement errors of the two instruments were independent, then, to obtain the least variable and unbiased measurement, we should weight together the measurements of the two instruments such that more weight is assigned to the more precise instrument. Holding constant the precision of the two instruments, one would prefer the measurement errors of the two instruments to be as negatively correlated as possible, such that the error of one instrument should be naturally corrected by the error of the other instrument. However, in our case, we get a case where the measurements taken by the subjects under A2/B2, B3 and A3 are positively correlated with the predictive model predictions. Thus, because of this, at the optimal compensation, the optimal compensation is higher under the A1//B1 treatments. Nonetheless, the net model usage at optimum (average) compensation [not shown] are quite similar across treatments.
3.8. External Validity Considerations

Just as is the case for any social science experiment, one needs to examine the potential transferability of the results of the experiment to other (often, non-experimental) circumstances.

3.8.1. Non-neutrality of the compensation scheme and environment

In "An Effort Based Analysis of the Paradoxical Effects of Incentives on Decision-Aided Performance" (Samuels and Whitecotton 2011), the researchers found that

[In contrast to the findings of prior research, our study shows that incentives do not necessarily decrease performance in the presence of decision aids. Rather, we demonstrate that the effect of incentives on decision-aided performance depends on other contextual factors such as the absence or presence of additional contextual information." (345)

Th[e]n it is quite possible that changing the compensation scheme or the context of the experiment may affect the findings and, therefore, may make the findings non-transferable. Here is a potential example of such non-transferability. Imagine we go back to our example of the use of predictive models in a marketing or sales context. Now, imagine that the objective of the users of the predictive model is not to provide their best assessment of a future statistic but, instead, to try to optimize sales (subject to some profitability constraints). Even further, imagine that there are negotiations with a third party involved. In that case, while the final selection of the user is expected to be influenced by the model output (when provided), there is no representation that the user selection is a best assessment and may instead represent the impact of external constraints imposed by the third party. There is a priori no reason to think that these two contexts would generate similar results. However, in a context where the final price of a transaction is rationally related to the expectation of a future profitability statistic and where compensation to the users of the decision-aid is related to their (individual) accuracy in predicting that future profitability statistic, then the context of the experiment and the pricing situation become alike enough to expect some level of transferability. For example, this would naturally occur in an insurance pricing context, but also in the pricing of many financial products.

4. CONCLUSION

At this point, we wish to interpret the results of the experiment from a business perspective (in particular for insurance or retail financial products, like mortgages). Assume that the interest of the business is to have the users of the decision-aid be as accurate as possible in predicting the future performance of a profitability statistic attached to a (sold) contract. In our closely related experimental setting, which decision-aid would the business prefer to provide to the users? Assuming that the difference in costs in providing the different decision-aids were not material, then the business would prefer to provide a decision-aid that presents either only the predictive model or the predictive model presented as a proposed deviation from a statistic that the users of the
decision-aid find relevant and credible (in the case of the experiment, that was the individual mean) as they generate the same accuracy in user predictions. Which decision-aid would the users prefer to receive? Presumably, the users would prefer to receive higher compensation, be more confident in their predictions and find the decision-aid they receive relevant, and it is unclear \textit{a priori} in what order. For our purposes, the choice of the subject should be insensible to the exact nature of the preference because the same basic ranking comes up under any weighting of the preferences: the users would prefer to receive either only a statistic they perceive as relevant and credible or the model presented as a proposed deviation from the same statistic, as these two decision-aids generate similar compensation, similar self-confidence and similar perceptions of relevance of information. So, altogether, this implies that the decision-aid that should be deployed is the decision-aid with the predictive model presented as a proposed deviation from a statistic that is perceived by the subjects to be relevant and credible.

Further, applied research needs to be done at the business level: our research has not identified what features of the individual mean statistic made it so attractive to subjects and, even if we had done that, it is unclear that this piece of the research would be transferable. Note, however, that the experimental framework that we used should be implementable in an applied (business) research framework at relatively little costs. This means that the applied researchers can conduct meaningful applied (business) research, with an experimental inclination, without needing to develop tools to the point where they are ready to be deployed in a production environment: it should be apparent from our research that significant insights can be gathered in a simplified, even to the point of being skeletal, framework.

The current research does leave open many empirically interesting questions. For example, what would be the best way to seek out the subjective opinion of the subjects to arrive at the most predictive best assessment? Should the opinion of the user be sought and then the final prediction be generated mechanically from the recorded prediction and the output of the predictive model? If the user always needs to have a final say in the recorded prediction, should the users be limited in their ability to deviate from the predictive model? For example, should the users be limited in their freedom to deviate from the model only globally or should there be a limit on the ability of users to deviate from the model that applies prediction by prediction (or both)? Is there a way to ensure that the deviations from the model reproduce some form of distribution known to hold for the underlying population (that the predictions relate to) as a whole?

Acknowledgments

Marc-André Desrosiers would like to thank his actuarial colleagues for their assistance in the design, field testing and interpretation of the results of the experiment. They have helped ground the experiment in the realities of a ratemaking actuary that is going through the process from data
modeling to final field implementation. He would also like to thank Prof. Justin Sydnor for financial support: the compensation for students came from his research fund. Prof. Sydnor was also an invaluable discussion partner in the preparation of this text.
Appendix A. On-line Survey Tool

Consent Block

*UNIVERSITY OF WISCONSIN-MADISON*

*Subject CONSENT to Participate in Research Study "Quarterback Rating Experiment"*

*Title of the Study*: Quarterback Rating Forecasts
*Principal Investigator* (*PI*): Justin Sydnor (phone: 608-263-2138, email: jsydnor@bus.wisc.edu)
*Mailing Address*: 5287 Grainger Hall, Wisconsin School of Business, 975 University Ave., Madison, WI, 53706

*Introduction*

You are invited to participate in this research study about forecasting. We are studying how people make predictions about quarterback ratings in upcoming NFL games. You are invited to take part because you are a student at UW-Madison. Note that you must be a citizen of the United States of America to participate in the study: this is because we can only provide compensation to American citizens. Your participation is voluntary.

*Procedures*

If you decide to participate in this research, you will be asked to forecast the quarterback rating of quarterbacks expected to start in the coming weekend of NFL activity. We will also collect information about you for this research study. This information includes gender, year of birth, citizenship, attained education level, major, GPA. We will also ask you about your familiarity with, interest in and understanding of football, sports statistics, and fantasy sports. We will also ask questions to assess some of your personality traits. This questionnaire will be conducted with an online Qualtrics-created survey.

*Risks/Discomforts*

The only risk of taking part in this study is that your study information could become known to someone who is not involved in performing or monitoring this study.

*Benefits*

You are not expected to benefit directly from participating in this study. Your participation in this research study may benefit other people by helping us learn more about how individuals make decisions. There are no direct benefits to you from participating in this research.

*Compensation*

You will receive a compensation that will be determined as a function of your forecasts and the actual outcomes in the coming weekend of NFL football. Your compensation will range from zero (0) to fifteen (15) dollars. If you choose to participate, the exact way to compute your compensation will be described within the survey. Once the actual values for the forecasts you made are known, we will tally up your compensation and send you an e-mail to let you know where you can pick them up on campus. Once you have received your compensation, your name and e-mail address will be removed from the databases related to the experiment.
*Confidentiality*

For compensation purposes, we will ask you to provide us with your name and your preferred e-mail address: we need this information because your compensation is determined based on your answers to the survey. We may provide your name and preferred e-mail address to support staff so that these persons can give you your compensation. You will also need to complete a Participant Payment Disclosure Form (i.e., Subject Log) in order to be paid. Once you receive your compensation, we will delete your name and e-mail address from our records. After that, all other data will be stored indefinitely on a secure location on campus in a faculty member or graduate student computer.

*Participation*

Your participation is voluntary. You do not have to continue with this on-line survey and you may refuse to do so. If you refuse to continue, however, you cannot take part in this research study. You may completely withdraw from the study at any time without penalty. You also may choose to cease participation or skip any questions that you do not feel comfortable answering.

*Questions about the Research*

Please take as much time as you need to think over whether or not you wish to participate. If you have any questions about this study at any time, contact the Principal Investigator Justin Sydnor at 608-263-2138. If you are not satisfied with response of research team, have more questions, or want to talk with someone about your rights as a research participant, contact the Social and Behavioral Science Institutional Review Board at the University of Wisconsin-Madison: 310 Lathrop Hall, 1050 University Avenue, Madison, WI 53706, phone: 608-263-2320.

1. STATEMENT OF CONSENT

By entering the information below, you acknowledge that:
   * You have read the above information.
   * You have received answers to the questions you have asked.
   * You consent to participate in this research.
   * You are an American citizen.
   * You are at least 18 years of age.

Type your first and last names

Type your preferred email address (so that we can communicate to you the exact details of where and when you can pick up your compensation).

2. Do you agree to participate in this study?

* Yes
* No
Demographics Block

The following questions are about demographics.

What is your citizenship?
* United States of America
* Other

What is your gender?
* Male
* Female

What year were you born?

Are you a ___?
* Freshman
* Sophomore
* Junior
* Senior
* Graduate
* Other

What is your major? (If more than one choice may apply, pick the one you most enjoy.)

What is your current Grade Point Average?

On a scale of 1 (very comfortable) to 5 (having difficulty), how would you rate your mathematical abilities?

Football Trivia

The following questions are football trivia.

Which team won the last Super Bowl (played in February 2012)?
* Cleveland Browns
* Green Bay Packers
* New York Giants
* University of Wisconsin-Madison Badgers
* New England Patriots

Who is a quarterback for the Green Bay Packers?
* Phil Esposito
* Aaron Rodgers
* Tom Brady
* Peyton Manning
* Tim Tebow

Which of these players is a defensive end who, in the 2011-2012 season, was a member of the Super Bowl winning team, went to the Pro Bowl, and lead his team for the number of sacks in the season?

* Jason Pierre-Paul
Familiarity with Football and Fantasy Sports

In the following questions, you will be asked about your familiarity with football and fantasy sports.

On a scale of 1 to 5 (where 1 is very familiar and 5 is very unfamiliar), how would rate your own familiarity with fantasy sports?

1 (very familiar)  2  3  4  5 (very unfamiliar)

On a scale of 1 to 5, how would rate your own familiarity with football?

1 (very familiar)  2  3  4  5 (very unfamiliar)

On average, how many days per week do you watch or read sports news?

1  2  3  4  5  6  7

Do you currently have a fantasy football team?

Yes  No

If you do have a fantasy football team, please rate your ability at fantasy football from 1 (very good) to 5 (very poor).

1 (very good)  2  3  4  5 (very poor)
Task Description and Compensation Scheme

This page describes the task we ask you to complete in this survey. We will also describe the exact formula that will be used to compute your compensation.

Task Description:

For the 30 or so quarterbacks that are expected to start the game in the upcoming weekend of NFL football, you will be asked to provide your forecast of the quarterback rating for the week for each of these quarterbacks.

Let us share with you some information about the statistic that we are asking you to forecast.

According to Wikipedia, passer rating <http://en.wikipedia.org/wiki/Passer_rating> is a measure of the performance of quarterbacks. Passer rating is calculated using each quarterback’s completion percentage, passing yardage, touchdowns and interceptions. A perfect passer rating in the NFL is 158.3. A perfect rating requires at least a 77.5% completion rate, at least 12.5 yards per attempt, a touchdown on at least 11.875% of attempts, and no interceptions.

Here are some facts about quarterback ratings for this season:
- the average for starting quarterbacks is about 90,
- on any given week, about half of the quarterback rating will be between 75 and 125, and
- about 75% of the time ratings fall within about 30 of the quarterback’s own season average.
You will be compensated for your participation based on how accurately you are able to forecast quarterback ratings this week. Here we describe the exact formula we will use to determine your compensation.

Your earnings in this experiment will be based on how accurate your predictions are. You can earn up to $0.50 for each prediction, for a total possible earnings of $15 if you manage to perfectly predict the rating of each quarterback this week. Any prediction that is off by more than 25 points will earn you no money for that prediction. So if you miss each prediction by 25 points or more, you will earn nothing in the experiment. Your goal here should simply be to try to make each prediction as accurately as you think possible.

The following table describes the compensation scheme. For the cases in between, the compensation will be obtained by interpolating between the values in the following table.

<table>
<thead>
<tr>
<th>Error</th>
<th>Your Compensation ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+/- 30</td>
<td>0.00</td>
</tr>
<tr>
<td>+/- 25</td>
<td>0.00</td>
</tr>
<tr>
<td>+/- 20</td>
<td>0.10</td>
</tr>
<tr>
<td>+/- 15</td>
<td>0.20</td>
</tr>
<tr>
<td>+/- 10</td>
<td>0.30</td>
</tr>
<tr>
<td>+/- 5</td>
<td>0.40</td>
</tr>
<tr>
<td>0</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Finally, your total compensation will be the sum of your compensation for each of your individual forecasts.

Suppose that you have made the following forecasts with the attached actual values, what would be your compensation?

<table>
<thead>
<tr>
<th>Player*</th>
<th>Your Forecast*</th>
<th>Actual QB Rating*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quarterback #1</td>
<td>100</td>
<td>133</td>
</tr>
<tr>
<td>QB #2</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>QB #3</td>
<td>90</td>
<td>95</td>
</tr>
</tbody>
</table>

* 0.1
* 0.3
* 0.5
* 0.7
* 0.9

The correct answer was 0.7.
Sensitize Info Set

In the coming pages, we'll ask to make two sets of predictions. The survey will randomly select what extra information, if any, you'll be provided with to make your forecasts. The extra information will be described at the top of the page.

Forecasts - Treatment 1

You will now be asked to make your forecast for half of the quarterbacks.

For each of the listed quarterback below, please enter your forecast for their quarterback rating in the coming game.

You will find in brackets, first, their own team, second, the team they are playing against.

_*You will also find the overall average of the quarterback rating for the starting quarterbacks (OAAvg)._*

* Tom Brady (Patriots, @49ers) _OAAvg_: 85.8

(...)

Forecasts - Treatment 1 - Retrospective Confidence

1 (very confident) 2 3 4 5 (not confident at all)

How confident do you feel about the credibility and relevance of the information you were provided with for the first series of predictions?

1 (highly confident in credibility and relevance) 2 3 4 5 (not at all confident in credibility and relevance)

We would be interested in learning about the strategies you used in making your predictions. Feel free to tell us about the strategies that supported your prediction choices.

Sensitize 2

In the next page, we'll ask to make the second set of predictions. The survey will randomly select what extra information, if any, you'll be provided with to make your forecasts. The extra information will be described at the top of the page.

_*Note that this extra information may *(or may not)* differ from the extra information you were provided with for the first selection*_.

(...)
Forecasts - Treatment 2

You will now be asked to make your forecast for the other half of the quarterbacks.

For each of the listed quarterback below, please enter your forecast for their quarterback rating in the coming game.

You will find in brackets, first, their own team, second, the team they are playing against.

*You will also find the overall average of the quarterback rating for the starting quarterbacks (OAAvg).

* Jay Cutler (Bears, @Packers) *OAAvg* : 85.8

Forecasts - Treatment 2 - Retrospective Confidence

How confident do you feel about your second series of predictions?

1 (very confident) 2 3 4 5 (not confident at all)

How confident do you feel about the credibility and relevance of the information you were provided with for the second series of predictions?

1 (highly confident in credibility and relevance) 2 3 4 5 (not at all confident in credibility and relevance)

We would be interested in learning about the strategies you used in making your predictions. Feel free to tell us about the strategies that supported your prediction choices.
Ex Post Demographics - MBTI

The following questions are aimed at facilitating the understanding of your personality type.

Below, you will find two descriptions. Select the set that best corresponds to you.

*A*
- I Make decisions objectively
- I appear cool and reserved
- I am most convinced by rational arguments
- I am honest and direct
- I value honesty and fairness
- I take few things personally
- I am good at seeing flaws
- I am motivated by achievement
- I argue or debate issues for fun

*B*
- I decide based on my values & feelings
- I appear warm and friendly
- I am most convinced by how I feel
- I am diplomatic and tactful
- I value harmony and compassion
- I take many things personally
- I am quick to compliment others
- I am motivated by appreciation
- I avoid arguments and conflicts

A    B
Below, you will find two descriptions. Select the set that best corresponds to you.

*A*

* I focus on details & specifics
* I admire practical solutions
* I notice details & remember facts
* I am pragmatic
* I live in the here-and-now
* I trust actual experience
* I like to use established skills
* I like step-by-step instructions
* I work at a steady pace

*B*

* I focus on the big picture & possibilities
* I admire creative ideas
* I notice anything new or different
* I am inventive
* I think about future implications
* I trust my gut instincts
* I prefer to learn new skills
* I like to figure things out for myself
* I work in bursts of energy

A  B

Below, you will find two descriptions. Select the set that best corresponds to you.

*A*

* I have high energy
* I talk more than listen
* I think out loud
* I act, then think
* I like to be around people a lot
* I prefer a public role
* I can sometimes be easily distracted
* I prefer to do lots of things at once
* I am outgoing & enthusiastic

*B*

* I have quiet energy
* I listen more than talk
* I think quietly inside their heads
* I think, then act
* I feel comfortable being alone
* I prefer to work "behind-the-scenes"
* I have good powers of concentration
* I prefer to focus on one thing at a time
* I am self-contained and reserved

A  B

Below, you will find two descriptions. Select the set that best corresponds to you.

*A*

* I like to have things settled
* I take responsibilities seriously
* I pay attention to time & am usually prompt
* I prefer to finish projects
* I work first, play later
* I seek closure
* I see the need for most rules
* I like to make & stick with plans
* I find comfort in schedules

*B*
* I like to keep my options open
* I am playful and casual
* I am less aware of time and may run late
* I prefer to start projects
* I play first, work later
* I may have difficulty making some decisions
* I question the need for many rules
* I like to keep plans flexible
* I want the freedom to be spontaneous

Ex Post Demographics - Other

For the following questions, you are asked to find the applicability of the statement to you on a scale of 1 ("I'm very much like that") to 5 ("I am not at all like that").

My friends would say I am cautious.

1 ("I'm very much like that") 2 3 4 5 ("I am not at all like that")

Being financially cautious is important to me.

1 2 3 4 5

I like statistics.

1 2 3 4 5

I like SPORT's statistics.

1 2 3 4 5

I enjoy reading the sports news.

1 2 3 4 5
I enjoy reading about sport statistics.

1 2 3 4 5

I quite often feel that things are set in my life and I can’t change them.

1 2 3 4 5

I’m aware that, while I can’t always control what happens around me, I do control my own reaction to said events.

1 2 3 4 5

I believe that when people find themselves in bad situations, it’s usually due more to unlucky circumstances.

1 2 3 4 5
5. REFERENCES


Biographies of the Authors

Marc-André Desrosiers is a Property/Casualty actuary working for a large insurer in Canada as a Research and Development project manager. He is interested in all aspects of ratemaking.