Framing effects in evaluation of accuracy of others' predictions

Saiwing Yeung (saiwing.yeung@gmail.com)
Institute of Education, Beijing Institute of Technology, China

Abstract

Most predictions can be partitioned into two components: the predicted outcome, and the chance that one considers the outcome will happen. We studied how people evaluate predictions with binary outcomes. These predictions can be conveyed in two equivalent ways: one predicting an outcome with some probability, and the other predicting the other outcome with the probability of the complement of the first outcome. Although these two ways of stating the predictions are mathematically interchangeable, we hypothesized that people would judge the congruently stated prediction, one that has the same qualitative component as the actual outcome, as more accurate. We tested this hypothesis in four experiments. Results suggested that this effect is consistent across a number of domains; depends on the frame in which the prediction is stated; is robust regardless of whether the ratings were elicited in positive or negative terms; holds for both rating and choice tasks.

Keywords: framing effects; probabilistic judgment; decision making.

Probabilistic predictions are frequently encountered in everyday life. For example, weather forecasts are often made in probabilistic terms (e.g. "chance of rain is 80%"). By comparing these statements against the actual outcomes, we can assess the predictors' skills at predicting these events. It is important to be able to accurately evaluate other people's predictions because it would then allow us to learn how good the predictors are in making predictions, to judge whether or to what degree should we trust the predictions, and to make decisions accordingly. For example, if a certain investment analyst predicts that there is a 99% chance that Acme Company will declare bankruptcy, and that we consider this analyst to be a good predictor, then it would be advantageous to sell stocks of Acme Company that we are holding.

In this paper, we focus on one particular aspect of evaluating predictions — how framing of predictions affect people's evaluations. Framing effect is an extremely well-researched topic and has led to numerous scholarly work. It refers to a phenomenon in which people's judgment, decisions, and actions are influenced by frames, or presentation of information and its context.

Framing effects have been found to influence people in various ways in different contexts. Levin, Schneider, and Gaeth (1998) proposed a typology that categorized them into three main types. The first type, risky choice framing effect, induces a choice reversal effect between two logically equivalent gambles (Tversky & Kahneman, 1981). In a prototypical setup, participants see one of the two gambles: either choosing between a sure gain and a risky gain, or choosing between a sure loss and a risky loss. Previous research has found that a majority of the people would prefer the sure gain choice in the gain condition, and risky loss choice in the loss condition.

The second type of framing effects was called attribute framing effects, as a single attribute within a given context presented in two logically equivalent frames has been shown to change people's evaluations about the subject. For example, in Levin and Gaeth (1988), beef that was labeled as "75% lean" was rated as better tasting and less greasy than beef that was labeled as "25% fat."

Goal framing effects is the third type in Levin et al.'s typology. Here negatively framed messages are found to be more persuasive than positively framed messages. Works by Meyerowitz and Chaiken (1987) demonstrated a typical setup of this problem. They found that women are more likely to perform breast self-examination (BSE) if they are told of the negative consequences of not performing BSE, compared to being told of the positive consequences of performing one.

In the present study we report a new type of framing effect, in which people's evaluation of a prediction with respect to the outcome is influenced by the frame in which the prediction is presented. We will focus on predictions in which there are clearly two possible outcomes (e.g. coin flips) and are stated with the subjective probability of said event happening (e.g. "80%"). Because there are exactly two outcomes, any predictions can be stated in two ways that are logically equivalent. For example, to say that there is a 99% chance that the world will be destroyed at end of 2012 is equivalent to a 1% chance that the world will not be destroyed at end of 2012.

We argue, however, that people evaluate these predictions differently. As demonstrated by the framing effects literature described earlier, people's judgments are often influenced by how information is presented. In the context of prediction evaluation, we suggest that people would overweight the qualitative component of the prediction (the stated outcome), relative to its quantitative component (the chance that one considers the outcome will happen). To differentiate this from previously discovered types of framing effects, we will call this *probabilistic statement framing effect* (PSFE). We will next describe four experiments that were carried out to investigate this hypothesized effect.

Pilot Experiment

The main objective of the Pilot Experiment was to establish initial evidence about PSFE. To ensure the realism of the stimulus, we used a cover story about the 2012 U.S. presidential election which had just ended a few weeks prior.

Methods

The participants were recruited using Amazon Mechanical Turk (MTurk). Only workers who were residing in the U.S., were at least 18 years old, and had a lifetime acceptance rate with MTurk of 95% or over were allowed to participate[1].

[1]The same requirements applied to all experiments in this paper. Moreover, we disallowed participants from participating in more than one experiment in this paper (except for two participants who

In order to detect participants who might have been bored or inattentive during the experiment, an attention check (AC) was employed in the experiment (Oppenheimer, Meyvis, & Davidenko, 2009). The AC took place before the actual experiment, and consisted of a paragraph of instruction followed by a question. The instruction began by asking participants to enter their favorite sports in the space below. However, at the end of the instruction we asked the participants to enter a different response: "To show that you have read this far, please enter candle below. To repeat, enter the word candle no matter what your favorite sports is." If participants had read the entire instruction, then they should have responded with the target word ("candle"). The other experiments in this paper employed AC's with exactly the same format with the exception of different target words.

The key content of the experiment would be next. The instructions were as follows, with the conditions marked by parentheses, and the differences between conditions marked by double brackets (⟦ and ⟧) and vertical lines (|||):

> Acme inc. is a company that conducts public opinion polls about the 2012 presidential election between Barack Obama and Mitt Romney. Before the election it had predicted that ⟦ (congruent) Mitt Romney had a 20 percent chance of winning ||| (incongruent) Barack Obama had an 80 percent chance of winning⟧.

All participants were then asked "If Romney had won, was Acme inc. wrong?" in a forced-choice question. The two conditions in this experiment represented the different ways in which predictions were framed. In the *congruent* condition, the qualitative component of the prediction was the identical to the hypothetical result stated in the stimuli (Romney winning), whereas it was the opposite in the *incongruent* condition.

We then asked participants to rate the prediction using a 9-point Likert scale on "How accurate was the prediction?" and "How useful was the prediction?" The participants then filled in a demographics survey, which included a question about their political orientation.

Results

There was a total of 93 responses. Eleven of them failed the attention check question and their data were discarded. Out of the resulting 82 data points, 56.1% were female, 81.7% had at least some college education. We recorded age information in brackets. Almost half of the participants were in the youngest bracket of under 25 (48.8%), but there were also significant portion of the participants in older brackets (22.0% between 26 and 35; 18.3% between 36 and 50; 11.0% 51 or over).

We first examined the forced-choice question on whether the participants regarded the prediction as wrong. Relatively fewer participants in the congruent group rated the prediction as wrong ($16/40 = 40\%$) than in the incongruent group ($22/42 = 52.4\%$). However, the differences were not significant ($\chi^2(1, N = 82) = 1.26$, $p = 0.26$, $\phi = 0.12$).

The congruent group rated prediction accuracy ($M = 4.58$, $s.d. = 2.21$) significantly higher than the incongruent group did ($M = 3.17$, $s.d. = 1.83$; $t(80) = 3.15$, $p < 0.01$, Cohen's $d = 0.71$). The congruent group also rated prediction usefulness ($M = 4.30$) higher than the incongruent group did ($M = 3.45$), although the difference was only marginally significant ($t(80) = 1.75$, $p = 0.09$, Cohen's $d = 0.39$).

As the stimuli in this experiment involved a question in politics, we also tested whether subjects' political orientation influenced their responses. There were more self-reported Democrats than Republicans, with 22 (26.8%) self-identified as strongly Democrat and 35 (42.7%) as moderately Democrat. Nonetheless, the participants' political orientations had a low correlation with their evaluation of accuracy at $r = 0.084$ and was insignificant ($t(80) = 0.75$, $p = 0.45$).

Discussion

In this experiment we found initial evidence supporting PSFE: Participants in the congruent frame rated the prediction as more accurate, although they did not consider the prediction less wrong. Nonetheless, there remains a number of unresolved issues. The two conditions represent differences at multiple attributes, including the prediction frame (whether the prediction was described in terms congruent with the actual result), framing of the result (whether the results were described using the same agent as the prediction frame), and valence of the evaluation (whether the evaluation is elicited in positive or negative terms). It remains to be established which of these attributes underlie this phenomenon. Moreover, the scenario used was based on a real event and this might have interfered with people's reasoning, especially because most of our participants self-identified as liberal. Therefore, we conducted the next two experiments to tease apart the pathways involved in bringing about this phenomenon.

Experiment 1

The first objective of Experiment 1 was to investigate PSFE using an artificial cover story in which the participants do not have a preference towards one of the two possible outcomes. The second objective was to investigate whether PSFE is driven by the prediction frame or result frame.

Methods

Participants were again recruited from MTurk. The experiment used a between-subject 2×2 design, crossing the prediction frame and the result frame. The stimuli in this experiment used the cover story of a college (American) football game. The stimuli were as follows:

> Imagine that you have just arrived a little early for a new class on the first day of the semester. Another student was already there. The two of you started talking and the conversation turned to an upcoming college football game between universities A and B. The other student predicted that ⟦ (Prediction frame: congruent) University B has a 30% chance of winning ||| (Prediction frame: incongruent) University A has a 70% chance of winning ⟧.

participated in two experiments because of a programming error). This ensures a broader representativeness of our samples.

The game took place later that week and ⟦ (Result frame: A) University A lost to University B ||| (Result frame: B) University B defeated University A⟧.

The conditions in the two prediction frames are so named because if we ignore the confidence levels in the predictions, the prediction in the congruent prediction frame (B winning) is congruent with the result (B won in all conditions in this experiment), while the prediction in the incongruent prediction frame (A winning) is incongruent. The conditions in the result frames are simply named after the agent in the frame.

The participants were then asked to state whether the predictions wrong, and how accurate was the prediction (9-point Likert scale). Finally the participants answered a demographics survey similar to that in the Pilot.

Results

There were a total of 112 participants (41.1% female), after discarding data from eight others for failing the attention check (6.7%). Average age was 28.62 (*s.d.* = 11.95) and 84.8% had at least some college education.

The main objective of Experiment 1 was to investigate whether the prediction frame or the result frame is driving the PSFE, and whether there is interaction. To examine the effect of the prediction frame, we performed a *t*-test to compare the evaluation of prediction accuracy between the two prediction frames. The mean rating in the congruent prediction frame was 4.91 (*s.d.* = 1.79), higher than that of the incongruent prediction frame at 3.34 (*s.d.* = 1.47), and the difference was significant ($t(110) = 5.08$, $p < 0.01$, Cohen's $d = 0.97$). This replicated the results from the Pilot.

In the Pilot, there was no significant difference between the two conditions in whether participants consider the predictions were wrong. Interestingly, this was significant in Experiment 1, in which 12 of 56 (21.4%) participants in the congruent condition judged the prediction as wrong, compared to 29 of 56 (51.8%) of those in the incongruent condition did so ($\chi^2(1, N = 112) = 11.12$, $p < 0.01$, $\phi = 0.32$).

One alternative hypothesis is that the differences were caused by the different result frames. We found the mean accuracy ratings to be 4.25 (*s.d.* = 1.96) for frame A and 4.00 (*s.d.* = 1.66) for frame B, respectively. There were no significant difference ($t(110) = 0.74$, $p = 0.46$, Cohen's $d = 0.14$). For the question on prediction wrong-ness, there were no significant difference between different result frames either ($\chi^2(1, N = 112) = 0.20$, $p = 0.66$, $\phi = 0.04$). Moreover, there were no interaction between prediction framing and result framing ($F(1) = 0.32$, $p = 0.57$, $\eta^2 = 0.00$). Figure 1 plots the results from Experiment 1.

Another alternative hypothesis is that having the same agent in the prediction frame and result frame would lead to higher accuracy ratings. We found this to not be the case. Mean accuracy ratings for participants who had the same agent in both frames was lower (4.05, *s.d.* = 1.69) than those with different agents (4.20, *s.d.* = 1.94), and the differences were not significant ($t(110) = 0.43$, $p = 0.67$, Cohen's $d = 0.08$). There were no significant differences in the forced-choice question either ($\chi^2(1, N = 112) = 0.00$, $p = 0.96$, $\phi = 0.01$).

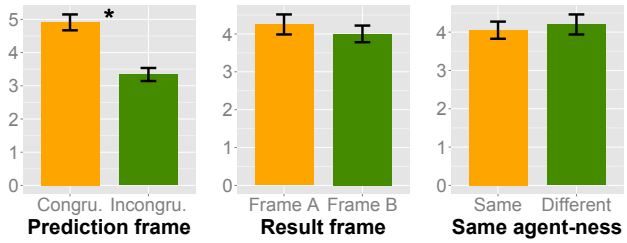

Figure 1: Results of Experiment 1. Each graph plots a comparison of accuracy ratings for a different factor. Error bars represent *s.e.*

Self-reported football knowledge was evenly spread over the 4-point scale. There were 32, 33, 23, and 24 responses, from the least knowledgeable to the most knowledgeable. To investigate whether there is an interaction between football knowledge and prediction frame, we carried out an ANCOVA analysis. The results indicated that there was no significant interaction ($F(1) = 0.696$, $p = 0.41$).

Discussion

Results of Experiment 1 suggested that framing of predictions significantly changes people's evaluation of predictions, whereas framing of results and whether the same agent is used in both frames has little effect. This not only replicated the results of the Pilot, but also suggested that prediction frame is what underlies the difference in how people evaluate how accurate predictions are. The results of this experiment correspond to the compatibility effects (Slovic, Griffin, & Tversky, 1990), which states that stimuli attribute that is compatible with the response mode would be overweighted.

Experiment 2

In both the Pilot and Experiment 1, the forced-choice question on evaluations were elicited in negative terms, i.e. we asked the participants whether the predictions were wrong. Therefore in Experiment 2 we tested whether PSFE also holds when the evaluations had a positive valence.

Methods

Participants were again recruited from MTurk and the procedures were mostly the same as the previous two experiments. The instructions were:

> Imagine that you have just arrived a little early for a new class on the first day of the semester. Another student was already there. The two of you started talking and the conversation turned to an upcoming college football game between universities A and B. The other student predicted that ⟦ (congruent) University B has a 30% ||| (incongruent) University A has a 70% ⟧ chance of winning.
>
> The game took place later that week and ⟦ (congruent) University B defeated University A ||| (incongruent) University A lost to University B ⟧.

As can be seen from the instructions, there were two conditions: congruent and incongruent. The major departure of Experiment 2 from the previous two lies in how we elicited the forced-choice response on about the prediction: we asked "Was the prediction made by the other student right?"

Note that in both condition, the agent remains the same in both the prediction frame and the result frame.

Results

Experiment 2 had a total of 78 participants (34.6% female), after discarding data from nine of them for failing the attention check (10.3%). Mean age was 29.03 ($s.d. = 12.73$) and 88.5% had at least some college education.

The main objective of this experiment was to test whether PSFE could be replicated when evaluations were elicited in positive terms. We first analyzed results of the forced-choice question in which the participants were asked whether the prediction was right. In the congruent condition, 22 of 40 (55.0%) responded affirmatively; whereas in the incongruent condition, 8 of 38 participants (21.1%) responded affirmatively. χ^2-squared test showed that the difference was significant ($\chi^2(1, N = 78) = 9.49,\ p < 0.01,\ \phi = 0.35$). The quantitative accuracy ratings for the two conditions reflected a similar picture. The mean accuracy rating for the congruent condition was 5.05 ($s.d. = 2.06$), compared to that of the incongruent condition of 3.39 ($s.d. = 1.72$). The difference was significant ($t(76) = 3.84$, $p < 0.01$, Cohen's $d = 0.88$).

Discussion

Experiment 2 focused on whether PSFE holds when the people are asked to evaluate the predictions in positive terms. Results indicated that this is indeed the case, suggesting that PSFE to be robust regardless of the valence in which evaluations were elicited.

Experiment 3

The first three experiments in this paper demonstrated that when people give accuracy ratings to predictions, predictions presented in a congruent frame as the actual result would be rated as more accurate. Experiment 3 investigated whether this phenomenon could be extended to choice tasks — when the two frames (congruent and incongruent) are presented at the same time as two choices and people are asked to judge which one is the more accurate one.

We also tested two factors that might shed light on the mechanism of PSFE. First, one potential reason that people rated predictions in the incongruent condition as less accurate might have been that the quantitative components of these predictions involve higher numerical probabilities (compared to those in the congruent condition), and this might have been perceived as being overconfident, which in turn led to participants down-adjusting their accuracy ratings. Second, many prior works have suggested that numeracy plays an important role in judgment and decision making. For example, Peters et al. (2006) found that participants who are higher in numeracy are less susceptible to attribute framing effects. To investigate the influences of these two factors, we also assessed perception of overconfidence and participants' numeracy.

Methods

Similar to the previous three experiments, all participants were recruited through MTurk. However, because this experiment is slightly longer than the previous three, we increased the reward from US$0.15 to US$0.20.

In the previous experiments, predictions in the two frames were given logically equivalent probability estimates (e.g. 75% vs. $100\% - 75\% = 25\%$). However, in Experiment 3 the participants would see both frames side-by-side, and therefore such a setup might seem contrived. Moreover, we wanted to test whether the congruent frame would be favored even when it is *logically inferior*. Hence we parameterized the congruent frame with a probability estimate of 15% (in the direction of the actual result), and the incongruent frame with 80% (opposite the direction of the actual result). The congruent frame is now logically superior because it predicts the outcome that turns out to be correct with $100\% - 80\% = 20\%$ confidence, compared to 15% in the congruent frame. Additionally, in order to make the scenarios more realistic, we added two detractor predictions to each option. The instruction for one of the conditions was as follows:

> Imagine that you are an analyst at an investment firm. Currently you are evaluating predictions made a year ago by two of your subordinates concerning a technology company called Acme Corp.
>
> Analyst A predicted that in the coming year:
>
> - Acme would buy out their supplier SuperTech Company.
> - Acme would expand into the European Union.
> - There was an 80% chance that Acme would become a public company.
>
> Analyst B predicted that in the coming year:
>
> - Acme would license crucial technology patents from their competitor CompX Company.
> - Acme would build another manufacturing plant within the U.S.
> - There was a 15% chance that Acme would not become a public company.

The probabilistic prediction shared by both analysts was whether Acme would become public or not. Each of the two analysts also made two detractor predictions additionally.

The participants then read about what actually happened. There were five total predictions: two unique detractors from each analyst, plus the common prediction. In all conditions, Acme would not become public. However, one of the two detractors from each of the analysts would come true.

In this counter-balance condition shown above, Analyst A predicted that there was an 80% chance of the target event (Acme became a public company) happening. Analyst B, in contrast, predicted that there was a 15% chance of the target

event not happening. If probabilistic statements could be inverted algebraically, it would mean Analyst B predicted that there was an 85% chance of the target event happening. As the target event did not happen, Analyst A should be evaluated as being more accurate, if prediction frames have no influence on people's judgment.

There were two counter-balancing conditions. First, the order of the congruent and incongruent options was randomized between subjects. Second, the detractors that came true were counter-balanced. For roughly half of the participants the supplier buy out and new U.S. manufacturing plant turned out to be true, while for the other half it was the opposite.

We then asked participants "Which analyst do you think made the better predictions?" and "Which analyst do you think was more confident about the predictions?" This was followed by a memory test. We asked the participants to indicate whether each of the five events happened in the actual outcome. Then to investigate the influence of participants' numeracy on their judgments, we added the 8-item abbreviated numeracy scale from Weller et al. (2012). After the numeracy section, participants answered a few demographics questions, including two questions about their level of knowledge concerning stock trading and technology.

Results

There were a total of 85 participants (60% female; one declined to self-identify). We discarded data from 29 (25.4%) participants: 27 for failing the AC and 2 for leaving over 80% of the answers blank[2]. Mean age was 33.1 ($s.d. = 12.87$) and 87.1% had at least some college education.

The portion of workers who failed the AC was higher than the previous experiments. We ran a 4 (experiment) $\times$ 2 (number of AC pass/failure) χ^2-squared test of independence and the results were significant ($\chi^2(3, N = 414) = 19.33$, $p < 0.01$, $\phi = 0.22$). However, there was no a priori reason to suspect that the workers in this experiment were different from those in the previous ones. In all four experiments, the AC was the second question in the entire experimental procedure, after only the question that elicited their MTurk ID. Therefore up to the AC, the experimental procedures of all four experiments were essentially the same. The monetary reward was the only difference between this experiment (US$0.20) and the previous ones (all three at US$0.15). However, Mason and Watts (2009) have found that financial incentives do not significantly impact the quality of MTurk experiments, even for amounts that differ by as much as 10 times. To further confirm the quality of the data, we checked the result of the memory test. The range of the memory score was from 0 to 5 (remembered perfectly). The mean memory score across all participants was 4.25, indicating that the participants remembered the details of the experiment well. Hence, we attribute the high AC failure rate to coincidence.

The main objective of this experiment was to test whether the PSFE could be extended to a choice task. More participants (56; 65.9%) chose the analyst in the congruent condition (15%) as more accurate, compared to the one in the incongruent (80%) condition (29; 34.1%). A χ^2-squared test indicated that it was significantly different from chance ($\chi^2(1, N = 85) = 8.58$, $p < 0.01$, $\phi = 0.32$).

We then examined whether perception of overconfidence was related to PSFE. There were 35 (41.2%) and 50 (58.8%) participants who judged the congruent and incongruent option, respectively, as more confident. The result was close to reaching significance ($\chi^2(1, N = 85) = 2.65$, $p = 0.10$, $\phi = 0.18$). This suggests that perception of predictors' overconfidence might play a small part in this effect and deserves further investigation.

The order of presentation had a big effect on choice. In conditions where the incongruent option was presented first, there were about the same number of participants who chose the congruent option ($N = 21$) as those who chose the incongruent option ($N = 20$) as more accurate. However, if the congruent option was presented first, 35 (vs. 9) participants judged the congruent option as more accurate. The interaction was significant ($\chi^2(1, N = 85) = 7.58$, $p < 0.01$, $\phi = 0.30$). This suggests that order of presentation significantly influenced evaluation of accuracy. However, order of presentation did not have a significant effect on evaluation of confidence ($\chi^2(1, N = 85) = 0.24$, $p = 0.62$, $\phi = 0.05$). The other counter-balancing condition — which pair of distractors turned out to be correct — had no significant effect on evaluation of accuracy ($p = 0.37$) nor confidence ($p = 0.79$).

We also investigated the effect of numeracy on people's judgments. As there are eight questions in Weller et al.'s numeracy scale, the range of the numeracy scores is from 0 to 8. No participant answered the mammogram question correctly. In fact, no answer came within 3 percentage point of the correct answer. This is not surprising because this question has been found to be a very difficult question (see Weller et al., 2012). The percentage of participants who answered each question correctly (Table 1) was in fact quite close to the result obtained by Weller et al. (2012). This suggests that the numeracy and motivation of the participants in this experiment were comparable to those in their experiment. This result also partly mitigated the concern raised by the high percentage of participants failing the AC.

The mean (and *s.d.*) of the numeracy score for participants who chose the congruent or incongruent options as more accurate were 4.43 (1.45) and 5.41 (1.45), respectively. We fitted a logistic model using the numeracy score as the independent variable, and participants' choices as dependent variable. Results indicated that the influence of numeracy was significant ($\beta = 0.75$, $z = 2.74$, $p < 0.01$). This suggested that participants who were lower on numeracy are more likely to consider the analyst in the congruent option — the normatively less accurate of the two — the more accurate predictor.

The effect of self-reported knowledge about stock trading and technology on choice of more accurate prediction was not significant in either case ($p = 0.80$ and $p = 0.56$).

[2]No other participants left more than one of the non-demographic answers blank.

Table 1: Percentage of participants correctly answering each item of the numeracy scale in Experiment 3 (E3), compared to the results from Weller et al. (2012).

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
E3	0.0	27.1	45.9	57.6	80.0	81.2	96.5	88.2
W	11.0	26.5	39.8	42.1	60.4	70.2	73.8	84.8

Discussion

The key objective of this experiment was to investigate whether PSFE would hold in a choice task. We also put PSFE to a stronger test because the congruent option was presented vis-à-vis a logically superior option. Our results found that significantly more participants would choose the congruent option, suggesting the robustness of PSFE. We also found that perception of overconfidence did not explain PSFE. However, numeracy was found to be a moderating factor. Like Peters et al. (2006), we found that people who are higher in numeracy to be less susceptible to framing effects.

General Discussion

We proposed a new phenomenon, probabilistic statement framing effect (PSFE), that occurs when predictions made in congruent frames (relative to eventual outcomes) are judged as more accurate, compared to logically equivalent or even superior predictions made in incongruent frames. Across four experiments, we found that this effect holds regardless of real world based event (Pilot Experiment) or hypothetical events (Experiments 1 to 3), and rating (Pilot and Experiment 1 & 2) or choice (Experiment 3) task. The effect held even when the congruent option was logically inferior (Experiment 3). Finally, we found numeracy to be a moderating factor.

The results from these experiments suggest that a majority of people do not evaluate the goodness of predictions in a normative manner. They overweight the qualitative component of a prediction while underweighting its quantitative component. This is especially true for people who are low in numeracy. The findings in this paper might have important implications in domains such as personal finance, medical decision making, and corporate strategic planning.

Among the three major types of framing effects, PSFE might be most closely related to the attribute framing effects. However, we argue that it is distinct for one major reason. Levin et al. (1998) demonstrated that attribute framing effects occurs because positive frames evoke favorable associations in memory; and vice versa for negative frames. However, PSFE can favor evaluations of negative frames (e.g. losing a game in sports), as long as the predictions are congruent to the outcome. This cannot be explained using the above framework and therefore we suggest that PSFE should be regarded as a separate phenomenon.

Although the effect seems to be robust across a broad range of conditions, its causal mechanism and cognitive processes are not well understood. Moreover, prior research has suggested that important personal decisions are less influenced by frames (Marteau, 1989). We are currently examining what roles information leakage (Sher & McKenzie, 2006), selective attention (Levin, 1987), and encoding of information (Levin & Gaeth, 1988), might play in relation to this effect.

All experiments here have been carried out through MTurk. This enabled us to collect data from a subject pool more diversified than one that of a university sample. Moreover, MTurk has been found to be able to yield high quality data (Buhrmester, Kwang, & Gosling, 2011), and be able to replicate a number of classical findings (Crump, McDonnell, & Gureckis, 2013). However, it might be interesting in the future to study this phenomenon in lab-based and field studies.

The findings in this paper demonstrate the psychological impact of prediction frames on how people evaluate predictions with respect to outcomes. When predictions are described in congruent frames as the eventual result, people consider them as more accurate than if they were described in incongruent frames. This observation is not captured by the previous literature on framing effects and highlights the need for a better understanding of the processes that underlies this phenomenon.

Acknowledgments. We thank the very helpful comments from four anonymous reviewers.

References

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, *8*(3), e57410.

Levin, I. P. (1987). Associative effects of information framing. *Bulletin of the Psychonomic Society*, *25*(2), 85–86.

Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of Consumer Research*, *15*(3), 374–378.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*(2), 149–188.

Marteau, T. M. (1989). Framing of information: Its influence upon decisions of doctors and patients. *British Journal of Social Psychology*, *28*(1), 89–94.

Mason, W., & Watts, D. J. (2009). Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD workshop on human computation* (Vol. 15, pp. 77–85).

Meyerowitz, B., & Chaiken, S. (1987). The effect of message framing on breast self-examination attitudes, intentions, and behavior. *Journal of Personality and Social Psychology*, *52*(3), 500–510.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*(5), 407–413.

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*(3), 467–494.

Slovic, P., Griffin, D., & Tversky, A. (1990). Compatibility effects in judgment and choice. In *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 5–27). University of Chicago Press.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2012). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*. doi: 10.1002/bdm.1751