Title
Like Schrödinger's Cat, the Impact Bias Is Both Dead and Alive: Reply to Wilson and Gilbert (2013)

Journal
Journal of Personality and Social Psychology, 105(5)

ISSN
0022-3514

Authors
Levine, Linda J
Lench, Heather C
Kaplan, Robin L
et al.

Publication Date
2013-11-01

DOI
10.1037/a0034340

Peer reviewed

Journal of Personality and Social Psychology
2013, Vol. 105, No. 5, 749–756

0022-3514/13/$12.00 DOI: 10.1037/a0034340

REPLY

Like Schrödinger's Cat, the Impact Bias Is Both Dead and Alive: Reply to Wilson and Gilbert (2013)

Linda J. Levine
University of California, Irvine

Heather C. Lench
Texas A&M University

Robin L. Kaplan
University of California, Irvine

Martin A. Safer
The Catholic University of America

In their comment on our article on affective forecasting (Levine, Lench, Kaplan, & Safer, 2012), Wilson and Gilbert (2013) criticized the meta-analysis, proposed alternative explanations for the empirical studies, and concluded that the impact bias is alive and well. Our reply demonstrates that, irrespective of the exclusion of effects and selective recoding of effects recommended for the meta-analysis, the pattern of results remains the same: Study participants' forecasts are more accurate when they report their feelings about a focal event, or immediately after a focal event, than when they report their feelings in general after a delay. New analyses rule out individual differences and focalism as alternative explanations for the results of our empirical studies. These studies show that people can accurately predict the intensity of their feelings about events. People overestimate in predicting the impact of events on their emotional state in general, but clarifying the meaning of the forecasting question reduces the magnitude of this bias. We conclude that the impact bias, which encompasses overestimating the intensity of feelings about events and overestimating the intensity of feelings in general, is both dead and alive. The importance of predicting feelings about events for decision making and the reasons people predict some features of emotion more accurately than others are discussed.

Keywords: affective forecasting, impact bias, intensity bias, emotion, prediction

Supplemental materials: http://dx.doi.org/10.1037/a0034340.supp

> There is a difference between a shaky or out-of-focus photograph and a snapshot of clouds and fog banks.
>
> —Erwin Schrödinger

Research on affective forecasting shows that people have a foggy view of the impact future events will have on their emotions. They often overestimate in anticipating the intensity, frequency, and duration of their feelings, collectively referred to as the impact bias (e.g., Gilbert, Driver-Linn, & Wilson, 2002). But are people really faced with clouds and fog banks when they try to foresee how events will make them feel? Or does the procedure commonly used in forecasting studies obscure features of emotion that people can accurately predict? This was the issue at the heart of our article on accuracy and artifact in affective forecasting research (Levine, Lench, Kaplan, & Safer, 2012) and Wilson and Gilbert's (2013) comment on our article. In this reply, we respond to their critique and argue that alternative research procedures present a clearer picture of the features of emotional experience that people can and cannot accurately predict.

Foggy Forecasting or Poor Measurement? The Levine et al. (2012) Argument

Levine et al. (2012) acknowledged that research on affective forecasting captures important truths about people's emotional experience. Events often elicit short-lived emotional responses that are followed by relatively quick adaptation. Over time, people's goals and expectations change, other events capture their attention, and they think about emotional events less often. People often fail to anticipate how quickly they will adapt (adaptation neglect), and they expect to think about events more than they actually do (focalism). As a result, they tend to overestimate the impact events will have on their overall emotional well-being (e.g., Gilbert,

Linda J. Levine, Department of Psychology and Social Behavior, University of California, Irvine; Heather C. Lench, Department of Psychology, Texas A&M University; Robin L. Kaplan, Department of Psychology and Social Behavior, University of California, Irvine; Martin A. Safer, Department of Psychology, The Catholic University of America.

Correspondence concerning this article should be addressed to Linda J. Levine, Department of Psychology and Social Behavior, University of California, Irvine, 4568 Social and Behavioral Sciences Gateway, Irvine, CA 92697-7085. E-mail: llevine@uci.edu

Pinel, Wilson, Blumberg, & Wheatley, 1998; Wilson & Gilbert, 2008; Wilson, Wheatley, Meyers, Gilbert, & Axsom, 2000).

Levine et al. (2012) argued, however, that people can accurately predict the intensity of emotion they will experience when they are thinking about events. We also argued that the procedure commonly used to assess forecasting accuracy is misleading. It inflates the extent to which people overestimate the impact events will have their overall emotional well-being. The procedure in question is this: In most affective forecasting studies, people are asked to imagine that an event has occurred and to rate how they will feel in general after a specified period of time. After the event has occurred and the period of time has elapsed, people are asked to rate how they are feeling in general without reference to the focal event. People may overestimate the impact of events on their overall emotional well-being because, in the context of having just been asked to imagine a specific future event, they interpret the request to predict how they will feel in general as asking how they will feel when they are thinking about that event. If people believe they are being asked to predict one feature of emotion (their feelings about an event) but are later asked to report another (their feelings in general), inaccuracy is virtually guaranteed. Levine et al. noted that this is not focalism. People are displaying focalism when they expect their general emotional state to be dominated by their reaction to a focal event. So focalism does not involve misunderstanding the general forecasting question. It involves giving the wrong answer to a question that was correctly understood.

Consistent with this argument, our meta-analysis showed that study participants' forecasts were more accurate when they reported their feelings about, or immediately after, a focal event than when they reported their feelings in general after a delay. Our empirical studies showed that participants accurately predicted the intensity of emotion they would experience when they were later thinking about an event. They overestimated only when asked to predict how they would feel in general and later report their feelings without reference to the event. We also found that most participants misinterpreted requests to predict their feelings in general as asking how they would feel when they were thinking about the focal event. Clarifying the meaning of the question dramatically reduced the impact bias. Levine et al. (2012) acknowledged that people are not always accurate when they predict the intensity of their feelings about events. They sometimes overestimate (e.g., Green et al., 2013) and sometimes underestimate (e.g., Lench, Safer, & Levine, 2011; Van Boven & Loewenstein, 2003). Our findings show, however, that people can anticipate the intensity of their future emotions more accurately than is commonly portrayed in the affective forecasting literature and that a procedural artifact contributes to biased forecasts.

In their comment on our article, Wilson and Gilbert (2013) criticized the meta-analysis, proposed alternative explanations for the empirical studies, and concluded that the impact bias is alive and well. Below, we show that the conclusions we drew from our meta-analysis and empirical studies are valid. We discuss the importance of the findings for decision making and why people are better at predicting certain features of emotion than others. In light of the issues raised by Wilson and Gilbert, we also discuss directions for future research.

Validity of the Conclusions Drawn From the Meta-Analysis

Wilson and Gilbert (2013) claimed that in our meta-analysis, we (a) omitted studies that were inconsistent with our hypotheses, (b) wrongly coded whether questions about emotional experience were immediate or delayed, (c) collapsed across moderator variables meant to reduce the impact bias, and (d) miscoded the direction of effects in some key studies. Below, we address problems with the general approach taken in Wilson and Gilbert's critique of our meta-analysis. We also show that, irrespective of the exclusion of effects and selective recoding of effects that Wilson and Gilbert recommended, the results support the conclusions we drew from the analysis. We disagree with Wilson and Gilbert's arguments about the inclusion and coding of individual studies and present a detailed evaluation of each of these arguments in supplementary material that is available online.

Misrepresenting the Aim of the Meta-Analysis

Wilson and Gilbert (2013) presented a straw man argument by refuting claims we did not make. They argued that our meta-analysis (Levine et al., 2012) was "designed to show that the impact bias does not exist" (Wilson & Gilbert, 2013, p. 743) and that this bias "was in large part the result of a procedural artifact" (Wilson & Gilbert, 2013, p. 746). They concluded that, when so-called errors are corrected, these claims are not supported. Our meta-analysis does not, and was not intended to, provide evidence for the nonexistence of the impact bias or for a procedural artifact. Rather, as stated in our article,

> We expected the effect size representing intensity bias to be large when study participants were asked a general question about their emotional experience after a delay. We expected the effect size to be significantly smaller when study participants were asked a specific question about experienced emotion, either immediately or after a delay, or a general question immediately after the focal event. (Levine et al., 2012, p. 594)

The meta-analysis provided robust support for this hypothesis. It showed that the type of question researchers ask about experienced emotion moderates the magnitude of the intensity bias. Importantly, the meta-analysis could not speak to the mechanisms underlying this finding, which may include focalism, artifact, or a combination of these and other factors. Our empirical studies did address mechanisms and showed that a procedural artifact contributes to bias when people are asked to predict and report their feelings in general after a delay.

Study Inclusion, Exclusion, and Coding

Wilson and Gilbert (2013) claimed that we omitted studies from our meta-analysis that were inconsistent with our hypotheses, and they questioned the coding of studies we included. The strength of a meta-analysis lies in the ability to *systematically* select and code a large set of investigations to examine effects and moderators of those effects. In our meta-analysis, we followed standard recommended procedures for such analyses. We defined the scope of the analysis as an investigation of effect sizes associated with the difference between forecast and experienced emotional intensity across question types (specific, general) and question timing (im-

mediate, delayed). We then used this scope to define inclusion and exclusion criteria for research reports. For example, given the scope, we searched for reports using keywords related to affective forecasting and also searched the references of those reports. We included reports that contained information about forecast and experienced emotion and assessed emotion intensity. We developed a coding scheme for effects and moderators and applied it consistently across all studies. Each of these steps was described in our article as recommended in the American Psychological Association guidelines for reporting meta-analyses (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008, pp. 848–849; also see Johnson & Eagly, 2000; Shea et al., 2007).

Wilson and Gilbert's (2013) claim that we omitted and miscoded studies suggests that they would have used different procedures. But they did not state the criteria they would have used. Instead of developing and consistently applying an alternative inclusion or coding scheme (Cooper, 2010; Shea et al., 2007), they took a piecemeal approach. They identified studies outside our search criteria that they claimed agreed with their hypothesis and selectively recoded studies that disagreed with their hypothesis. We welcome additional systematic reviews but do not believe a piecemeal approach is likely to advance understanding of this literature. Indeed, a major criticism of narrative literature reviews is that they take precisely this approach by selectively reviewing studies to make a theoretical point, whereas meta-analyses avoid this problem (Johnson & Eagly, 2000).

Coding Studies as Immediate or Delayed

Our meta-analysis showed that study participants were far more accurate when they reported their emotional experience about, or immediately after, a focal event than when they reported their emotional experience in general after a delay. This suggests that people are pretty good at predicting how they will feel when an event comes to mind. Wilson and Gilbert (2013) objected that some studies showed impact bias, instead of accuracy, even though participants were asked how they were feeling in general after just brief delays (such as 2 min), filler questions, or filler procedures (e.g., Wilson, Meyers, & Gilbert, 2003, Study 1). They argued that it is nonsensical to code such questions as delayed rather than immediate because participants are clearly still thinking about the focal event.

This is a critical issue, not just for evaluating the validity of our coding scheme but also for clarifying the procedures necessary to assess people's ability to predict the intensity of emotion they will feel when they are thinking about events. Our coding reflects the fact that brief delays and other tasks can direct people's attention away from personally significant features of events (e.g., Kalisch, Wiech, Herrmann, & Dolan, 2006). For example, Gilbert et al. (1998, Study 5) had forecasters read a summary of a newspaper article about a child's death and predict how bad they would feel after reading the whole article. Experiencers read the entire article, answered filler questions about its nonemotional features ("How well written was the story?" and "To what extent did the story keep your attention?"), and then reported how they felt. The results showed an impact bias. Forecasters overestimated how bad they would feel relative to the reports of experiencers. Wilson and Gilbert (2013) suggested that procedures of this sort are equivalent to asking participants to report their emotional experience immediately after a focal event, whereas we coded such procedures as delayed.

To empirically assess the effect of brief delays on the impact bias, we conducted a similar study. Forecasters ($n = 94$) read a summary and predicted how they would feel after reading an entire article about a child's death. Three different groups of experiencers read the entire article and rated their feelings either immediately afterward ($n = 93$), after answering Gilbert et al.'s (1998) filler questions ($n = 92$), or after watching a 2-min neutral video ($n = 86$). Consistent with Gilbert et al.'s findings, forecasters significantly overestimated how bad they would feel ($M = 4.92$ on a 9-point scale, $SD = 2.03$) relative to the reports of experiencers after answering filler questions ($M = 3.41$, $SD = 2.05$), Tukey-Kramer adjusted $t(184) = 5.26$, $p < .001$, $g = .74$, or after a 2-min delay ($M = 2.93$, $SD = 1.70$), $t_{TK}(178) = 6.81$, $p < .001$, $g = 1.06$. However, immediately after reading the article, experiencers felt about as bad as forecasters had predicted, and no significant impact bias was found ($M = 4.33$, $SD = 2.03$), $t_{TK}(185) = 2.09$, $p = .16$, $g = .29$. These findings suggest that brief delays and procedures can increase the impact bias and make it difficult to assess people's ability to predict their immediate emotional response. Thus, our decision to code such questions as "delayed" in our meta-analysis was appropriate.

We agree, though, with Wilson and Gilbert's (2013) broader point that question timing provides an imperfect proxy for the true question of interest—whether people can predict the intensity of emotion that will be evoked by thinking about a focal event. The strongest evidence that people can comes from our empirical studies (Levine et al., 2012, Studies 1 and 3). When we asked participants directly to predict and report their feelings *about* a focal event, which ensured that they were thinking about that event when they reported their emotional experience, they were quite accurate.

Collapsing Across Variables Expected to Moderate the Impact Bias

Wilson and Gilbert (2013) objected to our decision to collapse across variables expected to reduce, eliminate, or reverse the impact bias. Collapsing across moderators is a matter of controversy in the greater meta-analytic literature (e.g., Cooper, 2010; Johnson & Eagly, 2000; Rosenthal, 1991). If moderators can be included with sufficient sample size to permit analysis, collapsing is not recommended. This occurs when a literature is large and moderators are theoretically derived and well represented by studies in the sample (Lepper, Henderlong, & Gingras, 1999). In the affective forecasting literature, however, very few studies investigate the same moderators. Because no overarching theory provided a way to group the various moderators, it was not possible to include them as variables in the meta-analysis. Another option was to omit these studies, but recommendations about this approach differ (e.g., Cooper, 2010; Lepper et al., 1999; Rosenthal, 1991). We chose instead to include as many studies as possible and to examine the size of the effects both with and without studies that had moderators expected to lessen the impact bias. We noted that the approximate size of the effects and the inferences drawn from the meta-analysis did not change if effects based on moderators were excluded (Levine et al., 2012, p. 595).

Wilson and Gilbert (2013) argued that the inferences drawn from the analysis do change if studies with moderators expected to lessen the impact bias are excluded. Their argument reflects inappropriate reliance on *p*-values associated with a small set of effect sizes and misunderstanding of the stated aim of our meta-analysis. In the condition with specific questions presented after a delay that they took issue with, the effect size we reported with all studies included, a *g* of 0.37, is considered of small to moderate size; their recalculated effect size with moderators excluded, a *g* of 0.53, is considered of moderate size. Wilson and Gilbert were concerned that a *g* of 0.37 differs marginally from zero ($p = .07$) whereas a *g* of 0.53 differs significantly from zero ($p = .004$). But with only seven studies in this condition, whether an effect size differs significantly from zero is likely to be unstable, limiting the inferences that can be drawn. As we noted (Levine et al., 2012, p. 598), when individual effect sizes are based on a small number of studies, experts recommend caution in drawing inferences from the *p*-values associated with those effects (Cooper, 2010).

Importantly, the statistic of primary interest in our meta-analysis is the one comparing the size of effects based on question type and timing. This $Q_{Between}$ statistic changes from $Q_B(3) = 40.70$ ($p < .001$) with all effects included to $Q_B(3) = 27.18$ ($p < .001$) when effects are excluded from studies with moderators expected to lessen the impact bias. The effect associated with the general-delayed question remains significantly larger than the effect shown in the 34 studies with other question types. Thus, as we stated in our article, the pattern of results we hypothesized was evident regardless of whether studies with such moderators were included or excluded.

Selective Recoding of the Direction of Effects

Wilson and Gilbert (2013) used a variety of strategies to recode the studies in the specific-immediate condition, stating that this condition was critical for testing our claims. In fact, we claimed that specific *or* immediate questions should differ from general-delayed questions. They arrived at very different effect-size estimates for four of the seven studies, resulting in a larger effect-size estimate for that condition ($g = 0.50$, or $g = 0.35$ if they eliminated the controversial studies) than we reported ($g = -0.01$). We would welcome a reanalysis of the affective forecasting literature using different coding criteria from those we used, but a convincing reanalysis would require that researchers develop transparent coding criteria and consistently apply those criteria to all of the studies (Cooper, 2010).

A detailed critique of Wilson and Gilbert's (2013) alternative coding of individual studies and of their use of a within-subject effect size is provided in the online supplementary material. Briefly, however, we disagree with the recoding they performed for two articles that resulted in a major change in the effect size in the specific-immediate condition. To code the valence of the focal event as negative in Sevdalis, Harvey, and Bell's (2009) studies, Wilson and Gilbert assumed that all participants who bought an object considered the price they paid for it unfavorable, ignoring the fact that some participants were pleased and others were disappointed with the price they paid. To recode a study concerning forecasts about spring break vacation (Wirtz, Kruger, Scollon, & Diener, 2003), Wilson and Gilbert treated vacation as a positive outcome for predictions of positive emotion and as a negative outcome for predictions of negative emotion. This was the only study in which they defined the focal event as both a positive and a negative outcome. Meta-analyses are only informative if studies are coded systematically. It is not legitimate to select one study in one condition and recode it without developing a way to code the many other focal events in the meta-analysis that would be expected to elicit a mixture of positive and negative emotion. Indeed, had they applied this coding criterion to other events, it would have resulted in a reversal of the impact bias in many of their own landmark studies.

However, even if we use the other coding choices that Wilson and Gilbert (2013) made in the specific-immediate condition, the effect size associated with general-delayed questions remains larger than the effect size associated with specific-immediate questions, $Q_B(1) = 4.51$, $p = .03$. In addition, the effect size associated with specific-immediate questions does not differ from the effect size associated with the 27 other studies that used specific or immediate questions, $Q_B(1) = 0.40$, $p = .53$.

Limitations of Our Meta-Analysis

The inferences that can be drawn from a meta-analysis are limited by how well the underlying constructs of interest are measured and how well the moderators are described in the individual studies (Cooper, 2010; Johnson & Eagly, 2000). Our analysis was designed to contrast people's ability to forecast the intensity of emotion they would feel when they were later thinking about a focal event (specific and immediate questions) versus when they were not necessarily thinking about it (general-delayed questions). Very few studies in the forecasting literature have assessed experienced emotional intensity using specific questions, however. Among those that did, the specificity of the questions varied. For example, some questions we coded as "specific" clearly assessed how people felt when thinking about a discrete event such as a course grade. Others required participants to summarize their feelings about extended events that likely spanned several emotional and nonemotional episodes, such as a holiday or vacation. In addition, we coded question type and timing from authors' descriptions of their procedures. Because the studies were not designed to test the effect of these variables on the intensity bias, descriptions of question type and timing were sometimes absent or ambiguous. We addressed most limitations in study descriptions through coding decisions and supplemental analyses (e.g., coding question timing as unspecified; Levine et al., 2012, p. 595, end of Footnote 1). But some disagreements between our coding and that done by Wilson and Gilbert (2013) are possible because descriptions of key features of the methods or results in the original reports were lacking or ambiguous.

Despite these limitations, the hypothesized difference in the magnitude of bias in the general-delayed condition versus the specific or immediate conditions is robust and persists with the exclusion and selective recoding of studies recommended by Wilson and Gilbert (2013). The impact bias is significantly greater when study participants are asked a general question about their emotional experience after a delay than when they are asked a specific or immediate question about their emotional experience. Wilson and Gilbert's recommendations do result in a larger estimate of the intensity bias in the specific and immediate conditions, however. Our findings, using transparent and consistent inclusion,

exclusion, and coding criteria, suggest that the intensity bias in these conditions is small.

This disagreement highlights the need for research that is explicitly designed to address people's ability to forecast the intensity of their feelings when thinking about events. The empirical studies reported in Levine et al. (2012) were critical for establishing the effect of specific versus general questions on the intensity bias because they were designed to address that very question while holding other factors (e.g., the focal event, question timing) constant. These studies also assessed whether the procedure commonly used to assess forecasting accuracy inflates the magnitude of the intensity bias. Although Wilson and Gilbert (2013) claimed that the meta-analysis represents the broadest challenge to the affective forecasting literature, it is actually the empirical studies that elucidate the causal effect of question type on the impact bias.

Validity of the Conclusions Drawn From the Empirical Studies

Levine et al.'s (2012) empirical studies showed that people can predict the intensity of their feelings about events with a high degree of accuracy (Studies 1 and 3) and that a procedural artifact contributes to people's tendency to overestimate the intensity of their feelings in general (Studies 2, 3, and 4). Wilson and Gilbert (2013) proposed individual differences and focalism as alternative explanations for the findings. We conducted additional analyses to test these alternative explanations, and as described below, the results do not support them.[1]

People Can Accurately Predict the Intensity of Their Feelings About Events

In Study 1, participants predicted how happy they would feel in the days following Obama's victory in the 2008 U.S. presidential election. They overestimated the extremity of their emotional response when later asked to report how they were feeling in general without reference to the election. In contrast, when participants were asked directly how they felt about Obama being elected, no systematic tendency to overestimate emotional intensity was found, and the correlation between predicted and experienced feelings was very high: $r(198) = .89$, $p < .001$. Wilson and Gilbert (2013) found this accuracy puzzling because participants in a study they conducted overestimated in predicting how they would feel after an election (Wilson et al., 2003, Study 1). They speculated that we might have found greater accuracy because we did not control for individual differences in baseline happiness or because we did not limit participants to those for whom the election outcome was important (ratings of 6 or greater on a 9-point scale). But these procedural differences do not account for our findings. With the modifications suggested by Wilson and Gilbert, there was no systematic tendency to overestimate emotional intensity, and the correlation between predicted and experienced emotion remained high: $r(162) = .93$, $p < .001$. Why then did their findings and ours differ? Though Wilson et al. (2003, Study 1) reminded participants about the election outcome earlier in their survey, they later asked participants to report how they were feeling in general rather than how they were feeling about the election. This is precisely the procedure that we found was likely to show a robust impact bias.

Focalism Does Not Account for Misinterpretation of General Forecasting Questions

Levine et al. (2012) argued that, when the typical forecasting procedure is used, an artifact inflates the extent to which people overestimate the emotional impact of events. Studies 2 and 4 showed that most participants (57% to 81% across interpretation questions and studies) misinterpreted requests to predict their emotional state in general as asking how they would feel when thinking about the focal event. Moreover, participants who misinterpreted the general forecasting question predicted more intense emotion than those who correctly understood that they had been asked to predict their general emotional state. Study 3 showed that clarifying the meaning of the forecasting question dramatically reduced overestimation. Wilson and Gilbert (2013) objected that the artifact explanation is similar to focalism. They proposed instead that participants who expect an event to have a greater impact on them are more likely to focus on it, leading them to interpret the forecasting question as specifically about that event.

We assessed whether focalism could account for the results of Study 2 by examining participants' ratings of how important the election outcome was to them (a rough measure of the impact participants expected the event to have on them). Rating the election as more important was not associated with a greater likelihood of misinterpreting the general forecasting question, $r_{pb}(98) = 0.10$, $p = .32$. Wilson and Gilbert (2013) further suggested that, "rather than a specific interpretation of an event producing a more extreme forecast, it may be that an extreme forecast produces a more specific interpretation of the question" (p. 745). In one condition of Study 4, students predicted not only how happy they would feel in general but also how happy they would feel specifically about getting a low exam grade. Students who made more extreme forecasts in response to the specific forecasting question were not more likely to misinterpret the general forecasting question, $r_{pb}(79) = 0.03$, $p = .79$. Indeed, across each of the four interpretation questions in Studies 2 and 4, greater expected impact (as assessed by importance) and more extreme forecasts did not account for participants' tendency to misinterpret the general forecasting question (all ps $> .31$).

Levine et al. (2012, p. 600) also described a study expressly designed to test whether focalism accounts for misinterpretation of forecasting questions. We assessed both how much participants expected to think about a tragic event (focalism) and how they interpreted the general forecasting question. The results showed that expecting to think more about the focal event and misinterpreting the general forecasting question were distinct processes that contributed independently to more extreme forecasts.

[1] The additional analyses presented make use of participants' ratings of the importance of the focal event (Studies 1 and 2) and how much they expected to think about the focal event (Study 3). Responses to these questions were not included in Levine et al. (2012) because they did not directly concern the intensity bias. In the experimental sessions, these questions followed those concerning intensity bias, so they could not have influenced participants' responses reported in Levine et al.

Clarifying the Meaning of the General Forecasting Question Reduces Bias

Wilson and Gilbert (2013) also proposed focalism as an explanation for the results of Study 3. In that study, Levine et al. (2012) experimentally manipulated how participants interpreted the general forecasting question. Undergraduates predicted and reported how they would feel after receiving an exam grade. We compared how much they overestimated emotional intensity when asked a general forecasting question in the typical context (right after they were asked to imagine getting their grade) versus in a context designed to clarify the question's meaning. To clarify the meaning of the general forecasting question, we first asked, "How happy will you feel about your grade?" (a specific forecasting question) and "Do you think your grade will affect your overall mood?" We then asked, "In general, how happy will you feel?" People assume that questioners will not be redundant (Grice, 1975), so asking the specific question first encouraged students *not* to interpret the subsequent general question as referring only to their feelings about their grade. The second question was added so that students would not exclude their grade from consideration when forecasting their feelings in general (e.g., Schwarz, Strack, & Mai, 1991). It explicitly invited students to consider whether their grade would influence their overall mood, and most students thought it would (74%). The results showed that clarifying the meaning of the general forecasting question markedly improved forecasting accuracy.

Wilson and Gilbert (2013) objected that the additional questions in the clarifying context condition may have served to defocalize people. They compared our manipulation to the lengthy diaries that Wilson et al. (2000) had participants fill out to induce them to think about a wide range of events and activities other than the focal event. In contrast, the questions we added prior to the general forecasting question both referred directly to the focal event ("your grade") and did not refer to any other event. So, it is unlikely that these questions improved forecasting accuracy by reducing focalism. Furthermore, contrary to Wilson and Gilbert's suggestion that the clarifying context condition reduced focalism, students in this condition did not expect to think less about their grade ($M = 5.83$, $SD = 2.06$) than students in the typical forecasting procedure condition ($M = 5.65$, $SD = 2.22$), $t(179) = 0.56$, $p = .58$, $g = .08$.

In summary, Wilson and Gilbert's (2013) alternative explanations for the results of our empirical studies fail to account for the findings. People's accuracy in predicting the intensity of their feelings specifically about events persists after controlling for individual differences in baseline emotion and importance. Focalism does not account for people's tendency to misinterpret general forecasting questions. A reduction in focalism does not account for people's greater accuracy when the meaning of the question is clarified. These findings show that misinterpretation of the general forecasting question is both common and preventable. As Wilson and Gilbert rightly pointed out, dozens of studies have shown that people's forecasts are prone to an impact bias. Nevertheless, it is hard to evaluate their conclusion that the impact bias is alive and well if the procedure used in most of these studies promotes overestimation. In our empirical studies, no evidence of impact bias was found when people predicted the intensity of their feelings specifically about events. A robust impact bias was found when people predicted their emotional state in general, and clarifying the forecasting question reduced but did not eliminate this bias. With a nod to Schrödinger's (1935/1980) cat, then, we conclude that the impact bias is both dead and alive.

Importance of Accuracy and Artifact in Forecasting Research

Do our findings alter Wilson and Gilbert's (2013) basic claims about the impact bias? After all, if people expect an event to be on their mind more often than it is and expect it to influence their emotional state for longer than it does, then they are overestimating the enduring emotional impact the event will have on their lives. Does it matter that people can accurately predict how intensely they will feel during the possibly rare moments when the event comes to mind? We believe it does matter. Evidence of accuracy is rarely as captivating as evidence of bias (Funder, 1987), but people's ability to predict the peak intensity of emotion that events will elicit has important implications for the quality of their decisions.

People's predictions about the peak intensity, rather than duration, of their feelings often guide their choices (Fredrickson & Kahneman, 1993). Peak intensity provides an index of how good or how bad an experience will be, whether one has the resources to cope with it, and the amount of effort it is worth expending to achieve or avoid it (Fredrickson, 2000). Indeed, for many decisions, the intensity of emotion people expect to feel while an experience is occurring and when they are thinking about it later is the primary determinant of choice (e.g., musical performances, vacations, dentist visits, public speaking; Buehler & McFarland, 2001). Even when making life-altering decisions concerning career, marriage, and children, people consider the peak positive and negative feelings different choices will bring about, as well as longer term effects on their emotional well-being. Moreover, as a guide to decision making, peak intensity may often be superior to overall emotional well-being. For example, after September 11, 2001, anticipated horror about future attacks motivated many individual and national decisions. The quality of those decisions varied, but the fact that people's overall emotional state did not differ significantly a few months before versus after the attacks (Whalen, Henker, King, Jamner, & Levine, 2004) does not imply that precautions were unjustified. Similarly, years after the death of a child, parents may experience extreme distress only on the infrequent occasions that they think about their loss. Pointing out that their feelings in general differ little, if at all, from their preloss feelings will not encourage them to let their other children play in traffic. Accurately predicting emotional intensity bodes well for decision making even if people overestimate the duration of their feelings.

Our exchange with Wilson and Gilbert (2013) highlights important directions for future research. One is to investigate why people might be better at predicting some features of their emotional experience than others. Research on memory for emotion provides some clues. To predict how they will feel in the future, people draw on episodic memories of similar experiences in the past (Schacter, Addis, & Bruckner, 2008). The peak intensity of past emotional episodes is remembered more accurately than their duration (e.g., Fredrickson & Kahneman, 1993), providing a better basis for prediction. In addition, as Levine et al. (2012) argued, emotional arousal narrows the focus of attention to central, salient

features of events (Levine & Edelstein, 2009). Although focalism (i.e., focusing on central, salient features of events at the time of prediction) leads to bias when people predict the duration of their feelings (Wilson et al., 2000), focalism may promote accuracy when people predict emotional intensity.

This exchange also has important implications for the procedures used to assess forecasting accuracy. Wilson and Gilbert (2013) maintained that people overestimate in predicting both the peak intensity and the frequency or duration of their feelings. Levine et al. (2012) found that people can predict how intensely they will feel about events fairly accurately. Our findings suggest that mispredicting frequency and duration, rather than the intensity of emotion, may be largely responsible for people's tendency to overestimate the emotional impact of events. Testing this, though, will require more precise research methods that assess people's ability to anticipate distinct features of emotion. Moreover, regardless of one's views about the scope of the impact bias, ensuring that participants understand what they are being asked to predict and report can only be beneficial. At the very least, this would rule out question misinterpretation as a contributor to the impact bias. Our hope, however, is that more precise methods will allow researchers to better determine when and why biases emerge in people's forecasts of emotion.

Conclusions

We have argued that the impact bias and the procedure most commonly used to assess it mask important distinctions in people's affective forecasting abilities. As Wilson and Gilbert have long acknowledged, the impact bias encompasses several forecasting errors including overestimating the intensity as well as the frequency and duration of emotion. The typical forecasting procedure does not distinguish between these features of emotion and, for the majority of people who misinterpret the request to predict how they will feel in general, compares predicted feelings about an event to later overall mood. Like an out-of-focus photograph, then, this method does not provide researchers with a precise picture of people's forecasting strengths and weaknesses. When procedures are used that target specific forecasting abilities and are less likely to be misinterpreted, it appears that people are prone to making some errors but not others. Intensity bias can be reduced or eliminated by improving the methodology in affective forecasting studies, whereas frequency and duration biases may persist. Wilson and Gilbert's (2013) contribution to identifying biases in affective forecasting, as well as sources of these biases, cannot be overstated. But progress in this area will be facilitated by the use of measurement tools that are more precise. In Schrödinger's (1935/1980) famous thought experiment, a cat in a box, whose life depends on a concealed random event, is simultaneously dead and alive until scientists open the box and take a look. Is the impact bias alive and well? To check the status of the cat, researchers have to open the box.

References

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851. doi:10.1037/0003-066X.63.9.839

Buehler, R., & McFarland, C. (2001). Intensity bias in affective forecasting: The role of temporal focus. *Personality and Social Psychology Bulletin, 27,* 1480–1493. doi:10.1177/01461672012711009

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Fredrickson, B. L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion, 14,* 577–606. doi:10.1080/026999300402808

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65,* 45–55. doi:10.1037/0022-3514.65.1.45

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75–90. doi:10.1037/0033-2909.101.1.75

Gilbert, D. T., Driver-Linn, E., & Wilson, T. D. (2002). The trouble with Vronsky: Impact bias in the forecasting of future affective states. In L. F. Barrett & P. Salovey (Eds.), *The wisdom in feeling: Psychological processes in emotional intelligence* (pp. 114–143). New York, NY: Guilford Press.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 75,* 617–638. doi:10.1037/0022-3514.75.3.617

Green, J. D., Davis, J. L., Luchies, L. B., Coy, A. E., Van Tongeren, D. R., Reid, C. A., & Finkel, E. J. (2013). Victims versus perpetrators: Affective and empathic forecasting regarding transgressions in romantic relationships. *Journal of Experimental Social Psychology, 49,* 329–333. doi:10.1016/j.jesp.2012.12.004

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). New York, NY: Academic Press.

Johnson, B. T., & Eagly, A. H. (2000). *Quantitative synthesis of social psychological research* (CHIP Documents, Paper 13). Retrieved from http://digitalcommons.uconn.edu/chips_docs/13

Kalisch, R., Wiech, K., Herrmann, K., & Dolan, R. J. (2006). Neural correlates of self-distraction from anxiety and a process model of cognitive emotion regulation. *Journal of Cognitive Neuroscience, 18,* 1266–1276. doi:10.1162/jocn.2006.18.8.1266

Lench, H. C., Safer, M. A., & Levine, L. J. (2011). Focalism and the underestimation of future emotion: When it's worse than imagined. *Emotion, 11,* 278–285. doi:10.1037/a0022792

Lepper, M. R., Henderlong, J., & Gingras, I. (1999). Understanding the effects of extrinsic rewards on intrinsic motivation—Uses and abuses of meta-analysis: Comment on Deci, Koestner, and Ryan (1999). *Psychological Bulletin, 125,* 669–676. doi:10.1037/0033-2909.125.6.669

Levine, L. J., & Edelstein, R. S. (2009). Emotion and memory narrowing: A review and goal relevance approach. *Cognition & Emotion, 23,* 833–875. doi:10.1080/02699930902738863

Levine, L. J., Lench, H. C., Kaplan, R. L., & Safer, M. (2012). Accuracy and artifact: Reexamining the intensity bias in affective forecasting. *Journal of Personality and Social Psychology, 103,* 584–605. doi:10.1037/a0029544

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Thousand Oaks, CA: Sage.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2008). Episodic simulation of future events: Concepts, data, and application. *Annals of the New York Academy of Sciences, 1124,* 39–60. doi:10.1196/annals.1440.001

Schrödinger, E. (1980). The present situation in quantum mechanics (J. D. Trimmer, Trans.). *Proceedings of the American Philosophical Society, 124,* 323–338. (Original work published 1935)

Schwarz, N., Strack, F., & Mai, H. P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly, 55,* 3–23. doi:10.1086/269239

Sevdalis, N., Harvey, N., & Bell, A. (2009). Affective equilibria in the endowment effect. *Journal of Behavioral Finance, 10,* 89–100. doi:10.1080/15427560902720545

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., . . . Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7,* Article 10. doi:10.1186/1471-2288-7-10

Van Boven, L., & Loewenstein, G. (2003). Projection of transient drive states. *Personality and Social Psychology Bulletin, 29,* 1159–1168. doi:10.1177/0146167203254597

Whalen, C. K., Henker, B., King, P. S., Jamner, L. D., & Levine, L. (2004). Adolescents react to the events of September 11, 2001: Focused versus ambient impact. *Journal of Abnormal Child Psychology, 32,* 1–11. doi:10.1023/B:JACP.0000007576.94501.a0

Wilson, T. D., & Gilbert, D. T. (2008). Explaining away: A model of affective adaptation. *Perspectives on Psychological Science, 3,* 370–386. doi:10.1111/j.1745-6924.2008.00085.x

Wilson, T. D., & Gilbert, D. T. (2013). The impact bias is alive and well. *Journal of Personality and Social Psychology, 105,* 740–748. doi:10.1037/a0032662

Wilson, T. D., Meyers, J., & Gilbert, D. T. (2003). "How happy was I, anyway?" A retrospective impact bias. *Social Cognition, 21,* 421–446. doi:10.1521/soco.21.6.421.28688

Wilson, T. D., Wheatley, T. P., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 78,* 821–836. doi:10.1037/0022-3514.78.5.821

Wirtz, D., Kruger, J., Scollon, C. N., & Diener, E. (2003). What to do on spring break? The role of predicted, on-line, and remembered experience in future choice. *Psychological Science, 14,* 520–524. doi:10.1111/1467-9280.03455

Received April 18, 2013
Revision received July 29, 2013
Accepted August 5, 2013 ■