Overconfidence in Overconfidence

Kevin Dorst
kevindorst@pitt.edu
Draft—comments welcome!

January 2020

Abstract

Do people tend to be overconfident in their opinions? Many psychologists think so. They have run *calibration studies* in which they ask people a variety of questions, and then test whether their confidence in their answers matches the proportions of those answers that are true. Under certain conditions, an "overconfidence effect" is robust—for example, of the answers people are 80% confident in, only 60% are true. Psychologists have inferred that people tend to be irrationally overconfident. My question is when and why this inference is warranted. Although it may at first appear hopelessly flawed, I show that under controlled conditions it *is* a warranted inference. However, I argue that real-world studies standardly fail to meet these conditions—and, as a result, that *rational* people can often be expected to display the "overconfidence effect." Thus in order to test whether people are overconfident, we must first predict whether and to what extent a rational person will display this effect, and then compare that prediction to real people's performance. I show how in principle this can be done—and that doing so may overturn the standard interpretation of robust empirical effects. Upshot: we have much less evidence for overconfidence than many have thought.

1 The Question

Pencils ready! For each question, select your answer, and then rate how confident you are in that answer on a $50-100\%$ scale:

1) What city has a bigger population: Rome or Madrid?
2) What came first: the printing press, or Machiavelli's *The Prince*?
3) What is longer: an American football field, or a home-run lap?

If you're like most people, this test will reveal two results. First, it's likely that only one or two of your answers is correct. Second—and perhaps more worryingly—it's likely that your confidence in your answers does not match this probability of being correct. Among several dozen test-takers, the average confidence people had in their answers was 64%, while the proportion of correct answers was only 44%.[1]

That rather striking result—the so-called "overconfidence effect"—is common: on a variety of types of tests, people's average confidence in their answers exceeds the proportion that are correct.[2] Many have concluded that people tend to be overconfident in their opinions—i.e. more confident than it is rational for them to be, given their evidence. And they have invoked this conclusion to explain a variety of societal ills—from market crashes, to political polarization, to wars.[3] In the words of one widely-read textbook: "No problem in judgment and decision-making is more prevalent and more potentially catastrophic than overconfidence" (Plous 1993, 213).

So they say. But how—precisely—did we reach the conclusion that people tend to be overconfident?

The evidence comes in the form of calibration studies like the one you just took. We ask people a variety of questions, have them report their confidence in their answers, and then graph that confidence against the proportion of answers that are true. Say that a person is *calibrated* if exactly $x\%$ of the claims that they are $x\%$ confident in are true. They are *over*-calibrated if fewer than $x\%$ of such claims are true.[4] And they are *under*-calibrated if more than $x\%$ of such claims are true. Focusing on binary-question formats ("A or B?")—wherein people are always at least 50% confident in their answers—schematic graphs of these different **calibration curves** are given in Figure 1. Meanwhile, Figure 2 (p. 4) plots the results from my survey and those of a well-known study—both finding that people were substantially over-calibrated.

That's the evidence—namely, that people are often over-calibrated. How do we get from there to the conclusion—namely, that people are often overconfident? Well, if people are quite confident in their opinions, and yet so many of those opinions are false, it is natural to infer that they are *too* confident—*over*confident. Natural as it may be, that is an inference: it moves from "you are (mis)calibrated" to "you are (ir)rational." Call it the **calibration inference**. Psychologists conducting these studies have been presupposing that it is warranted.

The Question: Is it? More specifically: under what circumstances is the calibration inference warranted—and do these include the circumstances of real-world calibration

[1] Answers: Rome; *The Prince*; a football field. Confused? See §2, p. 6 for an explanation.

[2] For summaries, see Lichtenstein et al. (1982), Hoffrage (2004), and Glaser and Weber (2010).

[3] E.g. Howard (1984); Odean (1999); Glaser and Weber (2007); Johnson (2009); Myers (2010, 377); Johnson and Fowler (2011); Ortoleva and Snowberg (2015); van Prooijen and Krouwel (2019).

[4] "Over"-calibrated because their confidence in those opinions needs to be lower to be calibrated. In the graphs below, imagine the person controlling a left-right slider for their confidence; over-calibration is putting it too far to the right; under-calibration is putting it too far to the left.

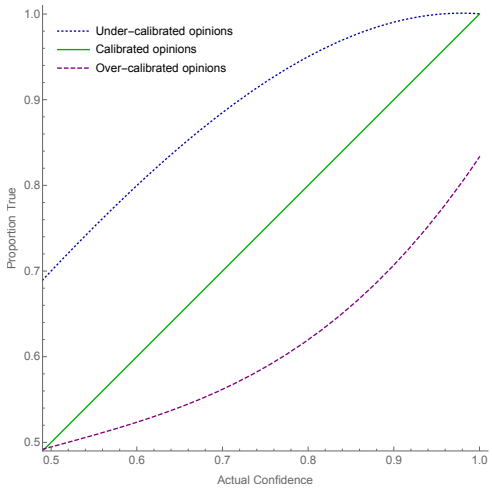

Figure 1: Schematic calibration curves

studies? I'm going to show that often the answer is 'no'.

The Plan: Although it may at first appear that the calibration inference is hopelessly flawed (§2), there *is* an important insight behind it—in certain simple scenarios, it works (§3). But in realistic variations of these simple scenarios, it can fail systematically—often rational people can be expected to be miscalibrated, meaning *mis*calibration is evidence for *rationality* (§4). Upshot: we cannot test whether people are overconfident by comparing their calibration curves to perfectly calibrated ones, as is standard—rather, we must compare them to the predicted rational *deviations* from calibration. I show how this can in principle be done, and that doing so has the potential to reverse standard interpretations of robust experimental effects (§§5-6). (For aficionados: I'll show that versions of the "hard-easy effect" (§5.2), the "base-rate effect" (§5.2) the "Dunning-Kruger effect" (§6.2) and the finding that confidence intervals are "too tight" (§6.3) are all to be expected from rational Bayesians.)

The Upshots: We'll see that if the theory developed here is correct, it has three main consequences. First, in order to explain why the calibration inference makes sense, philosophers must accept a theory that supports strong epistemic deference principles. Second, it is a live option for psychologists conducting calibration research to eschew its "bias-based" methodology in favor of a rationalist one that aims to identify subject-matters on which people tend to possess misleading evidence. Finally, rigorously applying philosophical tools for modeling rationality reveals—I think—that we have been too confident that people tend to be overconfident. This suggests that philosophers and psychologists can and should work together more closely in tackling the question of how irrational we truly are.

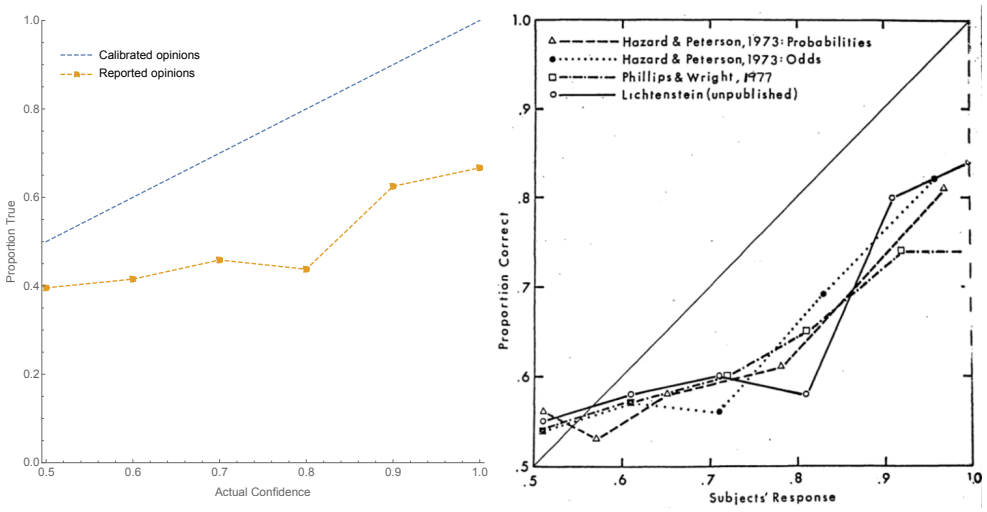

Figure 2: The "Overconfidence Effect". **Left:** My survey. **Right:** Lichtenstein et al. (1982).

2 The Problem

There is a problem here. The calibration inference involves three quantities:

(1) A person's actual degrees of confidence in some claims.

(2) The proportion of those claims that are true.

(3) The degrees of confidence it would be rational for them to have in those claims.

The only quantities that are directly observed are (1) and (2). The calibration inference uses these to infer something about (3): from the observation that (1) is higher than (2), it is inferred that (1) is higher than (3). Clearly this makes sense only if rational confidence—(3)—can be expected to align with proportion true—(2).

The point can be made graphically. What would it mean to say that people tend to be overconfident in a given domain?[5] It would mean that they are (on average) *more extreme* in their opinions in that domain than they would be if they were rational. If we plot actual degrees of confidence against rational degrees of confidence (on $50-100\%$ scale), people tend to be rational if (averaging across opinions) rational confidence matches actual confidence—the curve is diagonal; they tend to be overconfident if rational confidence is less extreme than actual confidence—the curve is tilted. (See the left side of Figure 3.) That's the overconfidence hypothesis. What is the evidence offered in its favor? It is that in a variety of settings, people are over-calibrated: if we

[5]The "in a given domain" rider is important, as patterns of miscalibration vary widely across different sets of questions, and psychologists point out that this is calls for a more nuanced normative assessment (see especially Koehler et al. 2002; Brenner et al. 2005). We'll introduce refinements to the empirical story in §4.1 and onwards—for now, I'll focus on tests for which the "overconfidence effect" is observed and the corresponding overconfidence hypothesis is taken to be supported.

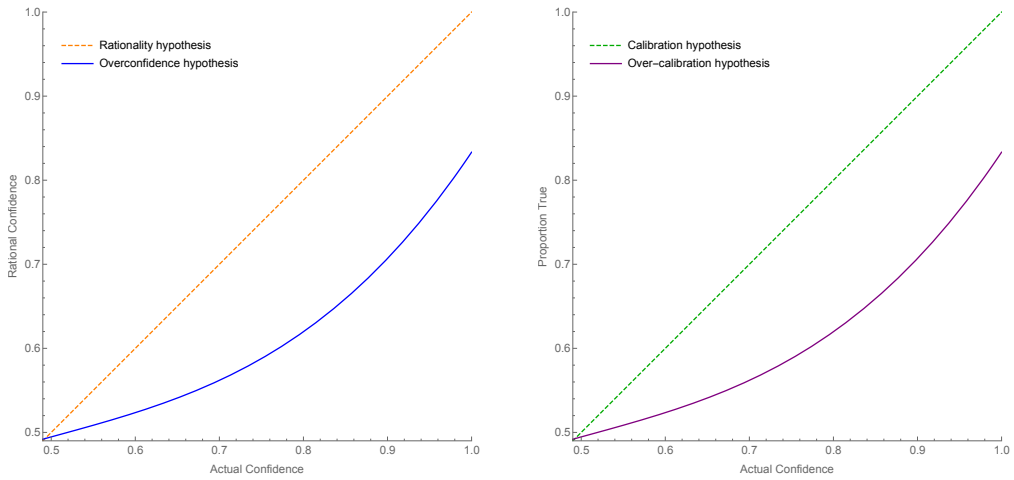

Figure 3: **Left:** Rationality vs. Overconfidence hypotheses. **Right:** Calibration vs. Over-calibration hypothesis

plot actual degree of confidence against *proportion true*, the curve is tilted—see the right side of Figure 3.

Simple point: although the two graphs look the same, the axes are different. It follows that the calibration inference is warranted when and only when you should expect the two axes to align—i.e. when you should expect a rational person's judgment to be calibrated on the given test. I know of no calibration study that states the assumptions needed to derive the result that we should expect rational opinions to be calibrated in their study.[6] In fact, I know of no study that explicitly represents the rational degrees of confidence for subjects to have as a variable to be investigated. Instead, all assume that we can investigate whether subjects' degrees of confidence match the rational degrees of confidence by simply investigating whether their confidence matches proportion true. But rational confidence is distinct from proportion true. That is the problem.

How serious is the problem? That is the question of this paper. To begin, we should note that—at least sometimes—rational confidence and proportion true can be expected to come arbitrarily far apart, meaning the calibration inference can fail completely.

Case 1: I have a coin in my pocket that's 60% biased toward heads. I'm about to toss it 100 times—how confident are you, of each toss, that it'll land heads? Write the number down—I'll look at it in a minute. First to toss the coin (...done). Turns out that

[6]Many studies simply assume that the rational opinions will be calibrated (e.g. Lichtenstein et al. 1982; Dunning et al. 1990; Vallone et al. 1990; Griffin and Tversky 1992; Kahneman and Tversky 1996; Hoffrage 2004; Glaser and Weber 2010; Ehrlinger et al. 2016; Magnus and Peresetsky 2018). Some explicitly derive this result for a given Bayesian agent (Brenner et al. 2005; Moore and Healy 2008; Merkle and Weber 2011)—but to do so they all implicitly assume that the Bayesian's prior beliefs match the objective frequencies on the test. As we'll see, this cannot in general be assumed.

(surprisingly) it landed heads only 30 times. Now to look at the confidence you wrote down... whoa—60%? Only 30% of the claims were true, but you were 60% confident in each of them. Have I gained evidence that you are overconfident? Obviously not—your 60% confidence was perfectly rational, given your evidence; you simply were unlucky in how the coin landed. Upshot: sometimes the calibration inference is not warranted.

Of course, in Case 1 the outcome of the coin tosses was a fluke. Does the calibration inference work in *normal* cases?

Not necessarily. *Case 2:* I have an urn of mis-printed coins: 60 of them are double-headed; the remaining 40 are double-tailed. I'm about to pull a single coin from the urn and toss it 100 times. How confident are you, of each toss, that the coin will land heads on that toss? 60%, I take it. Here goes... ...actually, on second thought, I don't need to bother. If I draw a double-headed coin, it'll land heads every time—100% of the things you're 60% confident in will be true. And if I draw a double-tailed coin, it'll land tails every time—0% of the things you're 60% confident in will be true. Either way, you'll be badly miscalibrated. So if the calibration inference were warranted, either way I'd be able to infer that you're irrational! Obviously I can do no such thing. I know (and you do too) that in *all* cases—normal or otherwise—you will be miscalibrated on this test, even though you're perfectly rational.

Very well. But in Case 2 *something* funny is going on, since each of the claims I'm testing you on ("Will the first toss land heads? The second?") will stand or fall together. When that's *not* the case—and there are no fluke-y coin tosses—does the calibration inference work?

Again, not necessarily. *Case 3:* Take some claims that have nothing to do with each other (or with coin tosses). Here are three—how confident are you in each?

i) Rome has a bigger population than Madrid.

ii) *The Prince* came before the printing press.

iii) A football field is longer than a home-run lap.

Those are easy, right? 100% in each! After all, I just told you the answers.

Confession: I lied. Each of (i)–(iii) is false; I inverted the answers on you. So 0% of the claims you were 100% confident in were true. Have I gotten evidence that you are overconfident? Obviously not. Sometimes the rational answers are (systematically) wrong. You were perfectly rational to believe me ("Who lies about trivia in an academic paper??"). I *tricked* you. And it's no sign of irrationality that you can be tricked.

In fact, that's a theorem. *Any* perfectly rational person can in principle be tricked, and thereby given a test on which they are expected to be over-calibrated. Here's a recipe. Let $A = \{p_1, ..., p_n\}$ be any set of claims you like. Let $R(p_i)$ be the degree of confidence you would have in p_i if you were perfectly rational. Perhaps R is calibrated on this test. No matter. That means (for instance) that of the claims in A that R assigns 80% confidence to, 80% are true—the remaining 20% of are false. So before

I give you the test, I'll (unbeknownst to you) remove all the true claims that are in A—leaving the false ones—to make a new test. Now I give this new test to you, asking for your degree of confidence in each claim. You—perfectly rationally—give the answer $R(p_i)$ for each one. When I learn that you are massively over-calibrated (for example, 0% of the claims that you were 80% confident in are true), do I get evidence that you are irrational? Obviously not—I've simply given you a trick test.

Put simply: the mere fact that someone is wrong more often than they expect to be does not necessarily give us any evidence that they are irrational—whether it does so depends completely on how the test was constructed. Often you should not expect rational people to be calibrated.[7]

To be clear, I am not suggesting that these toy cases undermine the calibration inference in practice. What they do show is that the calibration inference is not *always* warranted—and, therefore, that it's important to ask the question of why the calibration inference is ever warranted, so that we can assess whether it is warranted in real cases. As we'll see, in certain circumstances we *should* expect rational people to be calibrated (§3). However, the explanation for this also reveals that there are systematic reasons why we should *not* expect rational people to be calibrated in most real-life studies (§§4–5), calling into question the evidence offered in favor of overconfidence.

2.1 The bigger problem

But before diving into the details, let me step back and say something about the bigger picture here. Calibration research—like much of the research on judgment and decision-making—is at the interface between the normative and descriptive. The questions are: "How *do* people think?"; "How *should* they think?"; and—often the punchline—"To what degree does the former fall short of the latter?" Psychologists are experts on the first question. Philosophers are experts on the second. And to get to the punchline, we need to answer the second question properly. Yet—at least on questions of calibration and overconfidence—there has been little engagement between the fields.[8]

[7] For those familiar with certain bits of theory, a clarification may be helpful here. Any Bayesian will expect their own opinions to be calibrated (see below). But we are not them, and we know things that they do not. Therefore there is no theorem that *we* should expect them to be calibrated. Often we should not. (Why must *they* expect to be calibrated? Because a Bayesian's estimate of the proportion of truths amongst some particular set of claims will equal their average degree of confidence in them. Letting C be any probability function, $\mathbb{E}[X]$ be its expectation of a variable X ($\mathbb{E}[X] := \sum_t C(X = t) \cdot t$), and $I(q_i)$ be the indicator of q_i (1 if q_i is true, 0 if not), we have: $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} I(q_i)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[I(q_i)] = \frac{1}{n}\sum_{i=1}^{n} C(q_i)$. Thus our subject's estimate of the proportion of truths amongst the claims they are 80% confident in must be 80%. Moreover, so long as they treat the claims (relatively) independently, they will (by the weak law of large numbers) be confident that roughly 80% of those claims are true.)

[8] To my knowledge no philosophers have directly addressed the calibration inference, instead focusing on different questions about the epistemic significance of calibration. Some ask whether calibration can objectively vindicate a set of opinions (van Fraassen 1983; Dawid 1983; Seidenfeld 1985; Joyce 1998;

As a result, psychologists have controlled the narrative—and it has, by and large, been a narrative of irrationality. It is claimed that calibration studies show people's degrees of confidence to be "irrational" (Hoffrage 2004, 245; Magnus and Peresetsky 2018, 2), "unjustified" (Dunning et al. 1990, 579; Vallone et al. 1990, 588), "unreasonable" (Merkle and Weber 2011, 264), "biased" (Glaser 2010, 249; Koehler et al. 2002, 686), the result of "cognitive illusions" (Kahneman and Tversky 1996, 589), and so on.[9]

Why are people irrational in these ways? Perhaps because they do not take adequate account of the unknowns in a situation (Dunning et al. 1990; Vallone et al. 1990); or because they fall prey to confirmation bias or motivated reasoning (Koriat et al. 1980; Taylor and Brown 1988; Kunda 1990); or because they have a fear of invalidity (Mayseless and Kruglanski 1987); or because they are insensitive to the difficulty of the task (von Winterfeldt and Edwards 1986); or because they over-rely on the strength of the evidence and disregard its weight (Griffin and Tversky 1992; Koehler et al. 2002; Brenner et al. 2005); or because they do insufficient cognitive or meta-cognitive processing (Sniezek et al. 1990; Bol and Hacker 2012; Ehrlinger et al. 2016); or because of some combination of these factors (Tetlock and Gardner 2016). Regardless of the details, these stories agree that calibration studies have shown that ordinary people making ordinary judgments show a systematic tendency to form their opinions in an irrational and biased way. The narrative is an instance of the broadly irrationalist picture of human reasoning that has come out of much of the psychological literature over the past several decades (Kahneman et al. 1982; Krueger 2012). This picture has in turn entered widely-read psychology textbooks (Plous 1993; Myers 2010) and the broader culture (Fine 2005; Ariely 2008; Kahneman 2011; Tetlock and Gardner 2016), has driven the rise of behavioral economics (Thaler and Sunstein 2009), and has been widely cited in explanations of societal ills like polarization and intolerance (Sunstein 2000, 2009; Fine 2005; Lazer et al. 2018).[10]

Compared to psychologists, philosophers have been largely absent from these dis-

Dunn 2015; Pettigrew 2016); others ask whether a Bayesian agent's beliefs about their own long-run calibration are problematic (Dawid 1982; Belot 2013a,b; Elga 2016); and others ask how our your expectations about the calibration of your answers should affect your confidence in each (Roush 2009, 2016, 2017; White 2009b; Christensen 2010a, 2016; Sliwa and Horowitz 2015; Schoenfield 2015, 2016; Isaacs 2019).

[9] If you have any doubts about the normative interpretation of these studies, Kahneman and Tversky put it bluntly: "Our disagreement [with Gigerenzer (1991)] is normative, not descriptive. We believe that subjective probability judgments should be calibrated, whereas Gigerenzer appears unwilling to apply normative criteria to such judgments" (1996, 589).

[10] Of course, there has been some pushback within psychology to the irrationalist explanations of calibration studies (Gigerenzer 1991; Gigerenzer et al. 1991; Erev et al. 1994; Juslin 1994; Juslin et al. 2000; Moore and Healy 2008), and there has been quite a bit of pushback to the irrationalist picture more broadly—most notably, in the rise of "rational analysis" approaches to cognitive science (e.g. Anderson 1990; Oaksford and Chater 1994, 1998, 2007; Chater and Oaksford 1999; Tenenbaum and Griffiths 2006; Hahn and Oaksford 2007; Harris and Hahn 2011; Griffiths et al. 2012; Cushman 2019).

cussions.[11] Part of what I'm arguing is that they shouldn't be. Questions about how a rational person would think in a given situation are incredibly subtle, and any claimed demonstration of *ir*rationality will rely (implicitly or explicitly) on answers to those questions. Those answers should be heavily scrutinized—and philosophers are experts at developing and deploying tools for doing just that. So philosophers, too, have things to contribute to the study of human (ir)rationality.

That is the claim, at least. Let's dive into the details—perhaps you will agree.

3 The Insight

When you learn about the results of a calibration study, you get a *lot* of evidence: how (mis)calibrated many subjects were across many levels of confidence; what sorts of test items were used, and how they were selected and presented; what the researchers were looking for and responding to; etc. All this evidence makes things complicated.

Let's start by making things simple. Suppose you get very limited evidence. A single subject—Calvin—was given a calibration test; the questions were selected to be random and unrelated. All you are told is what proportion of claims amongst those he was 80% confident in—call those his **80%-opinions**—were true.

I claim that in this simple scenario, the calibration inference is warranted. If you learn that (roughly) 80% of Calvin's 80%-opinions were true, you get strong evidence that those opinions were rational; if you learn that far fewer (or far more) than 80% of these opinions are true—say, 60% (or 95%)—you get strong evidence that he was overconfident (underconfident). This is the insight behind calibration studies. Why is it correct?

Begin with a parable. Long ago, Magic Mary possessed a variety of magic coins—some were biased to come up heads almost every time; others to come up heads 90% of the time; others 80%, and so on.[12] The magic coins had a variety of special markings on them—on some, George Washington has a large nose and small ears; on others, he has a thin neck and bushy eyebrows; etc. In principle, if one knew how to decipher the markings, one could tell what the bias of the coin was just by looking at it.

Mary tossed the coins many, many times. She kept fastidious records: for each toss she wrote the details of the coin's markings on one side of a stone tablet, and the outcome of the toss (heads or tails) on the other.

Alas, Magic Mary and her magic coins are long gone—but many of the tablets remain, stored in various historical archives. And alas, no one can decipher the markings to tell which bias a given tablet corresponds to.

[11] Some exceptions: Cohen (1981); Stich (1985); Kelly (2004, 2008); Crupi et al. (2008); Fitelson and Hawthorne (2010); Koralus and Mascarenhas (2013); Nebel (2015); Icard (2017); Hedden (2018); Mandelbaum (2018).

[12] *Magic* coins because—it turns out—you cannot bias a real coin (Gelman and Nolan 2002).

. . . or so we thought! But now bias-busting Bianca claims that she can decipher the markings and determine the coins' biases. How can we test her claim, given that *we* don't know how to decipher them?

Here's a good strategy. Go to an archive that contains a representative sample of tablets; draw a tablet at random; show her the markings-side, having her announce her guess as to whether it landed heads or tails along with her confidence in that guess; write down whether she got it right (but don't tell her); then draw a new tablet and repeat. Suppose we do this with many, many tablets, and then I tell you this: "Of the guesses she was 80% confident in, 79% were correct!" How confident are you now that Bianca can reliably recognize the 80%-biased coins—i.e. those that are 80% biased toward heads and those that are 80%-biased toward tails? Quite confident, I take it. For—in brief—it is rather surprising that so nearly 80% of those coins landed the way she guessed; and if she can reliably decipher them, that would explain why this is so. Conversely, if I instead told you that only *60%* of the judgments she was 80% confident in were correct, you should—for parallel reasons—suspect that she can*not* reliably decipher the markings of the 80%-biased coins, and instead that she is likely over-estimating the strength of these coins' biases.

Call this inference—from "Bianca was (mis)calibrated in her 80%-opinions" to "she probably can(not) reliably decipher the 80%-bias markings"—the **bias inference**. Clearly it is warranted in this simple scenario. And clearly there is an analogy between Bianca's bias-deciphering test and Calvin's calibration test. If we can get clear on what exactly the analogy is and why the bias inference works for Bianca, it'll show us what needs to be the case for the calibration inference to work for Calvin.

In fact, that's one of the main claims of this paper: if we want to know whether and to what extent we can expect the calibration inference to work in a given scenario, imagine a parallel scenario for Bianca and her coins to see whether and to what extent the bias inference will work in that scenario.

So: Why does the bias inference work in this scenario? I said that it is because the hypothesis that she can(not) decipher the coins would help explain her calibration if she is (mis)calibrated—but what does that mean more precisely, and why is it true?

What it means more precisely is this. Before I tell you about Bianca's calibration, you should think to yourself:

> "If she can reliably recognize the 80%-biased coins, then the coins she says '80%' on will (on average) be 80%-biased in the way she predicts—and conditional on *that*, I'm confident that roughly 80% of those tosses will land the way she predicts. Meanwhile, if she *can't* reliably recognize whether a coin is 80% biased, it's much more likely that a different proportion will land the way she predicts—for example, if she's over-estimating the bias, probably only 70% or 60% of the coins she says '80%' on will land the way

she predicts."

Thus the evidence you received—that 79% of her 80%-opinions were correct—is much more likely given that she can decipher the 80%-biased coins than it is given that she cannot; so it provides reason to think she can do so. Conversely, if you learn that only 60% of her 80%-opinions were correct, this is much more likely given that she's over-estimating the bias of the coins, so it provides reason to think that she is over-estimating.

The driving force of the bias inference, then, is that hypotheses about whether she is recognizing the coins' biases, over-estimating them, or under-estimating them, each have direct and strong implications for how many of the coins you should expect to land the way she guesses.

Crucial question: why is this so? Answer: because hypotheses about the (average) biases of groups of coins have two very specific effects on how confident you should be in the outcomes of their tosses. First, you should **defer** to the average biases of the coins in setting your opinion for how a given coin will land: conditional on the coins corresponding to Bianca's 80%-opinions having an average bias of $x\%$ toward her predictions, you should be $x\%$-confident that each of those predictions will be true. Second, this deference is **independent**: regardless of how her other predictions turn out, it is still the case that conditional on the coins having an average bias of $x\%$ toward her prediction, you should be $x\%$-confident that her next prediction will be true.[13] Combined, these principles drive the bias inference by making it so that conditional on the coins having an average of $x\%$ bias toward Bianca's predictions, you're confident that roughly $x\%$ are true.

Upshot: for the calibration inference to work in Calvin's case, analogous deference and independence principles must hold. Let's now see what the analogy amounts to.

Bianca takes a bias-deciphering test in which she announces her best guesses about how coins with various markings landed, along with her confidence in those guesses. We want to use her resulting calibration score to draw conclusions about whether she is reliably deciphering the coins' biases, or over-estimating them, or under-estimating them. Meanwhile, Calvin takes a calibration test on which he announces his best guesses about the true answers to binary questions of various kinds, along with his confidence in those guesses. We want to use his resulting calibration score to draw conclusions about whether he is rational, overconfident, or underconfident.

For each tablet Bianca is shown, there is a fact about what the corresponding coin's bias was. Likewise, for each question Calvin assesses, there is a fact about the rational degree of confidence he should have in each of the possible answers.

[13] In standard setups of our case, these two principles follow from the well-known Principal Principle and its refinements (Lewis 1980, 1994; Hall 1994; Briggs 2009b). See below for formal statements of their analogues in Calvin's case.

We wanted to know whether Bianca can tell what the markings mean for the biases of the various coins. Likewise, we want to know whether Calvin can tell what his evidence means for the rational degree of confidence he should have in the various answers.

In Bianca's case, the bias inference went though because we should defer to the *biases* of the coins, and do so independently of how her other predictions turn out. Likewise, then, in Calvin's case: the calibration inference will go through when and because we should defer to the *rational* degrees of confidence for Calvin to have in his answers, and do so independently of whether his other answers turn out to be true or false.

What does this mean more precisely? Consider all of the guesses Calvin assigns 80% confidence to—his 80%-opinions. Label them $q_1, ..., q_n$, so q_i is the claim that *the ith claim that Calvin was* 80% *confident in on this test (whatever it is) is true.*[14] We can entertain different hypotheses for what the average *rational* confidence is for Calvin to have in these claims. Let $\overline{\boldsymbol{R}}$ be this quantity, whatever it is.[15] Perhaps Calvin's 80%-opinions are on average rational, in which case this quantity will be 80%: $\overline{R} = 0.8$. Or perhaps they are on average overconfident (or underconfident), in which case it will be lower (or higher) than 80%: $\overline{R} < 0.8$ (or $\overline{R} > 0.8$).

Let q_i be any of Calvin's 80%-opinions. If you learn what the average rational opinion for Calvin to have in those opinions is, how does that affect your opinion in q_i? For the case to be analogous to Bianca's, you must defer. Let $\boldsymbol{P}$ be a probability function representing *your* rational degrees of confidence. Then what we need is:

> **Deference:** Upon learning that the average rational confidence for Calvin to have in his 80%-opinions is x%, you should be x% confident in each of them.
> For all q_i: $P(q_i|\overline{R} = x) = x$.

How plausible is Deference? That depends heavily on the epistemological theory we accept.[16] It deserves far more discussion, but let me say just two things in its defense.

[14]For simplicity I assume you know that there are n such opinions. To generalize to the case where you don't know how many there are, we simply need to assume that learning how many there are would not affect our Deference and Independence principles below, and would not affect your confidence in what the (average) rational opinion for Calvin to have is. The inference will then go through by performing the reasoning described below, averaging over the various values n might take.

[15]Formally, let $R(q_i)$ be the Rational confidence for Calvin to have in any given claim q_i. Then $\overline{R} := \frac{1}{n}\sum_{i=1}^{n} R(q_i)$. I'll assume that the opinions that are rational for any given person can be modeled with a precise probability function. The same sort of reasoning may go through if the rational degrees of confidence were not unique (Schoenfield 2014) or not precise (Schoenfield 2012); for discussion of the (de)merits of such models, see White (2005, 2009a); Schultheis (2018); Carr (2019).

[16] Deference is an interpersonal, rationalized, and "averaged" generalization of the well-known Reflection principle (van Fraassen 1984; Briggs 2009a; Christensen 2010b; Mahtani 2017). §A.1 shows how this "averaged" version can be derived from a more familiar "point-wise" version. Whether *interpersonal* deference principles hold is highly dependent on the debate between uniqueness and permissivism (e.g. White 2005; Schoenfield 2014; Horowitz 2014b, 2019; Greco and Hedden 2016; Schultheis 2018). Whether *rationalized* deference principles hold is highly dependent on debates around higher-order evidence (e.g. Williamson 2000, 2018; Christensen 2010b; Lasonen-Aarnio 2013, 2015, 2019; Elga 2013;

First, in our setup you don't know what claims are expressed by Calvin's 80%-opinions—q_i is simply the claim that *the ith claim on this test that Calvin was* 80%-*confident in (whatever that is) is true*. Thus you have virtually no evidence about the q_i. Meanwhile, Calvin has strictly more evidence than you about these claims—he knows all you do about the setup of the test, plus he knows *which* claims he was 80%-confident in, and therefore knows which facts bear on their truth. So conditional on Calvin's (more informed) evidence making it rational for him to be (on average) x% confident in these claims, it seems reasonable for you to be x% confident in it.

Second, there is a strong intuition that the calibration inference is *sensible*: it in principle makes sense to run calibration studies to test for overconfidence. As we'll see, whether this is so depends on whether a principle like Deference holds. (Note, for example, that the reason the calibration inference failed in Cases 1 and 3 from §2 is that Deference failed: I knew the outcome of the coin tosses and the answers to the trivia questions, so I didn't defer to your rational opinions.) Thus anyone who thinks the calibration inference makes sense in principle is under pressure to accept an epistemological theory that can support strong deference principles. This argument for such epistemological theories is one of the main philosophical upshots I'll be defending.

Turn to the second assumption needed to make Calvin's case analogous to Bianca's: independence. This says that once you learn the average rational confidence for Calvin to have in his 80%-opinions, learning about whether some of those opinions were true or false doesn't affect your confidence in the others. Precisely:

> **Independence:** Given that the average rational confidence for Calvin to have in his 80%-opinions is x%, further learning that certain of these opinions are true or false shouldn't affect your opinion in the others.
> For all $q_{i_0}, ..., q_{i_k}$: $P(q_{i_0}|\overline{R} = x, q_{i_1}, ..., q_{i_l}, \neg q_{i_{l+1}}, ..., \neg q_{i_k}) \;=\; P(q_{i_0}|\overline{R} = x)$

How plausible is Independence? Again, there is much more to be said, but it is well-motivated as a first approximation—after all, you know the test questions were selected randomly, so learning whether some are true or false shouldn't (significantly) affect your deference to information about Calvin's rational opinions on others.[17] (Note that the reason the calibration inference failed in Case 2 from §2 is that Independence failed—I knew that if the first toss landed heads then the rest of them would as well, so I didn't treat them independently.)

Deference and Independence imply that the calibration inference is warranted in

Horowitz 2014a; Salow 2018; Dorst 2019a,b). Deference will be a theorem in our setup given uniqueness plus higher-order certainty, but it will still be approximately true under a variety of weaker theories.

[17] This is at best approximately true, as learning that *all* of Calvin's other 80%-opinions were false should make you suspect that the test is tricky. What's definitely true is that the q_i are *exchangeable* (order doesn't matter) given $\overline{R}$. Using this we could prove more general versions of the formula derived §A.2 by using beta-binomial distributions rather than binomial ones. The reasoning will be similar, and the closer the q_i come to being independent, the stronger the calibration inference will be.

our simple scenario: learning that Calvin's 80%-opinions were (mis)calibrated provides strong evidence that they were (ir)rational. This is because the assumptions make the case analogous to Bianca's: "(average) rational confidence for Calvin" plays the same epistemic role for you as "(average) bias of Bianca's coins." Just as the bias inference goes through in Bianca's case because you should defer (independently) to the biases of the coins, likewise the calibration inference will go through in Calvin's case when you should defer (independently) to the rational opinions for Calvin. In particular, conditional on Calvin's 80%-opinions being on average rational, you should be quite confident that roughly 80% of them will be true; and conditional on his 80%-opinions being on average overconfident (say, the average rational confidence is 60%), you should be quite confident that less than 80% (roughy 60%) of them will be true. Therefore when you learn that a given proportion of these opinions are true, that provides you with strong evidence about what the (average) rational confidence for Calvin to have is—i.e. about whether his actual opinions are (on average) rational.

To give a simple example, suppose you are initially equally confident that the average rational confidence for him to have in his 80%-opinions is any of 60%, 61%,..., or 99%. Suppose there are 50 such opinions. Let's say he is *substantially overconfident* if the average rational confidence in his 80%-opinions is less than 75%. Then you are initially 37.5% ($\frac{15}{40}$) confident that he is substantially over-confident. But if you were to learn that 70% of those opinions were true (and—suppose—that there are 50 such opinions), then the calibration inference is warranted: you should become 78% sure that he is substantially overconfident in these opinions.[18]

Upshot: despite a variety of concerns, the calibration inference can be put on a firm theoretical foundation: when Deference and Independence hold, it is warranted.

By the same token, however, when Deference *fails*, the exact same reasoning will show that the calibration inference fails with it. For example, suppose that conditional on the average rational confidence being 80%, you should be *70%* confident in each of Calvin's 80%-opinions: $P(q_i|\overline{R} = 0.8) = 0.7$. Then (if Independence holds) you should be confident that if Calvin's rational, 70% of his 80% opinions will be true—and thus finding out that 70% of such opinions are true (he's slightly over-calibrated) will be evidence that he's *rational*, rather than overconfident!

Thus we arrive at the key result:

> **Deference is Key:** Given Independence, the tenability of the calibration inference in a given scenario stands or falls with the tenability of Deference.

So the crucial question is: how robust is Deference to variations in our simple scenario? §4 argues that it is very fragile: there are common scenarios in which Deference systematically, and hence we should not expect rational people to be calibrated. However,

[18]The general formula for this update is given in §A.2; the numbers in the text were calculated using a uniform distribution over the stated $\overline{R}$ hypotheses.

§5 argues that these failures of Deference and the corresponding rational deviations from calibration are in principle predictable—meaning that a more nuanced type of calibration study is possible.

4 The Limits

The real world isn't like the simple scenario, for you know a whole lot more about the test: its content, how it was constructed, what the experimenters were trying to show, what sorts of subjects were involved, and so on. Each of these bits of information threatens to undermine Deference and Independence in certain situations—and exploring the contours of these threats is important for having a full theory of the calibration inference. Here I'll focus on just one type of information that a calibration study inevitably provides: our subject's full calibration curve—and, therefore, their overall proportion of true answers. Call that proportion their **hit rate**. Does knowing the hit rate cause a problem for the calibration inference?

Yes. A deceptively simple line of reasoning suggests that it *always* breaks the calibration inference. That's not right. But a refined line of reasoning shows that it sometimes does. In particular, when learning of Calvin's hit rate does not in itself provide strong evidence about his rationality, the calibration inference can be inverted—over-calibration is evidence of *rationality*. Moreover, I'll argue that in real studies, hit rates often do not provide strong evidence about rationality, meaning the calibration inference *is* often inverted (§4.1).

Start with the deceptively simple line of reasoning. Note that once you learn Calvin's overall hit rate, our Deference condition no longer holds. To see this, take an extreme example: suppose you learn that the hit rate is 0%. Then conditional on the average rational confidence in his 80%-opinions being 80%, how confident are you that a given one of those claims is true? 0%—you're sure it's false. More generally, if you learn that the hit rate is abnormally low, then conditional on $\overline{R} = x\%$, you should often be *less* than $x\%$ confident in any such claim being true—for learning that the hit rate is abnormally low gives you information which Calvin didn't have, and so which wasn't incorporated into $\overline{R}$.[19] (Likewise if you learn that Calvin's hit rate is abnormally *high*, then conditional on $\overline{R} = x\%$, you should often be *more* than $x\%$ confident in each of Calvin's 80%-opinions.) Compare Bianca: if you learn that only 20% of the tablets we pulled for the test were marked 'heads', then you no longer should be confident that 80% of the tablets corresponding to 80%-heads-biased coins landed heads—it's likely that an unusual number landed tails, since that would help explain why such a low overall proportion landed heads.

[19] See §6.1 for discussion of cases where Calvin *does* know the hit rate.

Thus suppose you know that Calvin's hit rate is abnormally low, say 50%. Then conditional on his 80%-opinions being rational, you should only be (say) 70% confident that each one is true (i.e. $P(q_i|\overline{R} = 0.8, H = 0.5) = 0.7$, where $\boldsymbol{H}$ is Calvin's hit rate); and conditional on the average rational confidence being 70%, you should be (say) 60% confident that each such claim is true ($P(q_i|\overline{R} = 0.7, H = 0.5) = 0.6$). As we've seen—since Deference is Key—so long as Calvin's 80%-opinions are still (roughly) independent, this means the reasoning from §3 now shows that learning that 70% of his 80%-opinions are true is evidence that he's rational.

(So far, this reasoning is correct; here comes the misstep.) Thus if you learn that Calvin's hit rate is low and that only 70% of his 80%-opinions are true, you gain evidence that his 80%-opinions are rational.

The final step is the one that is not (necessarily) sound. It involves the inference from "given that Calvin's hit rate is low, over-calibration provides evidence of rationality"[20] to "learning that *the hit rate is low and Calvin is over-calibrated* provides evidence of rationality."[21] The first claim can be true while the second is false if learning that Calvin's hit rate is low is *itself* evidence that he is over-confident.

To see this, take an extreme case in which you know that the *rational* hit rate—the hit rate that Calvin would have if all his opinions were rational—is moderate. Then learning that his hit rate is abnormally low is itself strong evidence that his opinions are overconfident. Now consider learning things in stages: first you learn that Calvin's hit rate is low—say, 50%—and then you learn that he is (slightly) over-calibrated in his 80%-opinions. We've seen above that at the second stage—once you know of his low hit rate—learning of his slight over-calibration should make you more confident that his 80%-opinions are rational. But since you know the rational hit rate is moderate, when you learn at the first stage that his hit rate is low, your confidence in his 80%-opinions being rational should plummet. If the rise at the second stage is smaller than the drop at the first, then the total information will still be evidence that Calvin is overconfident.[22] Upshot: if you should think that Calvin's being overconfident makes it substantially more likely that he'll have a low hit rate, then the calibration inference can still go through when you know his hit rate.

So the deceptively simple reasoning is too quick. However, there *is* a real problem here. To see it, consider Bianca, and suppose you know that she can always tell whether the coin is biased in favor of heads or tails—though she may mis-estimate the strength of this bias, she always guesses the way the bias of the coin indicates. In that case, learning that she has a low hit rate is *no* evidence that she can't decipher the coins—

[20] Precisely, letting $\overline{\boldsymbol{q}}$ be the proportion of truths amongst Calvin's 80% opinions: $P(\overline{R} = 0.8|\overline{q} = 0.7, H = 0.5) > P(\overline{R} = 0.8|H = 0.5)$.

[21] Precisely: $P(\overline{R} = 0.8|\overline{q} = 0.7, H = 0.5) > P(\overline{R} = 0.8)$.

[22] Precisely: if $P(\overline{R} \approx 0.8|H = 0.5) << P(\overline{R} \approx 0.8)$, then even though $P(\overline{R} \approx 0.8|\overline{q} = 0.7, H = 0.5) > P(\overline{R} \approx 0.8|H = 0.5)$, we may still have that $P(\overline{R} \approx 0.8|\overline{q} = 0.7, H = 0.5) < P(\overline{R} \approx 0.8)$.

the only explanation of her low hit rate is that the coins happened to land in ways not indicated by the bias, meaning you should no longer defer to the biases of the coins.

Turn now to a version of Calvin's case in which you know that he will guess rationally, and hence that his actual hit rate equals the the rational hit rate. (Letting $\boldsymbol{H_r}$ be the *r*ational hit rate: $P(H = H_r) = 1$.) Suppose now you learn that Calvin's hit rate is low and that he's slightly over-calibrated in his 80%-opinions. Does the calibration inference go through? No.

You know that Calvin will have the rational hit rate. Moreover, whether his 80%-opinions are overconfident is independent of what the *rational* hit rate is (whether Calvin is irrational doesn't affect what you should think a rational version of him would do!), i.e. $P(H_r = 0.5|\overline{R} < 0.8) = P(H_r = 0.5)$. It follows that learning that Calvin's hit rate is low is no evidence that he's overconfident:

$$P(\overline{R} < 0.8|H = 0.5) \;=\; P(\overline{R} < 0.8|H_r = 0.5) \;=\; P(\overline{R} < 0.8).$$

And given *this*, the simple line of reasoning *does* go through. Learning that the test has a low (rational) hit rate breaks our deference condition—making it so that (e.g.) conditional on Calvin's 80%-opinions being rational, you should be only (say) 70% confident in each of them; meanwhile, conditional on him being overconfident in these opinions, you should be *less* than 70% confident in each of them:

$$P(q_i|\overline{R} = 0.8,\ H = 0.5) = 0.7$$
$$P(q_i|\overline{R} < 0.8,\ H = 0.5) < 0.7$$

It follows that given the low hit rate, learning that 70% of Calvin's 80%-opinions are true is evidence that those opinions are rational. And since in this scenario learning that the hit rate is low does *not* provide evidence that Calvin is overconfident, the net effect of the overall evidence—*Calvin's hit rate was low and his* 80%*-opinions were slightly over-calibrated*—is to provide evidence that he's rational. The calibration inference is inverted.

Example: again suppose you are initially equally confident that the average rational confidence for him to have in his 80%-opinions is any of $60\%, 61\%, ..., 99\%$, and there are 50 of them. Suppose that learning that his hit rate was 50% does not affect your confidence in any of these hypotheses, but it has the effect of tempering your deference in each downward by 10%: $P(q_i|\overline{R} = x, H = 0.5) = x - 0.1$ (so, for example, if his 80%-opinions are rational you should be 70% confident in each of them). Say that Calvin is *approximately rational* if $0.75 \leq \overline{R} \leq 0.85$. Then you are initially 27.5% ($\frac{11}{40}$) confident that he's approximately rational, but upon learning that 70% of his 80%-opinions are true (he's slightly over-calibrated), you should *increase* this confidence to 61%. Meanwhile, you should *decrease* your confidence that Calvin is substantially overconfident ($\overline{R} < 0.75$) from 37.5% to 22%, inverting the effect from the end of §3.

In summary, we've arrived at the following result:

> **Hit Rates are Key:** Whether the calibration inference is warranted in a context in which you learn Calvin's hit rate depends on whether a high (low) hit rate is strong evidence of overconfidence (underconfidence). If it is, the calibration inference is warranted; if it's not, the inference is weakened or inverted.

The crucial question, then, is whether, in realistic scenarios, learning that Calvin has an extreme (low or high) hit rate on a test provides good reason to think he's irrational.

4.1 Hit rates and rationality

I claim that often it does not. Two reasons.

The first is simple: we know that rational hit rates will vary widely across tests, as they are fully determined by how the test was constructed (§2). As such, absent reason to expect that the test construction will lead to a specific rational hit rate, we should not be very opinionated about what the rational hit rate will be. That means that hit rates in themselves are at best very noisy evidence about rationality—if a person's hit rate on a test is low, it may well simply be because the *rational* hit rate is low.

The second reason that hit rates often do not provide good evidence about rationality requires returning to the question of what the overconfidence hypothesis and its alternatives might be (§2). We've been simplifying by focusing on the "overconfidence effect"—in fact, many studies find wildly different calibration curves for different types of questions. Sometimes people are over-calibrated at all levels of confidence; other times they are over-calibrated at high levels of confidence and under-calibrated at low levels of confidence; other times they are under-calibrated at all levels of confidence, and so on (more on this in §5; see Koehler et al. 2002; Brenner et al. 2005). Translating these calibration curves to corresponding (ir)rationality hypotheses, the varying types of possibilities are shown in Figure 4. In this figure, interpret the lines as averages: for example, the "over-extreme" hypothesis says that when a person's actual confidence is 80%, the confidence it is on average rational for them to have is merely 60% (as indicated by the red dot).

With these options on the table, the live (ir)rationality hypotheses are claims of the form, "For questions of type X, people's confidence obeys (ir)rationality hypothesis Y", where X is some specification of question-type, and Y is a curve having a shape like those in Figure 4 (Brenner et al. 2005). For example, *ecological models* have proposed that if X is "questions sampled randomly from a natural domain," then Y is the rational curve (Gigerenzer et al. 1991; Juslin 1994); meanwhile, *case-based judgment models* have proposed (among other things) that if X is "questions on which case-specific evidence is statistically weak and the base rate of truths is moderate," then Y is the over-extreme curve (Griffin and Tversky 1992; Koehler et al. 2002; Brenner et al. 2005).

Here is an important prediction of any such (ir)rationality hypothesis: if the alternative claims someone is guessing between are from the same domain, then *people's guesses*

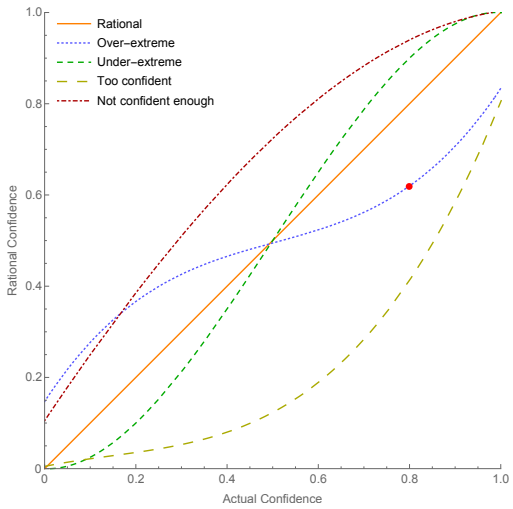

Figure 4: The Various (Ir)rationality Hypotheses

will tend to be rational.

Why? Note that all proposed (ir)rationality hypotheses have positive slopes, meaning that (on average) higher rational degrees of confidence correspond to higher actual degrees of confidence. Take any such (ir)rationality hypothesis, and consider a guess between a set of claims that it treats as in the same domain—say, "Which is bigger: Rome or Madrid?" The rational guess for Calvin is the one that he should assign higher confidence to: if $R(\textit{Rome}) > R(\textit{Madrid})$, it's rational for Calvin to guess *Rome*; and if $R(\textit{Rome}) < R(\textit{Madrid})$, it's rational for Calvin to guess *Madrid*. But since higher rational degrees of confidence correspond to higher actual degrees of confidence, the (ir)rationality hypothesis predicts that if the former, then Calvin's *actual* confidence will be higher in *Rome* than in *Madrid*—i.e. he'll guess *Rome*; and if the latter, Calvin's actual confidence will be higher in *Madrid*—i.e. he'll guess *Madrid*. Either way, the (ir)rationality hypothesis predicts that Calvin will guess rationally.[23]

Upshot: at least when people are guessing amongst claims that come from the same domain, all (ir)rationality hypotheses will agree that people's guesses will tend to be rational. People's hit rate is fully determined by their guesses—so if their hit rate is low, then (since their guesses will tend to be rational) this means that the *rational* hit rate is low as well. In other words, in many studies it is common ground amongst all

[23]Formally, let $C(q)$ be Calvin's actual confidence in q, and let an (ir)rationality hypothesis be a function $f : [0,1] \to [0,1]$ mapping actual degrees of confidence to (average) rational degrees of confidence: $R(q) = f(C(q))$. Any such function that is monotonically increasing ($f(x) > f(y)$ iff $x > y$) will be such that if $R(q) > R(p)$, then $f(C(q)) > f(C(p))$, hence $C(q) > C(p)$. Notably, since f is most plausibly interpreted as an average, there will be exceptions to this connection between rational and actual guesses. How common such exceptions will be depends on (1) how steep the slope of the (ir)rationality hypothesis is, and (2) how widely the deviations from f are distributed.

(ir)rationality hypotheses that we are in a situation in which we know that Calvin's hit rate will be (close to) the rational hit rate—and, therefore, that his hit rate does *not* provide strong evidence about whether he is rational. As stated in Hit Rates are Key, this is precisely the situation in which the calibration inference is weakened or inverted.

Let's sum up the results of §4. One piece of information that threatens to break the calibration inference in practice is the subjects' hit rate. I've shown that when we have reason to think that Calvin's hit rate will be close to the rational hit rate, then the calibration inference can be weakened or inverted when he has a low (or high) hit rate. And since we can now see that we often *do* have reason to think this, the strength of the calibration inference in practice is called into question.

5 The Implications

At this stage we've established that we often should not expect rational opinions to be calibrated when actual people's hit rates turn out to be low (or high). This has an important implication: in such contexts, we cannot evaluate the overconfidence hypothesis by simply checking whether people's opinions are calibrated—for we shouldn't expect *rational* opinions to be calibrated. That is my core critique of calibration studies. Such a methodology is standard practice; I claim that it is untenable.

What, then, would a sound methodology look like? Since the mistake is to assume that rational opinions will be calibrated, the solution is to make well-founded predictions about how and when rational opinions will *mis*calibrated. In particular, we must:

1) Build a plausible model of our test-construction procedure.
2) Use that model to predict the rational deviations from calibration on our test.
3) Compare those predictions to the actual calibration curves we observe.

This is the methodology I propose, and that am going to start to pursue in this section. After first getting clear on two of the core empirical factors that affect calibration curves (§5.1), I will use the above methodology to argue that these effects are to be expected given the rationality hypothesis (§5.2). I'll close in §6 by taking stock of some further empirical effects, and how the theory developed here may bear on their interpretation.

5.1 The hard-easy and base-rate effects

It turns out that the "overconfidence effect" is an overgeneralization: it is not the case that people are in general over-calibrated on binary-question tests. Rather, we can distinguish the tests that are *hard* from those that are *easy* based on the hit rate: an easy test is one with a hit rate of at least 75%; a hard test is one with a hit rate of less than 75%. The empirical generalization that subsumes the "overconfidence effect" is called the **hard-easy effect:** people tend to be over-calibrated on hard tests and

under-calibrated on easy tests—see the left side of Figure 5. (The reason we see the "overconfidence effect" on general-knowledge trivia tests is simply that most such tests turn out to be hard.) The hard-easy effect has been called "fundamental bias in general-knowledge calibration" (Koehler et al. 2002, 687), and is widely cited as one of the core pieces of evidence in favor of the overconfidence hypothesis (e.g Lichtenstein et al. 1982; Keren 1987; Gigerenzer et al. 1991; Griffin and Tversky 1992; Juslin 1994; Juslin et al. 2000; Koehler et al. 2002; Brenner et al. 2005; Hoffrage 2004; Moore and Healy 2008; Glaser and Weber 2010). The standard interpretation is that people do not make sufficient adjustments for task difficulty, leading them to be overconfident on hard tests and underconfident on easy ones.

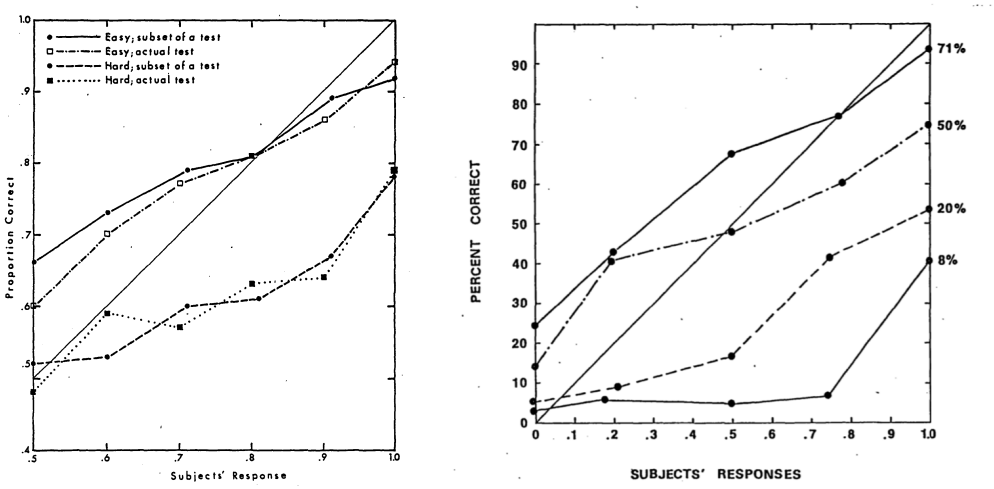

Figure 5: **Left:** The hard-easy effect. Top curves are easy tests; bottom curves are hard ones. **Right:** The base-rate effect. Base-rates generating different curves are labeled on the right. (Figures are from Lichtenstein et al. 1982.)

The second result I'm going to focus on is called the **base-rate effect**. It applies to tests of a slightly different format. Rather than the familiar binary-question tests—wherein subjects choose an answer from two options—the format is a *propositional test*, wherein subjects are presented with a series of claims and their job is simply to rate their confidence in those claims between $0 - 100\%$. The base-rate effect is the observation that on propositional tests, the overall proportion of truths on the test (the *base rate*) has a dramatic effect on people's calibration curves. In particular, low base rates tend to lead to over-calibration across much of the $0 - 100\%$ scale, high base rates tend to lead to under-calibration across much of the scale, and middling base rates often lead to under-calibration in low-confidence opinions and over-calibration in high-confidence opinions—see the right side of Figure 5.

The hard-easy and base-rate effects are two of the core pieces of evidence offered in

favor of various versions of the irrationality hypothesis. However, we now know that systematic patterns of miscalibration should sometimes be expected of *rational* people when the hit rate or base rate varies.[24] The question, then, is whether these effects would be surprising if people were being rational.

5.2 The rational predictions

How can we know what to expect rational calibration curves to look like on tests of various types? The way we're going to answer this question is by returning to our coin analogy with Bianca. We are now going to assume that she *can* decipher the tablet markings—and thus set her confidence equal to the biases of the coins—and go on to simulate what calibration curves we should expect from her as we vary the method of constructing the test, along with the difficulty or base rate. (Analogy: if we assume that Calvin can decipher his evidence so respond rationally, what should we expect his calibration curve to look like under various conditions?)

First, suppose the tablets are truly randomly distributed, meaning when we draw tablets from a given archive and show them to Bianca, we are in effect taking a random sample from all the original tosses of the coins. (Analogy: our test is taking a random sample from all the various opinions warranted by Calvin's evidence.) Model this by tossing each coin of varying biases (between $50-100\%$ for binary-question tests[25]; between $0-100\%$ for propositional tests) a random number of times, having Bianca announce her confidence that each will land heads, and record her calibration curve on that test. This is a single trial. Repeat this procedure thousands of times, and now look at the average results on tests that have various hit rates—what do we expect to see?

For all simulations, I'll display two versions. The **perfection model** assumes Bianca always gets the biases of the coins exactly right (analogy: Calvin is always perfectly rational). The **noise model** assumes that Bianca's announced confidence is a random perturbation of the bias of the coin—capturing the idea that she may be a reliable but imperfect at deciphering the coins' biases (analogy: Calvin's confidence may be a reliable but imperfect tracker of the degree of confidence his evidence warrants). The most plausible rationality hypotheses are ones in which there is some such error.[26]

[24]Note: learning the base-rate of a test breaks the Deference condition in the exact same way that learning the hit rate does (§4). Moreover, unlike the hit rate, there is no possibility that this base rate is evidence of (ir)rationality, since subjects have no control over the base rate.

[25]In the binary-question setting I simplify by tossing coins of biases $50-100\%$ and having her always announce heads, rather than tossing coins of biases between $0-100\%$ and having her first guess whether the coin lands heads or tails. The underlying statistics are the same.

[26]In all noise models below I assume the errors are normally distributed with mean 0 and (for illustration) standard deviation 0.2. This model takes inspiration from "error models" (Erev et al. 1994; Pfeifer 1994), but the interpretation is importantly different. Their models treats people's reported opinions as imperfect indicators of their *true* opinions, whereas mine treats people's reported opinions as imperfect indicators of the *rational* opinions. It is plausible that the latter errors will be larger than

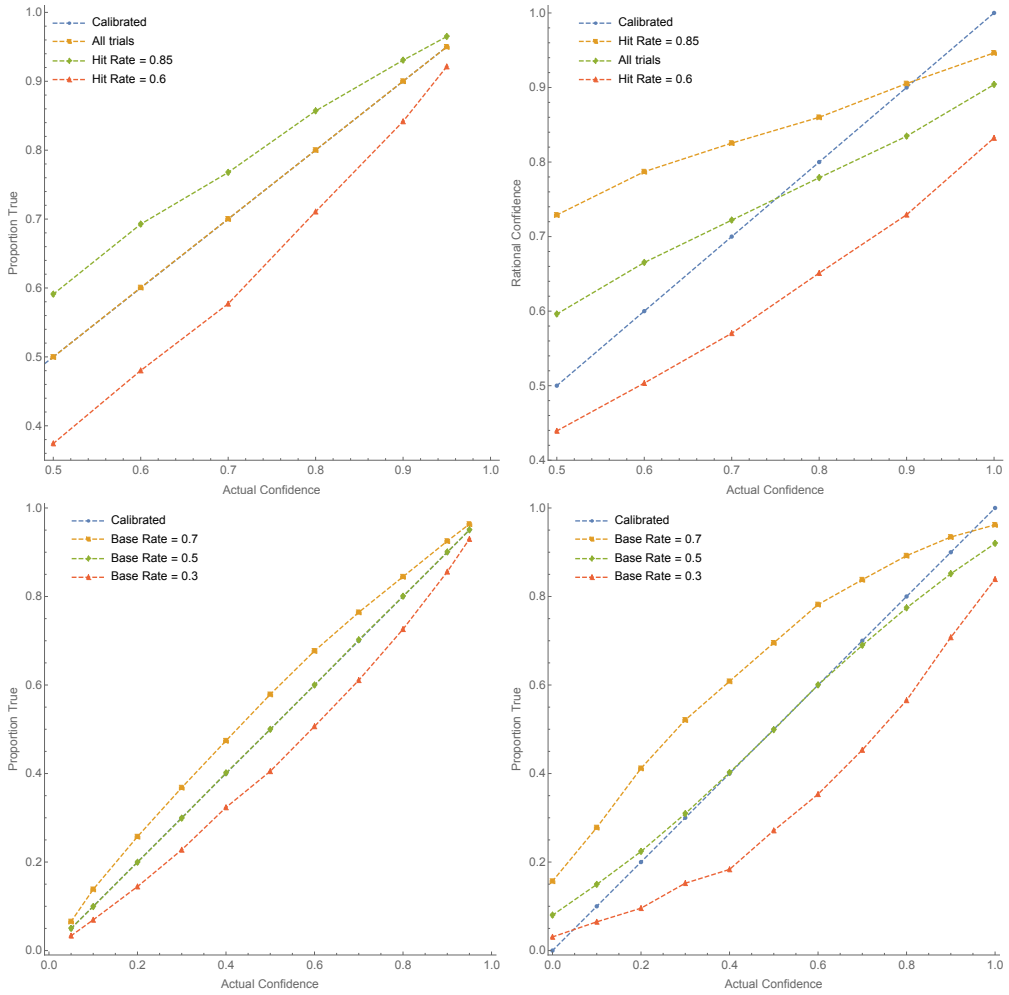

Figure 6: Random tests. **Top:** Binary-question, restricted to various hit rates. Left is perfection model (100,000 trials); right is noise model (50,000 trials). **Bottom:** Propositional, restricted to various base rates. Left is perfection model (2,000,000 trials); right is noise model (500,000 trials).

The expected calibration curves for Bianca at various hit rates and base-rates is displayed in Figure 6. In both binary-question and propositional tests, when we consider all trials together, Bianca is calibrated—perfectly so in the perfection model; slightly less so in the noise model due to "scale-end effects" (Juslin et al. 2000)—at the end-points of the confidence scale, errors can only go in one direction, resulting in the slight tilting

the former (it's harder to know what you *should* think than to know what you *do* think!). Moreover, while tests of the variance of people's reports have suggested that error in reporting their true confidence cannot account for the observed miscalibration (Budescu et al. 1997), these tests do not test for error in matching their reported confidence to the rational confidence.

of the curve.[27] But amongst tests where the hit rate or base rate is low (high), Bianca tends to be over-(under-)calibrated—just as observed empirically with the hard-easy and base-rate effects. Why?

Consider a given trial on which the proportion of heads was lower than usual. Why was it lower? One explanation is that this trial had an abnormally large proportion of coins that were biased against landing heads. A different explanation is that this test was unusual in the sense that more of the coins landed tails more often than you would expect, given their biases. Both are likely to play a role in any given trial with a low hit/base-rate. Bianca will account for the first factor in setting her degrees of confidence, since she can recognize the coins and see that more of them than usual have a low bias—but it is impossible for her to account for the second factor. The result? As we consider cases with more extreme hit/base-rates, Bianca will become increasingly miscalibrated. For example, take the perfection model—where Bianca is as sensitive to the biases of the coins as she could possibly be. On the binary-question test, on trials with a hit-rate of 75%, Bianca's average confidence was 75%; on trials with a hit rate of 90%, her average confidence is 77% (becoming under-calibrated); and on trials with a hit rate of 60%, her average confidence is 72.7% (becoming over-calibrated).

Upshot: even if the calibration tests contain questions that are random samples of the overall distribution of rational opinions (the best-case-scenario for the calibration inference, as seen in §3), we would still expect some form of the hard-easy and base-rate effects to emerge for rational subjects. Moreover, if they are merely approximately rational (the noise model), we should expect rational calibration curves that are qualitatively similar to the curves we observe empirically (compare the right side of Figure 6 to Figure 5).

The main reason to be worried about this particular model is the following: in these simulations, it is incredibly rare to deviate far from the mean hit rate (75% on binary-question tests) or base rate (50% on propositional tests). This is because we are randomly drawing coins of biases between $50-100\%$ (or $0-100\%$), meaning we expect to get an average bias around the midpoint and so are very likely to get a hit rate around 75% (or a base rate around 50%). This is like randomly sampling claims from all rational opinions between $50-100\%$ (or $0-100\%$ confidence), meaning we should (by Deference and Independence) be very confident than around 75% (or 50%) of the rational answers will be true. For example, in both binary-question models only around 0.06% had hit rates as low as 60%. Since such hit rates are not uncommon in real calibration tests (I constructed one with a hit rate of 44% on my first try!), it is not plausible that such

[27] In the binary-question simulations, I maintained the assumption that Bianca guesses rationally even in the noise model—so noise never pushes her confidence below 50%. This is appropriate in contexts in which the *direction* of the evidence (which option it favors) is unlikely to be mistaken, but the *force* of of the evidence (how strongly it favors it) is. It's worth investigating variations on this assumption.

tests are truly constructed from random questions.[28]

So we need some other explanation of the variability in hit- and base-rates. I'll close by offering two; the first is meant to address tests that choose a particular subject-matter and then present subjects with questions about that subject-matter (e.g. Dunning et al. 1990; Vallone et al. 1990; Brenner et al. 1996; Gigerenzer et al. 1991; Koehler et al. 2002; Brenner et al. 2005; Hoffrage 2004; Glaser and Weber 2007; Merkle and Weber 2011). The second is meant to apply to tests—like my questionnaire—that contain seemingly random and unrelated claims not drawn from a particular subject-matter (e.g. Lichtenstein et al. 1982; Juslin 1994; Ludwig and Nafziger 2011).

The idea behind the first model is this. Although we expect that the rational opinions will *on the whole* be calibrated, we should be aware that there will be random fluctuations in how calibrated they are across subject-matters. In some domains, a person's evidence will warrant misleadingly strong opinions (only 70% of their rational 80%-opinions are true); in others, it will warrant misleadingly weak opinions (90% of their rational 80%-opinions are true). We expect (perfectly rational) random fluctuations in these degrees of misleadingness. Moreover, it's plausible to expect that these random fluctuations will be correlated—for example, if only 50% of the rational 60%-opinions are true, it's likely that only 60% of the rational 70%-opinions are true.

Model: again there is a random number of coins of varying biases that Bianca can recognize, but this time there is random variation across tablet archives in how representative they are of the broader distribution of tablets—some archives tend to have the proportions of heads of a given type of tablet than would be expected from their bias; others tend to have lower proportion; others higher. Thus for each trial (visit to an archive), we generate a random misleadingness parameter and add it to the coin biases to determine how far the proportions of heads in this archive deviates from the biases of the coins.[29]

What should we expect rational calibration curves to look like across various hit/base rates on this way of constructing tests? The results are displayed in Figure 7. Again, we see a clear (and realistic) hard-easy effect and base-rate effect. Moreover, this model easily allows widely-varying hit rates—for example, 17.5% of binary-question trials had hit rates below 60%, and 15.6% of propositional trials had base rates as low as 30%. Thus the random-misleadingness model is a plausible explanation of how we can find such widely-varying hit/base rates in practice.

[28]Does the rarity of extreme hit- and base-rates in this model provide evidence that calibration tests *are* taking random samples of rational opinions, but that people's irrationality explains the variance? No. In propositional tests the base rates are set by the experimenters—yet in 2,000,000 trials, I found no instances of a base rate as low as 15%. Clearly researchers must be selectively generating tests with particular base rates to get ones as extreme as 8% (Figure 5).

[29]In the displayed simulations this parameter is normally distributed with mean 0 and standard deviation 0.2. Again, in binary-question simulations I assume that the variation in misleadingness is only in the magnitude—not the direction—of the evidence, so it never pushes below 50%.

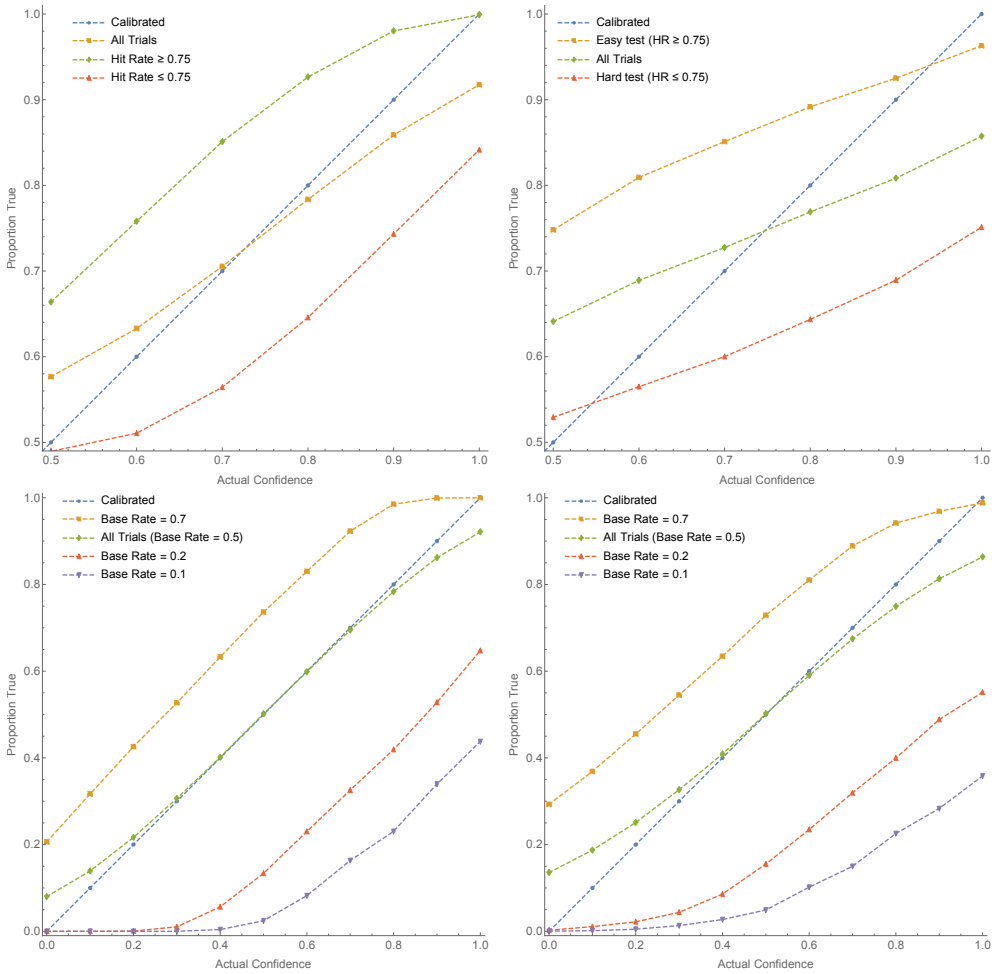

Figure 7: Tests with random misleadingness. **Top:** Binary-question. Left is perfection model; right is noise model (10,000 trials each). **Bottom:** Propositional. Left is perfection model, right is noise model (100,000 trials each).

Finally, turn to tests (like mine) that are not unified by any subject-matter. How could one generate tests like this with widely varying hit/base rates? A natural strategy is to randomly generate questions, but apply scrutiny to them as you do so—only keeping the ones that fit with the sort of test (hard or easy; high or low base rate) you want to end up with.

Model: we have a desired hit rate (or base rate) for our test of Bianca. For binary-question-tests: we randomly pull a tablet and predict which way Bianca will guess.[30]

[30]The binary-question model assumes we can predict which way she'll guess. This is a simplification, but not an outlandish one since we often know a fair bit about people's tendencies to guess on trivia questions—after all, we often share their evidence, and we know how *we* are inclined to guess.

If we predict that adding the tablet will move the overall difficulty toward (or keep it close to) our desired difficulty, we include it; if not, we discard it and draw again. (For propositional tests: we do the same procedure but without the guesswork—simply noting whether including the claim would move us toward the desired base rate.)

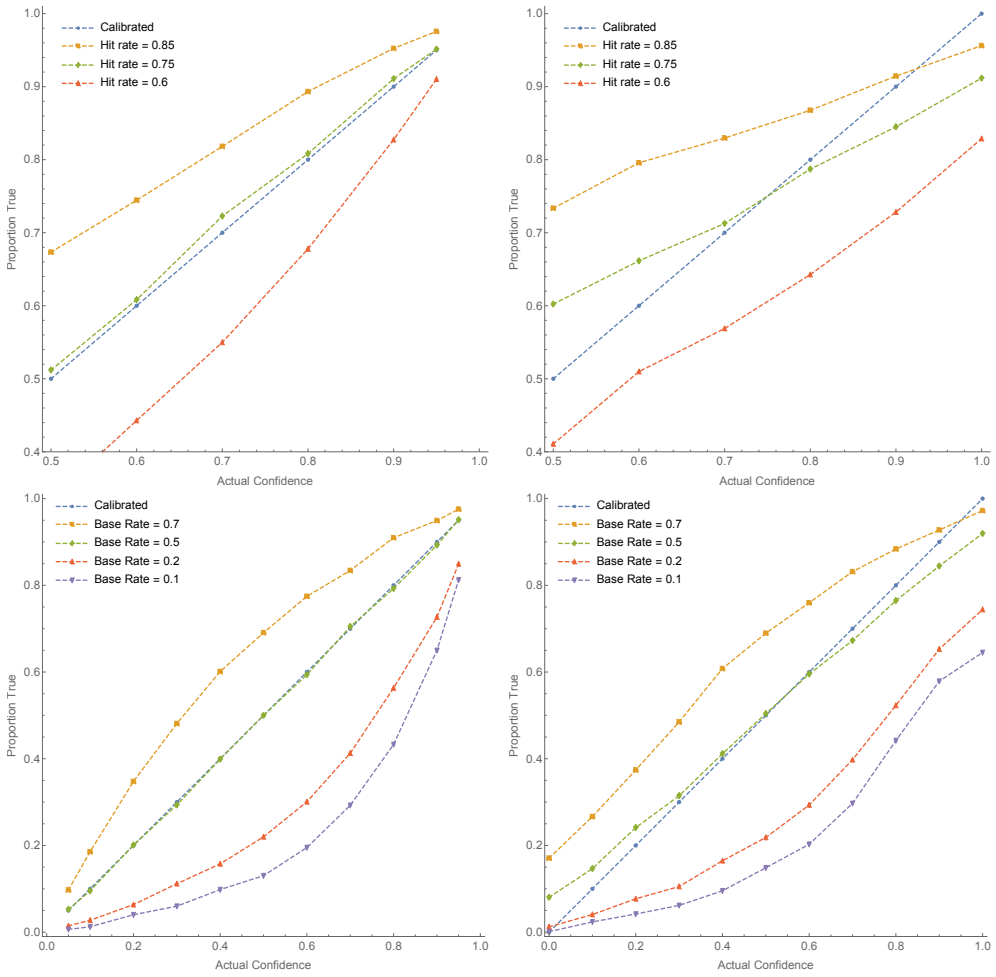

Figure 8: Random tests with scrutinized questions. **Top:** Binary-question. Left is perfection model; right is noise model (1,000 trials each). **Bottom:** Propositional. Left is perfection model; right is noise model (1,000 trials each).

The results for what we should expect rational calibration curves to look like across various hit/base rates are displayed in Figure 8. Again, we see realistic hard-easy and base-rate effects; and again, there is no problem with reaching widely varying hit rates (every trial achieved the desired hit/base rate to within 10% precision).

Upshot: I have given several models that take account of the fact that we should not

always expect rational people to be calibrated on the tests we give them. In particular, we've seen that in all such models, we should expect rational people's calibration to be systematically affected by the hit or base rate of the test—and in some of them, we should expect effects qualitatively quite similar to the empirically observed hard-easy and base-rate effects. As such, as things stand it is unclear whether these effects provide any evidence for irrationality. To properly assess this, studies must explicitly model their own test construction and the expected rational calibration curve that will result—only then can we make a fair comparison to see how far people's actual calibration curves are from what we would expect them to be if they were rational.

In closing, suppose we apply this methodology to *my* survey. I constructed it by randomly thinking of test questions, seeing how I was inclined to guess, and throwing them out if my guesses were accurate—so the appropriate rational model is the final one (random questions with scrutiny). Running the model with the observed hit rate (44%) yields the predicted rational calibration curve in Figure 9. Clearly we shouldn't put much weight on this particular result—it was a small ($n = 50$), informal survey. But it does illustrate that understanding how rational opinions can deviate from perfect calibration has the potential to reverse the standard interpretation of empirical effects.

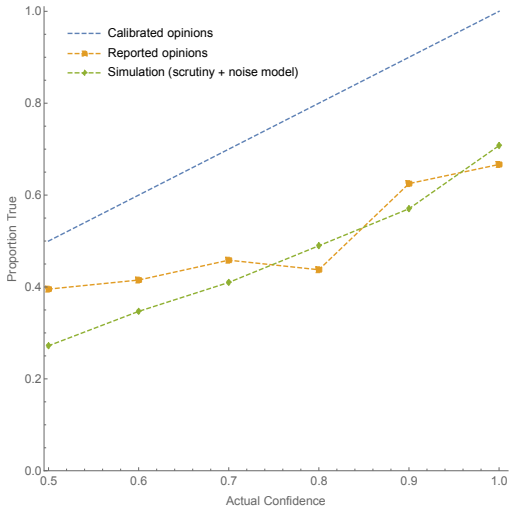

Figure 9: Simulation of rational and actual calibration curve in my survey (noise model).

6 The Further Effects

In this final section I'll briefly address how I take the theory developed in this paper to bear on the interpretation on several other empirical effects: base-right neglect (§6.1),

the Dunning-Kruger effect (§6.2), and the finding that interval estimates are often "too tight" (§6.3). Those mainly interested in the big picture may skip to §7.

6.1 Base-rate neglect?

The standard hard-easy and base-rate effects occur when subjects are not told about the difficulty or base-rate of the test they are taking—§5.2 shows what to expect of rational subjects in such a situation. But what should we expect of rational subjects when they *are* given such information?

According to the theory developed here, the difference is potentially crucial. What drives the calibration inference is Deference: upon learning how confident it is (on average) rational for Calvin to be in certain claims, you should adopt that level of confidence in those claims. Deference is plausible when you know Calvin has more relevant information and evidence about those claims than you do; it fails when you know things (such as the hit/base-rates) that he does not. Thus if you are certain that Calvin knows everything that you do—including the relevant hit/base-rate information—then you should still obey Deference, you should expect him to be calibrated if he's rational, and the calibration inference should be warranted.

What, then, should we make of studies that *do* attempt to provide hit/base-rate information to their subjects? First of all, if the theory developed here is right then such information *must* be provided in order for the calibration inference to be warranted—all studies should be conducted in this way. When they are, they do sometimes find that people are poorly calibrated (e.g. Kahneman and Tversky 1973; Dunning et al. 1990; Griffin and Tversky 1992; Koehler et al. 2002; Brenner et al. 2005). What to make of such results?

A straightforward interpretation is that people have an irrational tendency to ignore base rates—they exhibit "base-rate neglect." But this interpretation overgeneralizes: a wide variety of research show that people are often exquisitely—or even *overly*—sensitive to base rates (Peterson and Beach 1967; Edwards 1982; Tenenbaum and Griffiths 2006). Thus the plausible irrationalist interpretations have the form, "On questions from domain X, people tend to ignore base rates." Once we make this refinement, however, we see that a parallel rationalist interpretation is possible: "On questions from domain X, people tend to have misleading evidence about the probative value of base rates."

After all, the rational import of base-rate information is highly dependent on what evidence you have about the relevance of such base-rates to the case at hand. For example, if you know that personality is a weak indicator of career choice, then the fact that there are only a few lawyers in the room (a low base rate) should make you doubtful that Laura is a lawyer—even if she seems to have a stereotypical lawyerly personality. However, if you've been told that personality is a very strong indicator of career choice, then the fact that there are only a few lawyers in the room should have

very little dampening effect on your confidence that Laura is a lawyer—the evidence from her personality swamps the base-rate information. Thus the rational response to base rates depends on your background evidence. And—we now know—the mere fact that subject's are *wrong* about the probity of base-rate information in contexts like these does not show that they are *irrational* in processing such information.

What this means is that Deference can fail even when Calvin knows all that we know about the base rates, since we should be unsure what his information supports about the probity of such base rates. This in turn means that there are two different stories we can tell about instances of base-rate neglect. The irrationalist story says that people foolishly neglect base-rates in domain X, and that is why they are miscalibrated; the rationalist alternative says that people have misleading information about the probity of base-rates in domain X, and *that* is why they are miscalibrated. I'm not claiming that the rationalist explanation is better. But it would explain the miscalibration data, so it is a live option that is worth exploring.

6.2 The Dunning-Kruger effect

The Dunning-Kruger effect (Kruger and Dunning 1999) is the much-maligned finding that those who are comparatively unskilled in a given domain are also unable to accurately assess how comparatively unskilled they are. Precisely: the gap between a person's relative performance on a test (which percentage of test-takers did they outperform?) and their *estimate* of this number grows as relative performance decreases. For example, those in the 50th percentile may estimate that they are in the 60th percentile, while those in the 20th percentile may estimate that they are in the 50th percentile. This finding is routinely chalked up to a cognitive bias—a failure of the metacognitive ability to assess how competent one is (Dunning 2012). As far as I can tell, this is a mistake. We have already seen that for any set of rational opinions, there will be tests that are hard and easy for those opinions—in particular, that will lead to low or high hit rate (§2). We have also seen that in any such test, as the test gets harder for a rational person, they will become increasingly over-calibrated—meaning the gap between performance (actual hit rate) and estimated performance (average confidence, i.e. estimated hit rate) will grow (§5.2).

Upshot: Since even for rational people, the difficulty of a test will vary depending on their knowledge and skills, a straightforward consequence is that rational people who perform less well on a test will over-estimate their performance more than rational people who perform better. The Dunning-Krueger should be expected of them.[31]

[31] Compare to Moore and Healy (2008), and contra Merkle and Weber (2011)—who illicitly assume that Bayesians will have priors that match the objective frequencies on the test, and so be calibrated.

6.3 Interval estimation

A different type of calibration test asks people to state various confidence intervals for the true value of some unknown parameter, such as the length of the Amazon. Some claim that such interval tests reveal a different, more robust kind of overconfidence than the studies mentioned in the text (Moore and Healy 2008; Glaser and Weber 2010; Ortoleva and Snowberg 2015). I think that this is incorrect.

First note that the two types of test are inter-translatable (Tversky and Kahneman 1974): your 90% confidence interval for the length of the Amazon is "1000 to 5000 miles" iff you are 95% confident in both "The Amazon is at *L*east 1000 miles long" (L), and "The Amazon is at *M*ost 5000 miles long" (M); your confidence intervals tend to miss the true value too often iff you are over-calibrated on claims like L and M. Now, for someone to be calibrated in their interval estimates, items must fall outside the range of their 90%-confidence interval exactly 10% of the time. The result that supposedly supports the claim that people are more over-calibrated in interval-estimation is that some studies find "miss rates" as high as 50%, and almost never lower than 10% (Glaser and Weber 2010, 243). However, once we apply the translation, we see that these miss rates are no more extreme than we would expect given the standard hard-easy effect—and since the hard-easy effect is predicted of rational people, likewise are these miss rates for interval estimates.

Note that for hard binary-question tests, it is standard to see less than 75% of people's 95%-opinions being true—in fact, all of our rational noise models predict at least that much over-calibration that when hit rates are low. Now, by our above translation, Calvin's 90%-confidence interval for the Amazon's length in miles is "1000 to 5000" iff he is 95% confident in both"The Amazon is at *L*east 1000 miles" (L) and "The Amazon is at *M*ost 5000 miles" (M). The probability of both L and M being true (his interval covering the true value) is the probability of the former, multiplied by the probability of the latter given the former: $P(L \wedge M) = P(L) \cdot P(M|L)$. Supposing we expect a 75% hit rate for these two 95%-opinions L and M (so $P(L) = P(M) = 0.75$), then if they were independent we would expect a hit rate of $0.75 \times 0.75 = 56.25\%$ for their conjunction, i.e. a miss rate of around 44%. But they are (by definition) *not* independent: if L were false (the Amazon is less than 1000 miles), M would necessarily be true; hence learning that L is true necessarily lowers the probability of M: $P(M|L) < P(M) = 0.75$. Thus we should expect a hit rate for the conjunction $L \wedge M$ of *less* than 56.25%, and hence a miss rate of greater than 44%—50% is not at all unreasonable. Moreover, by parallel reasoning we should expect less-than-10% miss rates only if *more* than 95% of a person's 95%-opinions are true. Yet we've seen that (due to scale-end effects) this is virtually never the case (none of our binary-question noise models—even with easy tests—see such high rates). Hence both of these interval-estimate observations are unsurprising given the hard-easy effect—which we've seen is itself predicted by the rationality hypothesis.

7 The Upshots

Many have taken the results of calibration studies to support the view that people tend to be overconfident, but the theoretical foundations of this inference are shaky (§2). I've argued that we can secure them (§3)—but in doing so we find that there are systematic reasons why the calibration inference will often fail (§4). To account for this, we must predict the rational deviations from calibration on our tests before assessing people's actual calibration curves—and we've seen that doing so has the potential to overturn the standard interpretation of robust empirical effects (§§5–6). I want to close by defending what I think are three upshots of this discussion.

First, philosophers have had an extended debate over whether the proper theory of rationality will vindicate interpersonal, rationalized deference principles like Deference (see fn. 16). A straightforward consequence of the theory developed here is the following: if you think that learning that someone is (mis)calibrated, in general, provides evidence that they are (ir)rational—that is, if you think calibration studies are not hopelessly flawed—then you must endorse a theory of rationality that supports strong deference principles.

Second, while psychologists have standardly motivated calibration research from an irrationalist methodology, the theory developed here shows that a different, *rationalist* approach is also possible. All parties agree that if people's judgments are systematically wrong about a given subject-matter, that is a problem. The standard explanation is that people handle questions in that subject-matter irrationally—they are overconfident, and that is why they are so often wrong. The alternative explanation is that people handle questions as well as can be expected—but their evidence about that subject-matter is misleading, and that is why they are so often wrong. Both explanations fit with the calibration data. Both agree that calibration studies provides reason for people to become less confident in some of their opinions. Both agree that calibration research is necessary to identify *which* such opinions. But the latter—like all rationalist explanations—has an advantage. It does not require explaining why the most fantastically complex computational system in the known universe—a system that every minute of every day performs feats that the world's best engineers using the world's best supercomputers cannot duplicate (Pinker 2009)—would make systematic, predictable, and easily correctable errors in the ever-present and all-important domain of reasoning under uncertainty.

Finally, both psychologists and philosophers have been investigating questions about rationality—but often from radically different directions, and without substantial discussions. We've seen that the questions, methods, and tools from these investigations can be tied together in surprising and fruitful ways. That raises an exciting question: If we bring these investigations closer together, what other ties might we find?[32]

[32]Thanks to Lyle Brenner, Liam Kofi Bright, Thomas Byrne, Chris Dorst, Brian Hedden, Dmitri Gallow, Cosmo Grant, Harvey Lederman, Matt Mandelkern, Bernhard Salow, Ginger Schultheis, James

Appendix

A.1 Deriving Deference

Recall that $q_1, ..., q_n$ are the claims that Calvin assign 80%-confidence to, that R is the rational probability function for him to have overall, and that $\overline{R}$ is the average rational confidence in the q_i: $\overline{R} := \sum_{i=1}^{n} \frac{R(q_i)}{n}$. Recall Deference:

> **Deference:** Upon learning only that the average rational confidence for Calvin to have in his 80%-opinions is x%, become x% confident in each of them.
> For all q_i: $P(q_i|\overline{R} = x) = x$.

(For simplicity of notation I maintain focus on Calvin's 80%-opinions. Obviously, parallel principles and reasoning apply to the others thresholds.)

Deference follows from two further principles:

> **Point-wise Deference:** Upon learning the rational credence function for Calvin is δ, become $\delta(q_i)$-confident in each q_i.
> For all q_i : $P(q_i|R = \delta) = \delta(q_i)$.[33]

> **Equality:** Upon learning only that the average rational confidence for Calvin to have in his 80% opinions is x%, be equally confident in each of them.
> For all $q_i, q_j : P(q_i|\overline{R} = x) = P(q_j|\overline{R} = x)$.

Since Equality is extremely plausible in the situations we're considering (where you don't know anything more about the q_i than they they were claims that Calvin was 80% confident in), this shows that Deference follows from the more familiar Point-wise version.

To prove this, for any random variable X (a function from possibilities to numbers), let $\mathbb{E}[X] := \sum_t P(X = t) \cdot t$ be your rational expectation of X. (Assume a finite state space, for simplicity.) Note that $\overline{R}$ is a random variable; also note that if $I(q_i)$ is the indicator variable for q_i (1 if q_i is true, 0 otherwise), then $\mathbb{E}[I(q_i)] = P(q_i)$. Let $D_x = \{\delta_1, ..., \delta_k\}$ be the set of possible values of R such that $\sum_{i=1}^{n} \frac{\delta_j(q_i)}{n} = x$, so that $\overline{R} = x \Leftrightarrow R \in D_x$.

First, focus on your expectations of the proportion of truths, conditional on $\overline{R} = x$:

$$\mathbb{E}[\textstyle\sum \frac{I(q_i)}{n} \mid \overline{R} = x] \;=\; \sum_{\delta \in D_x} P(R = \delta |\; \overline{R} = x) \cdot \mathbb{E}[\textstyle\sum \frac{I(q_i)}{n} \mid R = \delta]$$

Shaw, and audiences at the Universities of Bristol, Pittsburgh, Oxford, and Sydney, and at MIT, for much helpful feedback and discussion.

[33] Here 'δ' is a rigid designator for a particular probability function (an assignment of numbers to propositions), whereas R is a definite description for "the rational credence function for Calvin, whatever it is"—so R can vary across possibilities but δ cannot.

By linearity of expectations, this equals

$$
\begin{aligned}
&= \sum_{\delta \in D_x} P(R = \delta |\ \overline{R} = x) \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[I(q_i)|\ R = \delta] \\
&= \sum_{\delta \in D_x} P(R = \delta |\ \overline{R} = x) \cdot \frac{1}{n} \sum_{i=1}^{n} P(q_i|\ R = \delta) && \text{(Definition)} \\
&= \sum_{\delta \in D_x} P(R = \delta |\ \overline{R} = x) \cdot \frac{1}{n} \sum_{i=1}^{n} \delta(q_i) && \text{(Point-wise Deference)} \\
&= \sum_{\delta \in D_x} P(R = \delta |\ \overline{R} = x) \cdot x && \text{(Definition of } D_x\text{)} \\
&= x.
\end{aligned}
$$

Therefore $\mathbb{E}[\sum \frac{I(q_i)}{n} \mid \overline{R} = x] = x$, so by linearity of expectations, your average rational credence in the q_i equals x: $\frac{1}{n} \sum_{i=1}^{n} P(q_i|\overline{R} = x) = x$. By Equality, since each of the values in this sum is equal, they must all be equal to x—therefore for all q_i: $P(q_i|\overline{R} = x) = x$, establishing Deference.

A.2 The calibration-inference formula

Here I show how to calculate what your posterior confidence should be that Calvin is overconfident in his 80%-opinions when Deference and Independence hold, you know that there are n such opinions, and you learn how (mis)calibrated they are. Recall:

Deference: For all q_i: $P(q_i|\overline{R} = x) = x$.

Independence: For all $q_{i_0}, ..., q_{i_k}$: $P(q_{i_0}|\overline{R} = x, q_{i_1}, ..., q_{i_l}, \neg q_{i_{l+1}}, ..., \neg q_{i_k}) = P(q_{i_0}|\overline{R} = x)$

Suppose you initially leave open that $\overline{R}$ will be any of $t_1, ..., t_m$, with prior probabilities $P(\overline{R} = t_i)$. Note that Deference and Independence imply that $P(\cdot|\overline{R} = t_i)$ treats the q_i as independent, identically-distributed Bernoulli variables with success probability t_i. Letting $\overline{q}$ be the proportion of q_i that are true, that means that conditional on $\overline{R} = t_i$, $\overline{q}$ is distributed according to a binomial distribution with parameters t_i and n; in particular: $P(\overline{q} = sn|\overline{R} = t_i) = \binom{n}{sn} t_i^{sn} (1 - t_i)^{n - sn}$.

Now suppose you learn that proportion $s \cdot n$ of the q_i were true. By Bayes formula, your posterior confidence in any $\overline{R} = t_i$ hypothesis should be:

$$
\begin{aligned}
P(\overline{R} = t_i|\overline{q} = sn) &= \frac{P(\overline{R} = t_i) \cdot P(\overline{q} = sn|\overline{R} = t_i)}{\sum_{j=1}^{m} P(\overline{R} = t_j) \cdot P(\overline{q} = sn|\overline{R} = t_j)} \\
&= \frac{P(\overline{R} = t_i) \cdot \binom{n}{sn} t_i^{sn} (1 - t_i)^{n - sn}}{\sum_{j=1}^{m} P(\overline{R} = t_j) \cdot \binom{n}{sn} t_j^{sn} (1 - t_j)^{n - sn}}
\end{aligned}
$$

References

Anderson, John R, 1990. *The Adaptive Character of Thought*. Erlbaum Associates.

Ariely, Dan, 2008. *Predictably irrational*. Harper Audio.

Belot, Gordon, 2013a. 'Bayesian Orgulity'. *Philosophy of Science*, 80(4):483–503.

———, 2013b. 'Failure of calibration is typical'. *Statistics and Probability Letters*, 83(10):2316–2318.

Bol, Linda and Hacker, Douglas J., 2012. 'Calibration research: Where do we go from here?' *Frontiers in Psychology*, 3(JUL):1–6.

Brenner, L. A., Koehler, D.J., Liberman, V., and Tversky, A., 1996. 'Overconfidence in Probability and Frequency Judgments: A Critical Examination'. *Organizational Behavior and Human Decision Processes*, 65(3):212–219.

Brenner, Lyle, Griffin, Dale, and Koehler, Derek J, 2005. 'Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment'. *Organizational Behavior and Human Decision Processes*, 97(1):64–81.

Briggs, R., 2009a. 'Distorted Reflection'. *Philosophical Review*, 118(1):59–85.

Briggs, Ray, 2009b. 'The Anatomy of the Big Bad Bug'. *Nous*, 43(3):428–449.

Budescu, David V, Wallsten, Thomas S, and Au, Wing Tung, 1997. 'On the importance of random error in the study of probability judgment. Part II: Applying the stochastic judgment model to detect systematic trends'. *Journal of Behavioral Decision Making*, 10(3):173–188.

Carr, Jennifer, 2019. 'Imprecise Evidence without Imprecise Credences'. *Philosophical Studies*, To appear.

Chater, Nick and Oaksford, Mike, 1999. 'Ten years of the rational analysis of cognition'. *Trends in Cognitive Sciences*, 3(2):57–65.

Christensen, David, 2010a. 'Higher-Order Evidence'. *Philosophy and Phenomenological Research*, 81(1):185–215.

———, 2010b. 'Rational Reflection'. *Philosophical Perspectives*, 24:121–140.

———, 2016. 'Disagreement, Drugs, etc.: From Accuracy to Akrasia'. *Episteme*, Forthcomin.

Cohen, L. Jonathan, 1981. 'Can human irrationality be experimentally demonstrated?' *Behavioral and Brain Sciences*, 4(3):317–331.

Crupi, Vincenzo, Fitelson, Branden, and Tentori, Katya, 2008. 'Probability, confirmation, and the conjunction fallacy'. *Thinking & Reasoning*, 14(2):182–199.

Cushman, Fiery, 2019. 'Rationalization is Rational'. *Behavioral and Brain Sciences*, To appear:1–27.

Dawid, A P, 1982. 'The Well-Calibrated Bayesian'. *Journal of the American Statistical Association*, 77(379):605–610.

Dawid, A. P., 1983. 'Calibration-Based Empirical Inquiry'. *The Annals of Statistics*, 13(4):1251–1273.

Dorst, Kevin, 2019a. 'Evidence: A Guide for the Uncertain'. *Philosophy and Phenomenological Research*, To appear.

———, 2019b. 'Higher-Order Uncertainty'. In Mattias Skipper Rasmussen and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, To appear. Oxford University Press.

Dunn, Jeff, 2015. 'Reliability for degrees of belief'. *Philosophical Studies*, 172(7):1929–1952.

Dunning, David, 2012. *Self-insight: Roadblocks and detours on the path to knowing thyself*. Psychology Press.

Dunning, David, Griffin, Dale W., Milojkovic, James D, and Ross, Lee, 1990. 'The Overconfidence Effect in Social Prediction'. *Journal of Personality and Social Psychology*, 58(4):568–581.

Edwards, Ward, 1982. 'Conservatism in Human Information Processing'. *Judgment under Uncertainty: Heuristics and Biases*, 359–369.

Ehrlinger, Joyce, Mitchum, Ainsley L., and Dweck, Carol S., 2016. 'Understanding overconfidence: Theories of intelligence, preferential attention, and distorted self-assessment'. *Journal of Experimental Social Psychology*, 63:94–100.

Elga, Adam, 2013. 'The puzzle of the unmarked clock and the new rational reflection principle'. *Philosophical Studies*, 164(1):127–139.

———, 2016. 'Bayesian Humility'. *Philosophy of Science*, 83(3):305–323.

Erev, Ido, Wallsten, Thomas S, and Budescu, David V, 1994. 'Simultaneous over-and underconfidence: The role of error in judgment processes.' *Psychological review*, 101(3):519.

Fine, Cordelia, 2005. *A Mind of its Own: How Your Brain Distorts and Deceives*. W. W. Norton & Company.

Fitelson, Branden and Hawthorne, James, 2010. 'The Wason Task(s) and the Paradox of Confirmation'. *Philosophical Perspectives*, 24:207–241.

Gelman, Andrew and Nolan, Deborah, 2002. 'You can load a die, but you can't bias a coin'. *American Statistician*, 56(4):308–311.

Gigerenzer, Gerd, 1991. 'How to make cognitive illusions disappear: Beyond heuristics and biases'. *European review of social psychology*, 2(1):83–115.

Gigerenzer, Gerd, Hoffrage, Ulrich, and Kleinbölting, Heinz, 1991. 'Probabilistic mental models: a Brunswikian theory of confidence.' *Psychological review*, 98(4):506.

Glaser, Markus and Weber, Martin, 2007. 'Overconfidence and trading volume'. *The Geneva Risk and Insurance Review*, 32(1):1–36.

———, 2010. 'Overconfidence'. *Behavioral finance: Investors, corporations, and markets*, 241–258.

Greco, Daniel and Hedden, Brian, 2016. 'Uniqueness and metaepistemology'. *The Journal of Philosophy*, 113(8):365–395.

Griffin, Dale and Tversky, Amos, 1992. 'The Weighing of Evidence and the Determinants of Confidence'. *Cognitive Psychology*, 24:411–435.

Griffiths, Thomas L., Chater, Nick, Norris, Dennis, and Pouget, Alexandre, 2012. 'How the bayesians got their beliefs (and what those beliefs actually are): Comment on bowers and davis (2012)'. *Psychological Bulletin*, 138(3):415–422.

Hahn, Ulrike and Oaksford, Mike, 2007. 'The rationality of informal argumentation: a Bayesian approach to reasoning fallacies.' *Psychological review*, 114(3):704.

Hall, Ned, 1994. 'Correcting the Guide to Objective Chance'. *Mind*, 103(412):505–517.

Harris, Adam J L and Hahn, Ulrike, 2011. 'Unrealistic optimism about future life events: A cautionary note.' *Psychological review*, 118(1):135.

Hedden, Brian, 2018. 'Hindsight Bias is not a Bias'. *Analysis*, To appear.

Hoffrage, Ulrich, 2004. 'Overconfidence'. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, 235–254.

Horowitz, Sophie, 2014a. 'Epistemic Akrasia'. *Noûs*, 48(4):718–744.

———, 2014b. 'Immoderately rational'. *Philosophical Studies*, 167:41–56.

———, 2019. 'The Truth Problem for Permissivism'. *The Journal of Philosophy*, 116(5):237–262.

Howard, Michael, 1984. *The Causes of Wars and Other Essays*. Harvard University Press.

Icard, Thomas, 2017. 'Bayes, Bounds, and Rational Analysis'. *Philosophy of Science*, 694837.

Isaacs, Yoaav, 2019. 'The Fallacy of Calibrationism'. *Philosophy and Phenomenological Research*, To appear.

Johnson, Dominic D P, 2009. *Overconfidence and war*. Harvard University Press.

Johnson, Dominic D.P. and Fowler, James H., 2011. 'The evolution of overconfidence'. *Nature*, 477(7364):317–320.

Joyce, James M, 1998. 'A Nonpragmatic Vindication of Probabilism'. *Philosophy of Science*, 65(4):575–603.

Juslin, Peter, 1994. 'The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items'. *Organizational Behavior and Human Decision Processes*, 57(2):226–246.

Juslin, Peter, Winman, Anders, and Olsson, Henrik, 2000. 'Naive empiricism and dogmatism in confidence research: A critical examination of the hardeasy effect.' *Psychological review*, 107(2):384.

Kahneman, Daniel, 2011. *Thinking Fast and Slow*. Farrar, Straus, and Giroux.

Kahneman, Daniel, Slovic, Paul, and Tversky, Amos, eds., 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kahneman, Daniel and Tversky, Amos, 1973. 'On the psychology of prediction.' *Psychological review*, 80(4):237.

———, 1996. 'On the reality of cognitive illusions.'

Kelly, Thomas, 2004. 'Sunk costs, rationality, and acting for the sake of the past'. *Nous*, 38(1):60–85.

———, 2008. 'Disagreement, Dogmatism, and Belief Polarization'. *The Journal of Philosophy*, 105(10):611–633.

Keren, Gideon, 1987. 'Facing uncertainty in the game of bridge: A calibration study'. *Organizational Behavior and Human Decision Processes*, 39(1):98–114.

Koehler, Derek J, Brenner, Lyle, and Griffin, Dale, 2002. 'The calibration of expert judgment: Heuristics and biases beyond the laboratory'. *Heuristics and biases: The psychology of intuitive judgment*, 686–715.

Koralus, Philipp and Mascarenhas, Salvador, 2013. 'The erotetic theory of reasoning: bridges between formal semantics and the psychology of deductive inference'. *Philosophical Perspectives*, 27:312–365.

Koriat, Asher, Lichtenstein, Sarah, and Fischhoff, Baruch, 1980. 'Journal of Experimental Psychology : Human Learning and Memory Reasons for Confidence'. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):107–118.

Krueger, Joachim I., 2012. 'The (Ir)rationality Proect in Social Psychology: A Review and Assessment'. In Joachim I. Krueger, ed., *Social Judgment and Decision Making*. Psychology Press.

Kruger, Justin and Dunning, David, 1999. 'Unskilled and Unaware of it: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments'. *Journal of Personality and Social Psychology*, 77(6):121–1134.

Kunda, Ziva, 1990. 'The case for motivated reasoning'. *Psychological Bulletin*, 108(3):480–498.

Lasonen-Aarnio, Maria, 2013. 'Disagreement and evidential attenuation'. *Nous*, 47(4):767–794.

———, 2015. 'New Rational Reflection and Internalism about Rationality'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 5, 145–171. Oxford University Press.

———, 2019. 'Higher-Order Defeat and Evincibility'. *Higher-Order Evidence: New Essays*, 144.

Lazer, David, Baum, Matthew, Benkler, Jochai, Berinsky, Adam, Greenhill, Kelly, Metzger, Miriam, Nyhan, Brendan, Pennycook, G., Rothschild, David, Sunstein, Cass, Thorson, Emily, Watts, Duncan, and Zittrain, Jonathan, 2018. 'The science of fake news'. *Science*, 359(6380):1094–1096.

Lewis, David, 1980. 'A subjectivist's guide to objective chance'. In Richard C Jeffrey, ed., *Studies in Inductive Logic and Probability*, volume 2. University of California Press.

———, 1994. 'Humean Supervenience Debugged'. *Mind*, 103(412):473–490.

Lichtenstein, Sarah, Fischhoff, Baruch, and Phillips, Lawrence D., 1982. 'Calibration of probabilities: The state of the art to 1980'. In Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under Uncertainty*, 306–334. Cambridge University Press.

Ludwig, Sandra and Nafziger, Julia, 2011. 'Beliefs about overconfidence'. *Theory and decision*, 70(4):475–500.

Magnus, Jan R. and Peresetsky, Anatoly A., 2018. 'Grade expectations: Rationality and overconfidence'. *Frontiers in Psychology*, 8(JAN):1–10.

Mahtani, Anna, 2017. 'Deference, respect and intensionality'. *Philosophical Studies*, 174(1):163–183.

Mandelbaum, Eric, 2018. 'Troubles with Bayesianism: An introduction to the psychological immune system'. *Mind & Language*, 1–17.

Mayseless, Ofra and Kruglanski, Arie W, 1987. 'What makes you so sure? Effects of epistemic motivations on judgmental confidence'. *Organizational Behavior and Human Decision Processes*, 39(2):162–183.

Merkle, Christoph and Weber, Martin, 2011. 'True overconfidence: The inability of rational information processing to account for apparent overconfidence'. *Organizational Behavior and Human Decision Processes*, 116(2):262–271.

Moore, Don A and Healy, Paul J, 2008. 'The trouble with overconfidence.' *Psychological review*, 115(2):502.

Myers, David G., 2010. *Psychology*. Worth Publishers, ninth edit edition.

Nebel, Jacob M., 2015. 'Status quo bias, rationality, and conservatism about value'. *Ethics*, 125(2):449–476.

Oaksford, Mike and Chater, Nick, 1994. 'A Rational Analysis of the Selection Task as Optimal Data Selection'. *Psychological Review*, 101(4):608–631.

———, 1998. *Rational models of cognition*. Oxford University Press Oxford.

———, 2007. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Odean, Terrance, 1999. 'Do Investors Trade Too Much?' *American Economic Review*, 89(5):1279–1298.

Ortoleva, Pietro and Snowberg, Erik, 2015. 'Overconfidence in political behavior'. *American Economic Review*, 105(2):504–535.

Peterson, CAMERON R. and Beach, LEE R., 1967. 'Man As an Intuitive Statistician'. *Psychological Bulletin*, 68(1):29–46.

Pettigrew, Richard, 2016. *Accuracy and the Laws of Credence*. Oxford University Press.

Pfeifer, Phillip E, 1994. 'Are we overconfident in the belief that probability forecasters are overconfident?' *Organizational Behavior and Human Decision Processes*, 58(2):203–213.

Pinker, S, 2009. 'How the Mind Works'.

Plous, Scott, 1993. *The psychology of judgment and decision making*. Mcgraw-Hill Book Company.

Roush, Sherrilyn, 2009. 'Second Guessing: A Self-Help Manual'. *Episteme*, 251–268.

———, 2016. 'Knowledge of Our Own Beliefs'. *Philosophy and Phenomenological Research*, 93(3).

———, 2017. 'Epistemic Self-Doubt'.

Salow, Bernhard, 2018. 'The Externalist's Guide to Fishing for Compliments'. *Mind*, 127(507):691–728.

Schoenfield, Miriam, 2012. 'Chilling out on epistemic rationality'. *Philosophical Studies*, 158(2).

———, 2014. 'Permission to Believe: Why Permissivism is True and What it Tells Us About Irrelevant Influences on Belief'. *Noûs*, 48(2):193–218.

———, 2015. 'A Dilemma for Calibrationism'. *Philosophy and Phenomenological Research*, 91(2):425–455.

———, 2016. 'An Accuracy Based Approach to Higher Order Evidence'. *Philosophy and Phenomenological Research*, To Appear.

Schultheis, Ginger, 2018. 'Living on the Edge: Against Epistemic Permissivism'. *Mind*, 127(507):863–879.

Seidenfeld, Teddy, 1985. 'Calibration , Coherence , and Scoring Rules'. *Philosophy of Science*, 52:274–294.

Sliwa, Paulina and Horowitz, Sophi, 2015. 'Respecting *all* the evidence'. *Philosophical Studies*, 172(11):2835–2858.

Sniezek, Janet A, Paese, Paul W, and Switzer III, Fred S, 1990. 'The effect of choosing on confidence in choice'. *Organizational Behavior and Human Decision Processes*, 46(2):264–282.

Stich, Stephen P., 1985. 'Could Man be an Irrational Animal?' *Synthese*, 64:115–135.

Sunstein, C, 2009. *Going to Extremes: How Like Minds Unite and Divide*. Oxford University Press.

Sunstein, Cass R, 2000. 'Deliberative trouble? Why groups go to extremes'. *The Yale Law Journal*, 110(1).

Taylor, Shelley E and Brown, Jonathon D, 1988. 'Illusion and well-being: a social psychological perspective on mental health.' *Psychological bulletin*, 103(2):193.

Tenenbaum, Joshua B and Griffiths, Thomas L, 2006. 'Optimal Predictions in Everyday Cognition'. *Psychological Science*, 17(9):767–773.

Tetlock, Philip E and Gardner, Dan, 2016. *Superforecasting: The art and science of prediction*. Random House.

Thaler, Richard H and Sunstein, Cass R, 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Tversky, Amos and Kahneman, Daniel, 1974. 'Judgment under uncertainty: Heuristics and biases'. *Science*, 185(4157):1124–1131.

Vallone, Robert P., Griffin, Dale W., Lin, Sabrina, and Ross, Lee, 1990. 'Overconfident Prediction of Future Actions and Outcomes by Self and Others'. *Journal of Personality and Social Psychology*, 58(4):582–592.

van Fraassen, Bas, 1983. 'Calibration: A Frequency Justification for Personal Probability'. In R.S. Cohen and L Laudan, eds., *Physics, Philosophy, and Psychoanalysis: Essays in Honor of Adolf Gr unbaum*, 295–318. D. Reidel Publishing Company.

———, 1984. 'Belief and the Will'. *The Journal of Philosophy*, 81(5):235–256.

van Prooijen, Jan-Willem and Krouwel, André P M, 2019. 'Psychological Features of Extreme Political Ideologies'. *Current Directions in Psychological Science*, 28(2):159–163.

von Winterfeldt, Detloff and Edwards, Ward, 1986. *Decision analysis and behavioral research*. Cambridge University Press.

White, Roger, 2005. 'Epistemic Permissiveness'. *Philosophical Perspectives*, 19(1):445–459.

———, 2009a. 'Evidential Symmetry and mushy credence'. *Oxford Studies in Epistemology*, 161–186.

———, 2009b. 'On Treating Oneself and Others as Thermometers'. *Episteme*, 6(3):233–250.

Williamson, Timothy, 2000. *Knowledge and its Limits*. Oxford University Press.

———, 2018. 'Evidence of Evidence in Epistemic Logic'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, volume To appear. Oxford University Press.