Being Rational and Being Wrong

Kevin Dorst
kevindorst@pitt.edu
Word count: 11,028

September 2021

Abstract

Do people tend to be overconfident in their opinions? Many think so. They've run studies to test whether people are *calibrated*: whether their confidence in their opinions matches the proportion of those opinions that are true. Under certain conditions, people are systematically "over-calibrated"—for example, of the opinions they're 80% confident in, only 60% are true. From this observed over-calibration, it's inferred that people are irrationally *overconfident*. My question: When and why is this inference warranted? Answering this question requires articulating a general connection between being rational and being right—something extant studies have not done. I show how to do so using the notion of *deference*. This provides a theoretical foundation to calibration research, but also reveals a flaw: the connection between being rational and being right is much weaker than is commonly assumed; as a result, rational people can often be expected to be miscalibrated. Thus we can't test whether people are overconfident by simply testing whether they are over-calibrated; instead, we must first predict the expected *rational deviations* from calibration, and then compare those predictions to people's performance. I show how in principle this can be done—and that doing so has the potential to overturn the standard interpretation of robust empirical effects.

1 The Question

Pencils ready! For each pair, circle the city that you think has a larger population (in the city proper), and then rate how confident you are in that guess on a $50-100\%$ scale:

1) Denver or Phoenix? Confidence: ____%
2) San Jose or Seattle? Confidence: ____%
3) Indianapolis or Columbus? Confidence: ____%

If you're like most people, this test shows two things. 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 200 test-takers, the average confidence people had in their answers was 75%, while the proportion of correct answers to questions like this was only 45% (see §5).

That rather striking result—the so-called "overconfidence effect"—is common: on a variety of tests, people's average confidence in their answers exceeds the proportion that are correct.[1] Many have inferred from this that people are often overconfident in their opinions: more confident than it is rational for them to be, given their evidence.[2] Many have used these (and related) results to paint unflattering pictures of the human mind as prone to pervasive irrationality and bias.[3] And many others have invoked overconfidence in particular to explain a variety of societal ills—from market crashes, to political polarization, to wars.[4] Daniel Kahneman summed it up bluntly: 'What would I eliminate if I had a magic wand? Overconfidence'.[5]

Fair enough. But how—exactly—did we conclude that people are overconfident? One of the most common types of evidence—the type I'll focus on, till §6—is binary-question ("2-alternative-forced-choice") calibration studies like the one you just took. Ask people a variety of such questions, have them guess the answer and report their confidence in those guesses, and then graph that confidence against the proportion of answers that are true. Say that a person is *calibrated* (at x) if $x\%$ of the claims that they are $x\%$ confident in are true. They are *over*-calibrated (at x), in the sense that their confidence needs to be lower in order to be calibrated, if fewer than $x\%$ of such claims are true. And they are *under*-calibrated (at x) if more than $x\%$ of such claims are true. Schematic graphs of these different **calibration curves** are given on the left of Figure 1. Meanwhile, the right side of Figure 1 plots the results of my study (§5), replicating the result that (on certain questions) people tend to be over-calibrated.

That's the evidence: people are often over-calibrated on binary-question tests. How does it support the conclusion—namely, that people are often *overconfident*? Well, it's natural to think that if people's confidence is rationally placed, their opinions will be right about as often as they expect them to be. Conversely, if they're confident in their opinions and yet many (or most!) of those opinions are wrong, this seems to suggest that they're *too* confident; *over*confident.

Though natural, this is a substantive inference: it moves from an empirical observation ('you are miscalibrated') to a normative conclusion ('you are irrational'). Call it the **rational-to-right inference** since it presupposes that whether your opinions are right as often as you expect can be used to figure out whether they're rational.

The Questions: What *is* the connection between being rational and being right? More specifically: When is the rational-to-right inference warranted? When is it not? And what does that tell us about how to interpret the results of calibration studies?

[1] Lichtenstein et al. 1982; Harvey 1997; Hoffrage 2004; Glaser and Weber 2010; Moore et al. 2015b.

[2] E.g. Lichtenstein et al. 1982; Dunning et al. 1990; Vallone et al. 1990; Griffin and Tversky 1992; Kahneman and Tversky 1996; Budescu et al. 1997; Brenner 2000; Koehler et al. 2002; Brenner et al. 2005; Glaser and Weber 2010; Merkle and Weber 2011; Brenner et al. 2012; Moore et al. 2015b; Ehrlinger et al. 2016; Magnus and Peresetsky 2018.

[3] E.g. Plous 1993; Fine 2005; Ariely 2008; Hastie and Dawes 2009; Myers 2010; Kahneman 2011b; Thaler 2015; Lewis 2016; Tetlock and Gardner 2016.

[4] E.g. Howard 1984; Odean 1999; Glaser and Weber 2007; Johnson 2009; Johnson and Fowler 2011; Kahneman 2011a; Ortoleva and Snowberg 2015; van Prooijen and Krouwel 2019.

[5] Shariatmadari (2015). Some authors argue that the "overconfidence effect" is compatible with rationality (e.g. Gigerenzer 1991; Hoffrage 2004; Angner 2006; Moore and Healy 2008)—see §2. But the majority of the literature thinks it is evidence of irrationality.

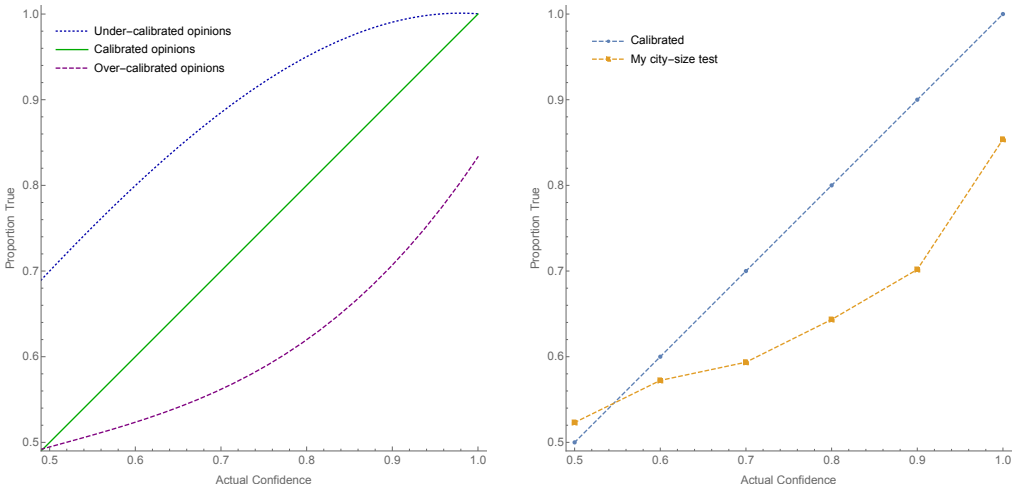

Figure 1: Left: Schematic calibration curves. **Right:** "Overconfidence effect" (i.e. over-calibration) in my study. (See §5 for study details.)

The Plan: I'll first say what sort of connection the rational-to-right inference assumes, and explain why the existing literature has failed to articulate it (§2). I'll then use the notion of *deference* to show why we should expect such a connection in certain simple cases (§3). However, it turns out this connection will break in predictable ways—meaning that often *mis*calibration is evidence for *rationality* (§4). I argue that this result provides both a foundation for and a refinement to the standard calibration-study methodology: in testing whether people are rational, the null hypothesis should not be that they'll be calibrated; rather, we must first predict the expected rational *deviations* from calibration, and then compare people's performance to those predictions. I'll show how in principle this can be done, and that doing so has the potential to change the interpretation of robust empirical effects (§§5–6). Although I'll make this argument specifically for binary-question calibration tests, §6 will suggest that many of the considerations carry over to other formats, such as placement (Kruger and Dunning 1999) and interval-estimation (Moore et al. 2015b) tests. In short: although calibration scores are a great guide to how *wrong* your past judgments were—and therefore, how to do better in the future (Tetlock and Gardner 2016)—they are a poor guide to how *rational* they were. *Mantra:* Use calibration for deliberation, not for evaluation.

The Upshot: If this is correct, it shows that certain philosophical and psychological literatures are closer than has been realized. Philosophical debates about deference principles can inform the methodology of calibration studies; meanwhile, the methods of such studies suggest that, although there is (arguably) no *necessary* connection between being rational and being right,[6] there often is an *evidential* connection between the two. Thus this paper supports the growing interest in connecting philosophical accounts of rationality with psychological investigations of it.[7]

[6] Joyce (1998); Littlejohn (2012, 2018); Gibbons (2013); Schoenfield (2016); Horowitz (2014b, 2019b); Wedgwood (2017); Lord (2018); Rinard (2019); Comesaña (2020); Staffel (2020).

[7] E.g. Kelly 2004, 2008; Crupi et al. 2008, 2009; Fitelson and Hawthorne 2010; Koralus and Mascar-

2 The Problem

There's a problem here. The rational-to-right 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 observed are (1) and (2). The rational-to-right inference draws a conclusion 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[8])? I'll take it to mean that they're (on average) *more extreme* in their opinions 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 2.) That's the overconfidence hypothesis. What is the evidence offered in its favor? It's that in a variety of settings, people are over-*calibrated*: if we plot actual degree of confidence against *proportion true*, the curve is tilted (right side of Figure 2).

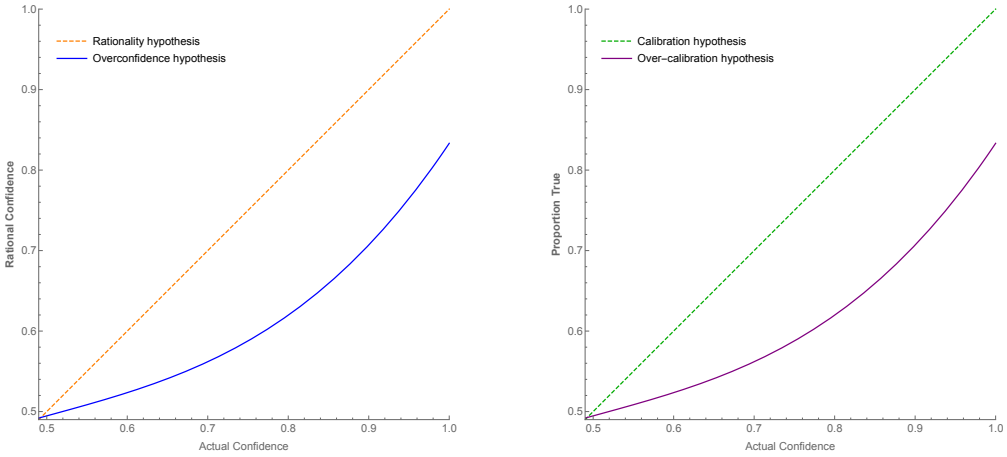

Figure 2: Left: Rationality vs. Overconfidence hypotheses. **Right:** Calibration vs. Over-calibration hypothesis.

The graphs look the same, but the axes are different. It follows that you're warranted in drawing the rational-to-right inference only when you should expect the two axes to align—i.e. when you should expect a rational person to be calibrated on the given test.

enhas 2013; Nebel 2015; Icard 2017; Mandelbaum 2018; O'Connor and Weatherall 2018; Hedden 2019; Singer et al. 2019; Doody 2020; Quilty-Dunn 2020; Dorst and Mandelkern 2021; Karlan 2021; Kinney and Bright 2021; Thorstad 2021.

[8]The "in a given domain" rider is important, as patterns of miscalibration vary widely—see §4.1.

Moreover, their average confidence presumably equals their estimate of the proportion of their answers that are right. (It does for a Bayesian.[9]) If so, the rational-to-right inference is warranted only when you should expect that, if (certain of) the person's estimates are rational, those estimates are right.

What do I mean by "rational"? Well, what do *you* mean? Calibration studies often overlook the fact that they presuppose a substantive notion of rationality. After all, to say that someone is overconfident in a set of opinions $q_1, ..., q_n$ is to say that they are, on average, more confident than they *should* be: that there is some number c that represents their average confidence, some other number r that represents the average confidence it would be rational for them to have, and that $c > r$. Thus calibration studies presuppose that there are rational degrees of confidence (r_i) that people ought to have in the claims they evaluate (q_i), which may differ from their reported degrees of confidence (c_i). Nevertheless, I know of no study that represents the rational degrees of confidence r_i as variables to be investigated. None of the studies cited in this paper do so.[10] As a result, none state the assumptions needed to derive the result that we should expect the rational opinions r_i to be calibrated in their study.[11] In other words: I know of no study that states what assumptions it is making such that we should expect the two y-axes in Figure 2 to align. Yet observing that people's judgments are miscalibrated provides no evidence that they are irrational unless we have reason to expect these axes to align. That is the problem.[12]

Is it easily answered? No. Although any Bayesian will expect any particular set of *their own* opinions to be calibrated,[13] we are not them—so there's no theorem that *we* should expect their opinions to be calibrated. Often we should not. As is often pointed out by the philosophical literature (see footnote 6), there is no necessary connection between being rational and being right at any level of statistical generality. *Case 1:* Rajat uses all his information rationally. His information seems a lot like yours or mine. As a result he's sure that he has hands, confident he's healthy, and suspects he'll soon grab

[9]Let Q be any set of opinions each of which they're x%-confident in; they're calibrated on Q iff x% of those claims are true. Let C be their probability function, $\mathbb{E}[X]$ be their estimate of random variable X ($\mathbb{E}(X) := \sum_t C(X = t) \cdot t$) and $\mathbb{1}_q$ be the indicator for q (1 if true, 0 if false). Then their estimate for the proportion of claims in Q they got right is $\mathbb{E}[\frac{1}{|Q|}\sum_{q\in Q}\mathbb{1}_q] = \frac{1}{|Q|}\sum_{q\in Q}\mathbb{E}[\mathbb{1}_q] = \frac{1}{|Q|}\sum_{q\in Q} x = x$.

[10]Including those cited in footnotes 2, 3, and 5. Some studies invoke objective probabilities, *true* (vs. reported) confidence, or the confidence of differing agents (Gigerenzer et al. 1991; Erev et al. 1994; Juslin et al. 1997, 1999, 2000; Moore and Healy 2008). None of these can stand in for rational confidence.

[11]Some derive this result for a given Bayesian agent (Brenner et al. 2005; Moore and Healy 2008; Merkle and Weber 2011; Benoît and Dubra 2011)—but they all implicitly assume that the Bayesian's prior beliefs match the frequencies on the test. As we'll see, this can't in general be assumed.

[12]Might this indicate that these studies aren't interested in rationality? No. They are peppered with normative assessments of people's confidence as '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" (Koehler et al. 2002, 686; Glaser and Weber 2010, 249; Moore et al. 2015b, 182), and so on. Kahneman and Tversky (1996) 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" (589).

[13]Their estimate of the proportion of truths will equal their average degree of confidence in them since $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(q_i)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{1}(q_i)] = \frac{1}{n}\sum_{i=1}^{n}C(q_i)$. So their estimate of the proportion of truths amongst the claims they are (say) 80% confident in must be 80%. If they treat the claims independently, they will (by the law of large numbers) be confident that roughly 80% of them are true.

lunch. But though rational, Rajat is wrong on all these fronts (and many others)—for, unbeknownst to him, he's a brain-in-a-vat being deceived by a mad scientist. If we ran a calibration study on Rajat, he would be over-calibrated—many of his confident opinions would be false. Yet this doesn't suggest he's irrational.

Mundane cases make the same point. *Case 2:* Georgie is quite confident—and quite wrong—in most of her geographical opinions. Does this provide evidence that she's overconfident? Not if we know that, contrary to her evidence, her geography teacher gave her an outdated textbook. She's misinformed, not irrational. Obviously we can imagine scenarios in which the entire population of test-takers are the same position: we'd expect *every* student in Georgie's geography class to be similarly over-calibrated.

Likewise, it's easy to construct cases in which we know someone has high-quality evidence, and yet the rational-to-right inference fails. *Case 3:* 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 on that toss? Write that number down—I'll look at it in a second. First to toss the coin (...done). Turns out it landed heads only 30 times. Now to compare that to your confidence. You were 60% confident that each toss would land heads, but only 30% of those claims were true. Is this evidence that you're overconfident? Obviously not; it's just evidence that you're unlucky.

Similarly, sometimes we can *know beforehand* that your rational opinions will be systematically wrong. *Case 4:* I have an urn of mis-printed coins—60 of them are double-headed, and 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 I draw will land heads on that toss? 60%. Yet you know that either I'll draw a double-headed or a double-tailed coin. If the former, all the tosses will land heads—100% of the things that you're 60% confident in will be true. And if the latter, then none of them will land heads—0% of the things that you're 60% confident in will be true. So we know that, either way, the rational opinions will be badly miscalibrated.

Finally: we can almost always expect that *certain classes* of the rational opinions will be miscalibrated. *Case 5:* Suppose you're about to take a test drawn randomly from a representative set of your generally-accurate, rationally-processed evidence. Though rational, your opinions aren't perfect—sometimes you'll be wrong. Consider the set of guesses $\mathcal{W}$ you'll be wrong about, and the set $\mathcal{R}$ you'll be right about. You won't know what they are until the answers are revealed. But you *know* you'll be miscalibrated on them—0% of the claims in $\mathcal{W}$ will be true, but your average confidence in them will be higher than that; and 100% of the claims in $\mathcal{R}$ will be true, but you average confidence in them will be lower than that. More generally, we should *always* expect that people will be over-calibrated on sets like "the answers they tended to get wrong" and under-calibrated on sets like "the answers they tended to get right" (§5).

Upshot: it's easy to imagine scenarios in which rational people are systematically miscalibrated. When we run the rational-to-right inference, we must somehow be discounting concerns from such scenarios. My question is what justifies us in doing so.

To be clear: I'm not claiming that these toy examples shed doubt on the rational-to-right inference in practice (nor that they will be of any surprise to the researchers

conducting calibration studies!). What I'm claiming is that these cases make salient a conceptual question. As we've seen, whether we should expect the rational opinions to be calibrated on a given set of questions depends completely on the evidence that the test-taker has and how that set was determined. Surely, in some sense, we should expect that the opinions their evidence warrants will tend to be right—that's the point of evidence, after all—and thus that rational degrees of confidence will *tend* to be calibrated. The question is: *When, why, and in what sense should we expect this?*

My goal is to answer this question. In §3 I'll articulate a general, probabilistic connection between being rational and being right that explains why the rational-to-right inference works in certain simple cases. But it also reveals (§4) that it'll fail in systematic ways. §5 will use this fact to refine the methodology of calibration studies.

But before moving on, I should say how this project relates to a variety of theoretical points made in the calibration literature.[14] *Ecological* approaches argues that subjects may have misleading information, so we must try to control for this by choosing representative questions from a natural domain (Gigerenzer 1991; Gigerenzer et al. 1991; Juslin 1994; Juslin et al. 2000; Hoffrage 2004). *Error-model* approaches argue that regardless of how questions are selected, there will be stochastic errors ("noise") in both the selection of items and in the subject's reporting of their confidence that can lead to them being miscalibrated on a given test even if their opinions are calibrated overall (Erev et al. 1994; Pfeifer 1994; Juslin et al. 1997, 1999, 2000). Similar points have been made using information asymmetries between subjects (Moore and Healy 2008; Jansen et al. 2018). Given this, these researchers have built models of how people may arrive at their confidence in an apparently-rational way, and yet nonetheless we might expect to see the sorts of miscalibration that we in fact observe (cf. Benoît and Dubra 2011).

I agree with all of this—but the point I'm making is a broader one. These researchers have proposed particular, rational-seeming mechanisms for forming opinions,[15]and shown that they can lead to miscalibration. What I'm going to show is that *no matter the mechanism* a person uses to form degrees of confidence—whether it is Bayesian, or not—rational opinions should be expected to be miscalibrated in systematic ways. Demonstrating this becomes possible once we explicitly represent the rational opinions as variables to be investigated. Interestingly, these rational deviations from calibration turn out to be broadly consistent with some of the main empirical trends (§5). But more importantly, they show that *we've been using the wrong yardstick*. In assessing whether people are overconfident, we should never simply compare their calibration curves to

[14]What about precedents in the *philosophical* literature? While many have addressed the general connection between being rational and being right (see fn. 6), to my knowledge no philosophers have addressed the rational-to-right inference as formulated here, in terms of the interpretation of calibration studies. Instead, they've asked questions like: whether calibration can objectively vindicate a set of opinions (van Fraassen 1983; Dawid 1983; Seidenfeld 1985; Joyce 1998; Dunn 2015; Pettigrew 2016); or whether a Bayesian agent's beliefs about their own calibration are problematic (Dawid 1982; Belot 2013a,b; Elga 2016); or how your expectations about your calibration should affect your confidence in your answers (Roush 2009, 2016, 2017; White 2009b; Christensen 2010a, 2016; Lam 2011, 2013; Sliwa and Horowitz 2015; Schoenfield 2015, 2018; Isaacs 2019).

[15]Which, in turn, have been criticized on a variety of grounds (Kahneman and Tversky 1996; Budescu et al. 1997; Brenner 2000; Koehler et al. 2002; Brenner et al. 2005; Merkle and Weber 2011).

the diagonal calibrated line—rather, we must compare them to the predicted rational *deviations* from this calibrated line. I'll show how we might in principle predict these rational deviations without making any assumptions about mechanism.

3 The Insight

When are we warranted in performing the rational-to-right inference? That is: when should we expect the rational opinions for a given person to be calibrated?

Stard by making things simple. Suppose you get very limited evidence. A single subject—Calvin—was given a calibration test about geography; the questions were selected at random from a textbook. He knows everything you do about the setup of the test, and you have no reason to suspect that his evidence is misleading. Consider all the claims that Calvin was 80% confident in—call them his **80%-opinions**. All you're told is what proportion of them were true. I claim that in this simple scenario, the rational-to-right inference is warranted: if you learn that (roughly) 80% of Calvin's 80%-opinions were true, you get 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 evidence that he was overconfident (or underconfident). Thus even without a necessary connection between being rational and being right, there is often a robust *evidential* connection between the two. This is the insight behind calibration studies.

Why is it correct—and when can it fail? Here's the basic idea of this paper. Calvin has more information about his 80%-opinions than you do. Thus *absent any defeaters* you should defer to the opinions his evidence warrants: conditional on his evidence indeed warranting being confident that around 80% of them are true (conditional on him being rational), you should be confident that around 80% of them are true; meanwhile, conditional on his evidence warranting being confident that only around *70%* of them are true (conditional on him being overconfident), you should be confident than around 70% of them are true. Thus when you *learn* that only 70% of them are true (he's over-calibrated), you learn something that's more likely if he's overconfident than if he's rational—and so get evidence that he's overconfident. This is why the rational-to-right inference is warranted when you have no defeaters. But things change if you *do* have defeaters: if you know something Calvin does not, you should no longer defer to his rational opinions; learning that he's rational to expect that 80% of those opinions are true doesn't necessarily mean that *you* should expect that. As a result, the rational-to-right inference breaks down.

This section explains how this reasoning works when you don't have any defeaters; §4 explains why it fails when you do.

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. The coins had a variety of special markings on them—on some, Washington has a large nose and small ears; on others, he has a thin neck and bushy eyebrows; etc. In principle, if you knew how to decipher the markings, you could tell what the bias of the coin was just by looking.

Mary tossed the coins many, many times. She kept fastidious records: for each toss she drew a picture 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.

. . . 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 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 tell which coins are 80%-biased—i.e. which that are either 80% biased toward heads or toward tails? More confident, I take it. For it's rather surprising that so nearly 80% of the guesses she was 80% confident were correct, and if she can reliably decipher the coins, that would explain why this is so: she's 80%-confident in her guess only when the coin is 80% biased in that direction. 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 tell which coins are 80%-biased, and instead that she is likely over-estimating the strength of some coins' biases.

Call this inference—from "Bianca was (mis)calibrated in her 80%-opinions" to "she probably can(not) reliably decipher which coins are 80%-biased"—the **deciphered-to-right** inference, since it moves from her rates of being right to whether she has deciphered the markings. Clearly it's warranted in this simple scenario. If we can get clear on why it is—and what the analogy is between Bianca's case and Calvin's—it'll tell us what needs to hold for the rational-to-right inference to be warranted.

Why does the deciphered-to-right inference work in this scenario? 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 be (on average) around 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.

Thus the driving force of the deciphered-to-right inference is that hypotheses about whether she is deciphering the coins' biases, over-estimating them, or under-estimating them, each have direct implications for how many of the coins *you* should expect to land the way she guesses. This is because, first, you should **defer** to the average biases of the coins in forming your opinion about how a given coin will land: conditional on the coins having an average bias of $x\%$ toward Bianca's prediction (heads or tails), 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. (In our case, these two principles follow from the Principal Principle; see Lewis 1980, 1994; Hall 1994; Briggs 2009b.) Combined, these principles make 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 rational-to-right inference to work in Calvin's case, analogous deference and independence principles must hold.

What does the analogy amount to? For each tablet Bianca was shown, there was 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 his answer. We wanted to know whether Bianca could tell what the markings mean for the biases of the various coins—not necessarily in the sense of being certain of what they mean, but of being able to reliably line up her credences in how they'll land with the coins' biases. 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—not necessarily in the sense of being certain what his evidence supports, but of being reliably able to line up his credences with the degree of confidence it warrants. In Bianca's case, the deciphered-to-right inference went through because we should defer to the *biases* of the coins, and do so independently of how her other predictions turn out. Thus, in Calvin's case, the rational-to-right 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 guesses turn out to be true or false.

Now more precisely. As we're focusing on binary-question tests, I'll assume that when presented with a pair of possible answers, $\{p, p'\}$, he's certain that one of them is correct and forms his degrees of confidence in each; then he guesses the one that he thinks is more likely to be true (picking randomly if he's 50–50), and reports his confidence in that guess. Thus I'll assume that for binary questions his guess is determined by his degrees of confidence ("credences")—and that these degrees of confidence are the "opinions" that we're assessing the rationality of. (See §6.2 for discussion of guesses about questions that aren't binary.)

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.* (I assume we know that there are exactly n such opinions—the reasoning generalizes if we're unsure.) We can entertain different

hypotheses about the *rational* opinions for Calvin to have. Let $R(q_i)$ be the *R*ational confidence for Calvin to write down in q_i (I make no assumption about how it's determined). Now let $\overline{R} := \frac{1}{n}\sum_{i=1}^{n} R(q_i)$ be the *average* rational opinion for Calvin to have in the q_i, i.e. the average confidence he *should* have in the claims he's in fact 80%-confident in. 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, in which case it will be lower than 80%: $\overline{R} < 0.8$. (Or perhaps he's underconfident, in which case $\overline{R} > 0.8$.)

Let q_i be any of Calvin's 80%-opinions. How should learning about what's rational *for Calvin* affect your opinion in q_i? For the case to be analogous to Bianca's, you must again defer—not to Calvin's actual opinions, but to the opinions that his evidence makes *rational* for him to have. Let $\boldsymbol{P}$ be a probability function representing *your* rational degrees of confidence.[16] Then what we need is:

> **Deference:** Conditional on 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: $P(q_i|\overline{R} = x) = x$.

Deference is an interpersonal, rationalized, and "averaged" generalization of the well-known Reflection principle (van Fraassen 1984; Briggs 2009a; Christensen 2010b; Mahtani 2017).[17] It tells you to defer to the opinions it is *rational* for Calvin to have, not the opinions he in fact has. In our setup you don't know exactly what claims he was evaluating in forming his 80%-opinions—q_i is simply the claim that *the ith claim on this geography test that Calvin was 80%-confident in (whatever that is) is true*. Thus you have little 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. Since you have no reason to think Calvin's geographical evidence is misleading (i.e. absent any defeaters), it thus seems reasonable for you to be disposed to adopt the level of confidence you learn it's rational for him to have.

Whether Deference holds in a given case obviously depends massively on the details. My aim is not primarily that it *does* hold in a particular class of cases. Rather, my aim is to establish a biconditional: modulo a few complications, the rational-to-right inferences is warranted iff, given your evidence about the test, you should obey Deference—i.e. be disposed to defer to the opinions that are rational for Calvin. As is true of deference principles generally (Mahtani 2017), this will usually hold only if you don't have relevant information about the test that Calvin does not—a point we'll return to in §4.

In addition to the details of the case, whether Deference should generally be expected to hold depends on hard epistemological questions. Whether *interpersonal* deference

[16] Why are *you* in the picture at all? Of course, whether Calvin is rational has nothing to do with you; but whether *you have evidence* that he's rational does—and that is our question. For simplicity I'll assume that your rational opinions can be modeled with a precise probability function—but the reasoning will generalize. For discussion of the (de)merits of such models, see White (2005, 2009a); Schoenfield (2012, 2014); Horowitz (2014b); Schultheis (2018); Carr (2020).

[17] Appendix A.1 shows how this "averaged" version can be derived from a more familiar "point-wise" version *if* we assume—what I don't assume generally—that R is a probability function.

principles hold is highly dependent on the debate between uniqueness and permissivism (e.g. White 2005; Schoenfield 2014, 2019; Horowitz 2014b, 2019a; Greco and Hedden 2016; Schultheis 2018). Whether *rationalized* deference principles hold is highly dependent on debates around higher-order evidence (e.g. Williamson 2000, 2019; Christensen 2010b; Lasonen-Aarnio 2013, 2015, 2019; Elga 2013; Horowitz 2014a; Salow 2018; Dorst 2020a,b). Deference will be a theorem in our setup given uniqueness plus higher-order certainty; it'll be approximately true under some (but not all) weaker theories—see §6.1.

But wait. Isn't there a simpler problem? How are we supposed to use deference to get evidence about whether Calvin is rational? Doesn't applying it require *knowing* what the (average) rational opinions for Calvin are?

No. Deference is a claim about *conditional* opinions—it says that conditional on the claim that the average rational confidence for Calvin is $x\%$, adopt $x\%$ confidence yourself. Even if you never learn that (say) the average rational confidence for him is 70%, Deference tells you what to expect about q_i *if* that hypothesis is true—and thus it constrains how your confidence in that hypothesis should be affected by learning about the q_i, such as that about 70% of them have landed heads. Compare: here's a coin of unknown bias. The Principal Principle (Lewis 1980) is a claim about conditional opinions—it says that conditional on the claim that the bias of the coin is $x\%$, you should be $x\%$ confident that it'll land heads on a given toss. Even if you never learn that (say) the bias of the coin is 70%, the Principal Principle tells you what to expect about the coin flips *if* that hypothesis is true—and thus it constrains how your confidence in that hypothesis should be affected by learning about the coin flips, such as that around 70% of them have landed heads.

So Deference *can* be used to learn about whether Calvin is rational, i.e. to ground the rational-to-right inference. Indeed, the failure of Deference explains why the rational-to-right inference fails in many of our initial cases (§2). You shouldn't defer to Rajat (Case 1) or Georgie (Case 2), because you know something they don't—namely, that he's a brain in a vat, and she has an outdated textbook. Similarly, when I saw that my 60%-biased coin landed heads only 30 times (Case 3), I had evidence that you didn't when you formed your (rational) opinions, so I shouldn't defer to them. Likewise for Case 5—I shouldn't defer to your opinion about q_i if I know that it's in the set $\mathcal{W}$ of guesses you were wrong about, since you (of course) didn't know that.

But Deference doesn't explain why the rational-to-right inference fails in our case of the misprinted coins (Case 4). I haven't yet drawn the coin from the urn, so I defer to your rational opinions, yet I know you'll be miscalibrated. This is where we need our second assumption: 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 $q_{i_0}, ..., q_{i_k}$: $P(q_{i_0}|\overline{R} = x, q_{i_1}, ..., q_{i_l}, \neg q_{i_{l+1}}, ..., \neg q_{i_k}) \;=\; P(q_{i_0}|\overline{R} = x)$

This explains why the rational-to-right inference fails in the case of the misprinted

coins—we know that if one toss lands heads, they all will. How plausible is Independence in Calvin's case? There's much more to be said, but it's at least well-motivated as a first approximation since you know the test questions were random and unrelated.[18]

Deference and Independence imply that the rational-to-right inference is warranted in our simple scenario. 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." In particular, conditional on Calvin's 80%-opinions being on average rational, you should think it quite likely that roughly 80% of them will be true; and conditional on his 80%-opinions being on average quite overconfident (say, the average rational confidence is 60%), you should think it quite likely that less than 80% (roughy 60%) of them will be true. Since things are evidence for the hypotheses that make them likely, it follows that learning that 80% of his opinions are true is strong evidence that the (average) rational confidence was 80% (he was rational); meanwhile, learning that 60% of them are true is strong evidence that the (average) rational confidence was *less* than 80% (he was overconfident).

Toy example: suppose he has 50 different 80%-opinions, and you're initially equally confident that the average rational confidence ($\overline{R}$) for him to have in them is any of 60%, 61%,..., or 99%. Say he is *substantially overconfident* if the average rational confidence in his 80%-opinions is less than 75% ($\overline{R} < 0.75$). Then if you learn that 70% of those opinions were true, the rational-to-right inference is warranted: your confidence that he's substantially overconfident should jump from 37.5% to 78%. (See §A.2.)

Upshot: the rational-to-right inference can be put on a firm theoretical foundation. Whenever Deference and Independence hold, it's warranted.

But by the same token: when Deference *fails*, the exact same reasoning will show that the rational-to-right 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; 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 (and subject to minor exceptions), the rational-to-right inference is warranted in a given scenario iff Deference holds.

Upshot: in order to perform the rational-to-right inference on a given test, we must make the case that Deference holds for it.

How easy will it be to make this case? How robust is Deference to variations in our simple scenario? §4 argues that it is very fragile: that we should often expect Deference to systematically fail, and hence we should not expect rational people to be calibrated. §5 then argues that these rational deviations from calibration are in principle predictable—meaning that a more nuanced type of calibration study is possible.

[18]This is at best approximately true—for example, learning that *all* of Calvin's other 80%-opinions were false should make you suspicious. 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 given in §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 rational-to-right inference will be.

4 The Limits

The real world isn't like our simple scenario, for you know a whole lot more about the test: its content, how it was constructed, etc. If Calvin doesn't know them, each of these bits of information threatens to undermine Independence and Deference. Thus the first and simplest methodological point is that we must try to inform subjects as much as possible about the test.

But some information *can't* be given ahead of time. Here I'll focus on just one such piece of information: Calvin's full calibration curve, which provides his overall **hit rate** on the test—the proportion of his answers on the *whole* test (not just amongst his 80%-opinions) that were true. For example, perhaps across all his answers you learn that only 50% were correct. How should that affect your Deference to the rational confidence for Calvin to have in his 80%-opinions in particular?

It breaks Deference. Instead of being disposed to completely adopt whatever opinions were rational for him to have, you should instead temper these conditional opinions toward the hit rate. After all, Calvin didn't know his hit rate was he was forming his guesses, so you have evidence about the truth-values of his answers that he did not—so Deference fails once you now this. And because (as I'll argue in §4.1) hit rates often don't themselves provide strong evidence about Calvin's rationality, this failure of Deference causes a failure of the rational-to-right inference.

To see why, start with a simple version of Bianca's case (which I owe to [XXX].) Suppose in our archive all the tablets come from one of two coins—one that is 60% biased towards heads, the other that's 90%. Suppose we know that Bianca *can* decipher the coins (so, by analogy, we expect her to be calibrated iff we expect the rational opinions for Calvin to be calibrated). If all you know is that we've chosen random tablets from the archive, then you should expect her to be calibrated—but you should *also* expect her hit rate to be around 75%. After all, she'll always guess heads (every coin is biased toward heads), we expect roughly 90% of the 90%-biased coins to land heads and 60% of the 60%-biased coins to do so, and we expect they'll be roughly a 50-50 split between them (and $0.5 \cdot 0.6 + 0.5 \cdot 0.9 = 0.75$).

Now we learn Bianca's hit rate and—surprise!—it's below 75%. Should we still expect her to be calibrated? (Analogy: should we still expect the rational opinions for Calvin to be calibrated once we know his overall hit rate?) No. This is easy to see in extreme cases: if the hit rate is *very* low (say, 50%), we know she won't be calibrated, since the lowest credence she'll assign is 60%. Similarly if it's less extreme: whenever the hit-rate is below 75%, we should expect Bianca to be over-calibrated. The connection between the biases of the coins and the frequency with which they land heads is probabilistic and therefore loose. Thus learning that the coins landed heads less often than you'd expect provides evidence that this is one of the cases where the biases and the frequencies came apart. That means that even upon learning that the bias of a given toss was 60% (90%), you should temper your deference downward and be *less* than 60% (90%) confident that it landed heads.

Likewise for Calvin: when you learn that his hit rate was lower (or higher) than

you expected, this provides evidence that it was one of the scenarios in which there's a gap between being rational and being right—that fewer (or more) of Calvin's guesses were correct than he'd be rational to expect. You now know more than he did when he formed his opinions. Thus you should temper your deference downwards (or upwards): conditional on the average rational confidence in his answers being $x\%$, you should be less than (more than) $x\%$ confident in a given answer. So long as Calvin's hit rate is not *itself* significant evidence about his rationality (as argued in §4.1), this failure of Deference leads to a failure of the rational-to-right inference

Now more carefully. Let's assume—what I'll argue for in §4.1—that on many binary-question tests, Calvin's hit rate, H, will be (approximately) equal to the *rational* hit rate, H_r, i.e. the hit rate he'd have if his degrees of confidence were rational and he guessed accordingly. In other words, assume that, when faced with the question, "Which is true: p or p'?", he'll usually be more confident of p than p' iff he'd be *rational* to be more confident of p than p'. (He may of course still be more confident of p than he should be—perhaps the rational degrees of confidence are 60-40 in p-vs.-p', but his degrees of confidence are 80-20.)

Granting that $H = H_r$ makes Calvin's case analogous to the simple version of the Bianca scenario from this section, where we knew she'd guess the way she would if she knew the bias of the coin since all the coins are biased toward heads. Under this assumption, we can see why learning hit rates will break the rational-to-right inference. Consider whether Calvin's 80%-opinions are rational. Learning his hit rate does not significantly affect your opinion on this question—after all, learning (merely) the *rational* hit rate shouldn't affect your opinion whether his 80% opinions are rational, and we're granting that his hit rate equals his rational hit rate.[19] So learning his hit rate doesn't shift your opinions in his rationality. But it *does* shift your opinions in the truth-values of his answers—for example, if his hit rate is low, you know many of his answers are wrong. This gives you information he didn't have when he formed his opinion. Therefore this shift in your opinions about the truth-values breaks Deference.

For instance, suppose you learn that Calvin's hit rate is 50%. (75% is normal, since it's the average of $50-100\%$.) Then upon learning that the average rational confidence for him is 80%, you should *less* confident than that, since you know that more of them are false than he (rationally) expected:

$$P(q_i|\overline{R} = 0.8, H = 0.5) < 0.8, \quad \text{even though}$$
$$P(q_i|\overline{R} = 0.8) = 0.8$$

Thus conditional on his 80%-opinions being rational, you should only be (say) 70% confident in each being true. And conditional on his 80% opinions being *over*confident, you should be even less confident—say, 60%—in each one being true. If so, then—by exactly parallel reasoning to that at the end of §3—learning that only 70% of his 80%-opinions are true (he's slightly over-calibrated) is evidence that he's *rational*. After all, that's exactly what you'd expect if he's rational, given his low hit rate! The rational-to-

[19] Precisely: for any t, s, $P(\overline{R} = s|H = t) \approx P(\overline{R} = s|H_r = t) = P(\overline{R} = s)$.

right inference is inverted. (If you learn that Calvin's hit rate is abnormally *high*, the inference will be inverted in the other direction.)

Toy example: suppose he has 50 different 80%-opinions, and you're initially equally confident that the average rational confidence ($\overline{R}$) is any of $60\%, 61\%, ..., 99\%$. Learning his hit rate was 50% doesn't affect your confidence in any of these hypotheses, but it does temper your deference downward by (say) 10%: $P(q_i|\overline{R} = x, H = 0.5) = x - 0.10$. (So conditional on his 80%-opinions being rational, you should be 70% confident in each.) Then upon learning that 70% of his 80%-opinions are true, you should *decrease* your confidence that Calvin is substantially overconfident ($\overline{R} < 0.75$) from 37.5% to 22%—and *increase* your confidence that he's approximately rational ($0.75 \leq \overline{R} \leq 0.85$) from 27.5% to 61%—inverting the effect from the end of §3.

In summary:

> **Hit Rates are Key:** The rational-to-right inference works only when (rational) hit rates are moderate—on any set of questions on which (rational) hit rates are high (or low), rational deviations from calibration should be expected.

This qualitative claim raises a quantitative question: *how much* deviation from calibration should we expect as hit rates vary? In §5 I'll show how we can answer this question under the assumption that people's actual hit rates match the rational ones; so first, we need to clarify why this is often a reasonable assumption.

4.1 (Rational) Hit Rates

Here I'll make the case that in many binary-question tests the default hypothesis should be that Calvin will tend to guess the way he would were his opinions rational—and, therefore, that his hit rate should be expected to be close to the rational hit rate.

So far we've been focusing on the "overconfidence effect"—but 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 3.

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). 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 3 (Brenner et al. 2005).

Notice a prediction of any such hypothesis: if the alternative claims someone is guessing between are from the same domain, then *people's guesses on binary-choice questions will tend to be rational.* All proposed (ir)rationality hypotheses have positive slopes: on average, higher rational degrees of confidence correspond to higher actual degrees of confidence. Take any such (ir)rationality hypothesis, and consider a guess

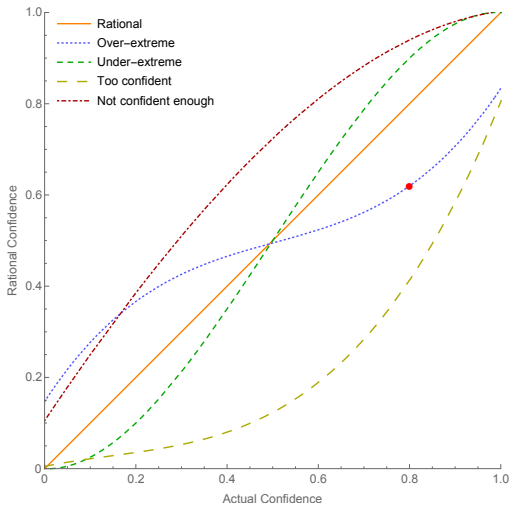

Figure 3: The Various (Ir)rationality Hypotheses

between a pair of claims that it treats as in the same domain—say, "Which is bigger: Denver or Phoenix?" The rational guess for Calvin is the one that he should assign higher confidence to: if $R(Denver) > R(Phoenix)$, it's rational for Calvin to guess *Denver*; and if $R(Denver) < R(Phoenix)$, it's rational for Calvin to guess *Phoenix*. 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 (usually) be higher in *Denver* than in *Phoenix*—i.e. he'll guess *Denver*; and if the latter, Calvin's actual confidence will (usually) be higher in *Phoenix*—i.e. he'll guess *Phoenix*. Either way, the (ir)rationality hypothesis predicts that Calvin will guess as he would if he were rational.[20]

Upshot: in binary-choice calibration studies, so long as both answers come from the same domain, all (ir)rationality hypotheses will agree that people will tend to guess the way they would if they were rational. Since guesses determine hit rates, their actual hit rate should be expected to be close to the rational one—and thus should not be taken to provide significant evidence about rationality

5 The Implications

We often shouldn't expect rational opinions to be calibrated on sets of questions for which the hit rates turn out to be high or low. Thus no matter how carefully we construct our test, we cannot evaluate whether people are overconfident simply by checking

[20]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 $C(q) > C(p)$. Since f is an average, there will be exceptions to this connection between rational and actual guesses. But if we use the *average* hit rate (across subjects) on a test—as experiments usually do—such deviations from rationality should cancel out, and the average hit rate should be close to the average rational one.

whether they're calibrated.

What should we do instead? My proposal is that we use simulations of the Bianca scenario to predict the rational *deviations* from calibration given our test setup, and then compare observed calibration curves to those predictions.[21] Three steps:

1) Choose a test-construction procedure, along with a hypothesis about the degree to which this procedure will lead to deviations from Deference and Independence.
2) Translate that hypothesis into the Bianca analogy and use it to build a simulation of the rational opinions.
3) Compare the predicted calibration curves for rational opinions from this simulation to the actual calibration curves we observe.

I'll spend the rest of this paper illustrating how this methodology can work, arguing that it calls into question the standard interpretation of certain empirical effects.

Studies do not always find the "overconfidence effect". 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 Figure 4. The hard-easy effect has been called "fundamental bias in

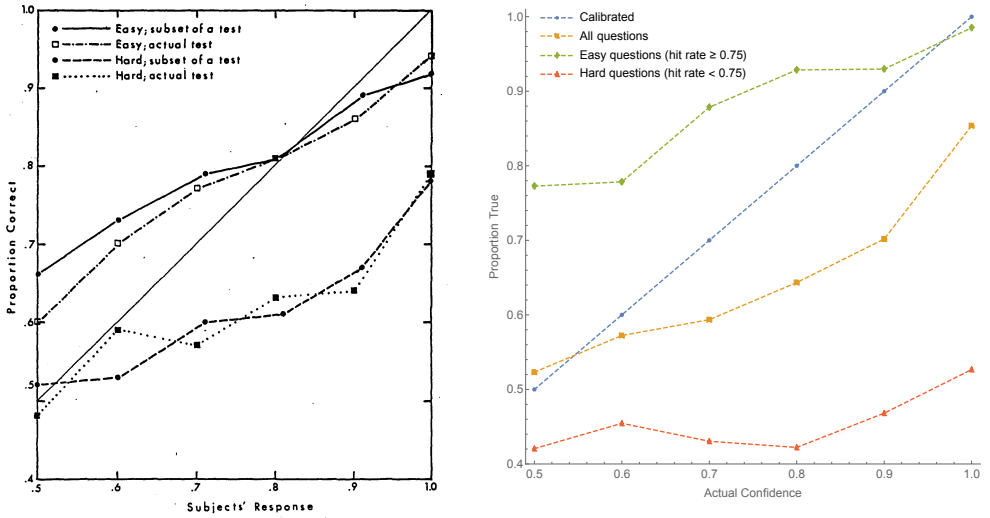

Figure 4: The hard-easy effect. In both graphs, top curves are easy sets of questions; bottom curves are hard ones. **Left:** Lichtenstein et al. (1982). **Right:** My study.

general-knowledge calibration" (Koehler et al. 2002, 687). The idea is that it shows that people do not sufficiently adjust for task-difficulty, leading them to be overconfident on

[21]The simulation of rational opinions to predict miscalibration was pioneered by Erev et al. 1994; Pfeifer 1994; Juslin et al. 1997, 1999. What I'm arguing here is that such simulations are not the special purview of those testing particular rational models of confidence-formation—rather, they are a necessary precondition for figuring out what the null hypothesis figuring out what to expect the *rational* calibration curves to look like on our tests.

hard tests and underconfident on easy ones. Thus it's widely cited as one of the core pieces of evidence in favor of irrational explanations of miscalibration.[22]

Should it be? Is the hard-easy effect evidence for irrationality? By analogy: would finding a hard-easy effect for *Bianca* be evidence that she cannot properly decipher the coins' biases, i.e. that she tends to mis-categorize them? The answer depends on what we should expect if she *can* (at least usually) decipher them. If the hard-easy effect is unexpected given that she can do so, then the effect is clear evidence that she can't decipher them; if it is expected, then the effect does not provide clear evidence that she can't decipher them. Likewise for Calvin: to know whether the hard-easy effect is evidence for his irrationality—i.e. that he tends to mis-categorize propositions into their proper levels of confidence, given his evidence—we have to know whether to expect it given that he *is* (at least approximately) rational.

To answer this question, let's model Bianca. Assume that she *can* usually decipher the tablet markings—setting her confidence at least approximately equal to the biases of the coins—and go on to simulate what calibration curves we should expect from her as we vary the difficulty of various sets of questions.

Step 1 is to choose our test-construction procedure, and form a hypothesis about how this procedure will sample from the rational opinions and the right opinions. In particular: (i) how likely are we to include a question on which the rational credence in the answer is 50%? 60%? Etc. (ii) And on any given test we give, do we defer to the rational opinions? If so, how robust is that deference—does Independence hold, or would learning of false (true) answers temper our deference away from the rational confidence? These questions matter because they affect (i) how often our simulations present Bianca with coins of various biases, and (ii) how robustly the bias of the coins lines up with our expectations about how many of them land heads.

For illustration, focus on the simplest case: a test on which we can reasonably suppose that (i) the questions we pull are *equally* likely to have any level of rational confidence in their guess, between $50 - 100\%$; and on which (ii) our deference is quite robust. One way to try to form such a test is to make one on which we pull questions randomly from a well-defined, representative domain on which we can expect that the accuracy of people's evidence will not be systematically correlated across questions.

This turns out to be a difficult criterion to meet, but I'll take a standard paradigm from the literature (Gigerenzer et al. 1991), pulling pairs of American cities randomly from the top-20 most populous cities, and ask people which they think has a bigger population. On a test like this, it's a reasonable first-pass assumption that the rational opinions in answers will be reasonably uniformly distributed. (The results don't depend heavily on this assumption.) How robust your deference should be is a more vexed question—if we discover Calvin is wrong about whether San Francisco is larger than Phoenix, should that temper our deference to his evidence about whether San Jose is bigger than Austin? Perhaps—but let's ignore that for just a moment.

[22] 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.

Given this, we can perform Step 2: model (and then simulate) our test using the Bianca analogy. We toss a number of coins (equal to the number of questions on our test), selecting them uniformly at random from coins of varying biases between $50-100\%$[23], have her guess how they'll land and rate her confidence in that guess, and record her calibration curve. This is a single trial. Repeat this procedure thousands of times, and now look at the average results on trials (sets of questions) that have various hit rates. What do we expect to see for question-sets of various hit-rates?

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 the rational confidence).[24] The most plausible rationality hypotheses are ones in which there is some such error—though of course it's worth emphasizing that whenever their *is* such error, the person by hypothesis is not fully rational (cf. Brenner 2000). Yet if such deviations are randomly distributed, there's still a good sense in which people are approximately rational.

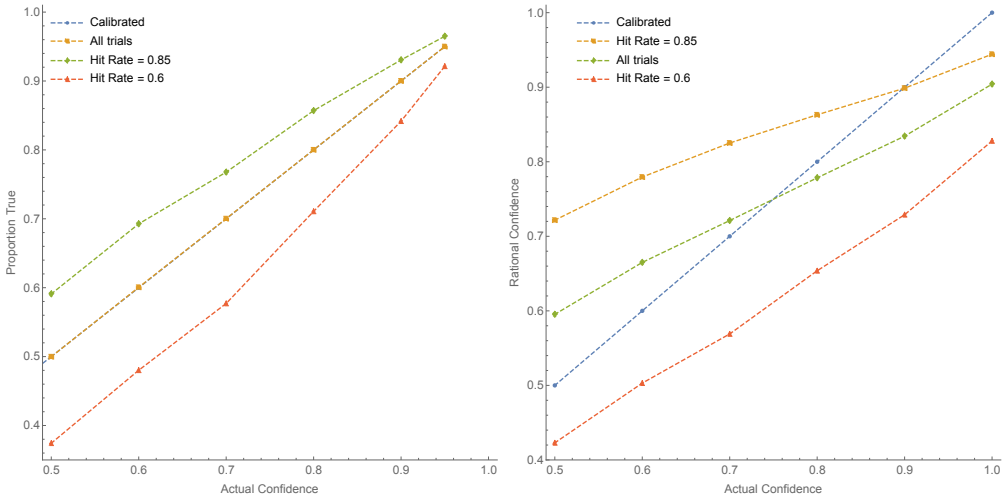

Figure 5: Random tests, restricted to various hit rates. **Left:** Perfection model. **Right:** Noise model (100,000 trials each).

For illustration, the expected calibration curves for Bianca at various hit rates are displayed in Figure 5. When we consider all trials together, Bianca is calibrated—

[23] I simplify by tossing coins of biases 50–100% and having her always guess heads, rather than coins of biases between 0–100% and having her first guess heads or tails. The statistics are the same.

[24] I assume the errors are normally distributed with mean 0; Figure 5 uses standard deviation 0.2. This model takes inspiration from "error models" (Erev et al. 1994; Pfeifer 1994; Juslin et al. 1997, 1999), but the interpretation is importantly different. Their models treat people's reported opinions as imperfect indicators of their *true* opinions (or, in some cases, objective frequencies), whereas mine treats people's reported opinions as imperfect indicators of the *rational* opinions. While tests of variance suggest that error in reporting true confidence cannot account for the observed miscalibration (Budescu et al. 1997), these tests don't examine deviations between reported confidence and rational confidence.

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 tilting of the curve. But amongst tests where the hit rate is low (high), Bianca tends to be over- (or under-)calibrated—just as observed empirically with the hard-easy effect.

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 more of the coins landed tails than you'd usually expect, given their biases. Absent further information, both are likely to play a role in any given trial with a low hit 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 she *can't* account for the second factor. Thus as we consider cases with more extreme hit-rates, Bianca becomes 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 in the best-case test-construction scenario, we should still expect some form of the hard-easy effect 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 5 to the left of Figure 4). Thus the the hard-easy effect in itself is not clear evidence that Bianca is mis-categorizing the coins, nor is it clear evidence that Calvin is irrationally setting his confidence—it's what we should expect even if they weren't.

Let's now perform Step 3 and apply this model to *my* study (pre-registration available at [XXX].) To be clear, this experiment is intended only as a proof of concept, and an illustration of the type of methodology I'm advocating. To give a proper empirical assessment of (something like) the hard-easy effect, we would want to use more systematic and sophisticated empirical and statistical methods.

I generated all pairs from the 20 most-populous U.S. cities, and recruited 200 U.S. residents through Prolific (90 F, 107 M, 3 Other; mean age 34.7). After giving them standard instructions about how to use the 50–100% confidence scale, I presented each with 21 pairs—20 randomly selected from the 190 pairs, and 1 attention check. (I excluded the 1 participant who failed the attention-check.)

I pooled subjects' answers, and divided the questions into those that were *easy* (more than 75% of answers correct) and those that were *hard* (less than 75% correct). Figure 4 (page 18) above shows the calibration curves from my study overall, amongst the hard questions, and amongst easy ones. The hard-easy effect was observed as expected—though it was especially stark. Amongst hard questions the average confidence was 75.1%, while the proportion true was only 45.2%.[25] Meanwhile amongst easy questions,

[25] Average confidence in hard questions ($M = 0.751$, $SD = 0.165$) was above proportion true of hard

the average confidence was only 84.7%, while the proportion true was 92.1%.[26] Unexpectedly, the test overall was slightly hard, with an average confidence of 79.8% and a proportion true of only 68.0%—hence the over-calibration observed for all questions.[27]

We can compare these results to both the perfection-model and noise-model predictions. As pre-registered, I generated these simulations by setting the number of questions Bianca faces to the size of the easy/hard/all-questions set, simulating millions of trials, and then removing trials with high/low hit rates until the mean hit rate matched the actual hit rate in the easy/hard/all-questions sets. The perfection model has no free parameters; its comparisons to the data from my study are displayed on the left of Figure 6. Though the predicted curves deviate substantially from calibration, they do have steeper slopes than the observed curves.

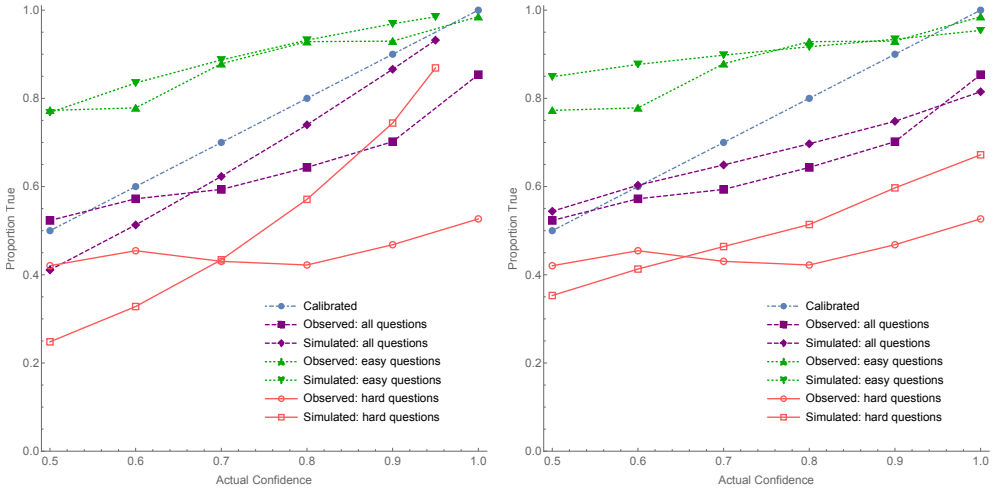

Figure 6: Random tests run with the observed hit-rates in my study. **Left:** Perfection model (5 million trials). **Right:** Noise model (8 million trials, noise parameter = 0.3).

Meanwhile, the noise model has a free parameter for the standard deviations from rational confidence. As pre-registered, I ran versions of the simulations with the parameter varying from 0 – 0.3, and chose the one with the resulting predicted calibration curves that minimized mean squared divergence between the model predictions (amongst hard and easy subsets) and the actual curves. This set the noise parameter to 0.3 (admittedly, a worryingly-high value), and the resulting comparison between simulations and the data in my study is displayed on the right of Figure 6. The predictions generated from the rational-credence-plus-noise model are generally close to the observed mis-calibration.

But remember that these simulations assumed Deference was quite robust (i.e. Independence held fully). Looking at the data, this seems wrong. It was incredibly difficult

questions ($M = 0.452$, $SD = 0.498$), with $t(2487) = 25.82, p < 0.001$, and $d = 0.808$ (one-sided).

[26] Average confidence in easy questions ($M = 0.847$, $SD = 0.162$) was below proportion-true of easy questions ($M = 0.921$, $SD = 0.270$), with $t(3156) = 10.37, p < 0.001$, and $d = 0.333$ (one-sided).

[27] Average confidence across all questions ($M = 0.798$, $SD = 0.170$) differed from proportion true ($M = 0.680$, $SD = 0.467$), with $t(5013) = 14.97$, $p < 0.001$, $d = 0.336$ (two-sided).

to find simulation-runs leading to hit rates as extreme as observed in the real study (of 8 million trials, only 183 had hit rates at or below 0.515, while my study's hard questions had a hit rate of 0.452). A natural explanation is that for tests that share a *common subject-matter*—such as my city-comparison tests, and many others[28]—we need to revise our assumption of Independence, since each person's evidence will be highly correlated across questions. (Recall Georgie: if she's wrong about several geographical opinions, we should think it more likely that she has misleading evidence about others.)

In particular, though we should expect that the opinions warranted by their evidence will *on the whole* be calibrated, we should also expect that a there will be random fluctuations in how calibrated they are across subject-matters. For instance, in my city-comparison test, some subjects will have evidence that warrants misleadingly strong opinions (only 70% of the opinions they should be 80% confident in are true), while others will have evidence that warrants misleadingly weak opinions (90% of the opinions they should be 80% confident in are true). Moreover, we expect these fluctuations in evidence to be correlated for a given person on a given subject-matter—if only 50% of the opinions Calvin ought to be 60% confident in on my test are true, we should expect that (say) only 60% of the ones he ought to be 70% confident in are true.

Here's a simple 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 have higher proportions of heads from a coin of a given bias than would be expected; other have lower proportions. 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]

Although I had constructed these models before running my city-calibration test, it only occurred to me that they were an apt model of it after running the test and seeing how extreme the variation in hit-rates were. (The empirically-observed curves looked familiar...) As a result, these comparisons were not pre-registered and should be taken with several grains of salt. But it turns out to be *much* easier to find hit-rates as extreme as the ones we observed using this model, lending the model some support. Running the same analysis as above yields the optimal noise parameter at 0.15, and yields the comparisons in Figure 7 (page 24).

These statistical methods are preliminary, and the models are by no means a perfect fit. Nevertheless, qualitative effects like the hard-easy effect *are* predicted—and when we incorporate the possibility of either noise in subject's judgments *or* random misleadingness in subject's evidence (or both), the observed calibration curves are close to what we should expect from rational people. It is the (much smaller) deviations from *these* predicted curves that we must study systematically—not the deviations between people's actual confidence and the perfectly-calibrated line.

[28] E.g. Dunning et al. 1990; Vallone et al. 1990; Brenner et al. 1996; Koehler et al. 2002; Brenner et al. 2005; Hoffrage 2004; Glaser and Weber 2007; Merkle and Weber 2011; Brenner et al. 2012.

[29] In the displayed simulations this parameter is normally distributed with mean 0 and (for illustration) standard deviation 0.2. In these 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%.

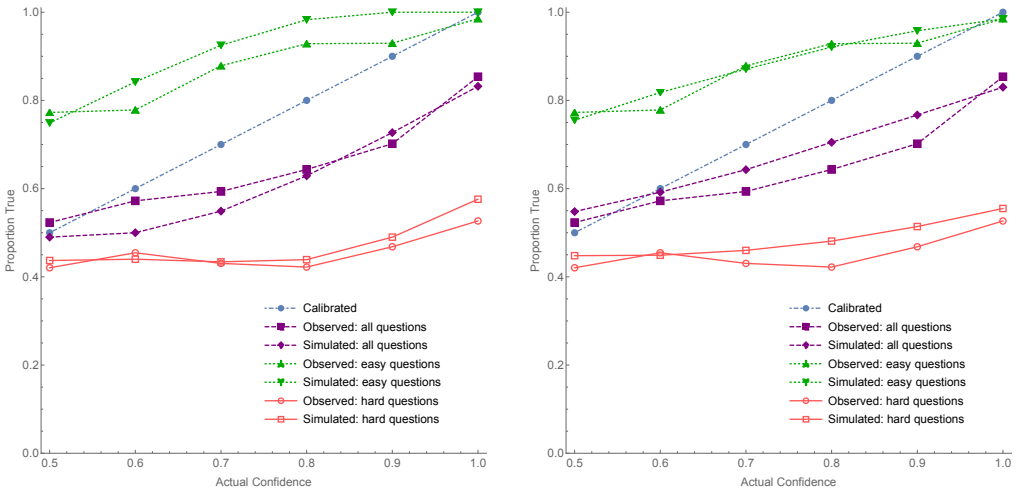

Figure 7: Tests with random misleadingness (SD = 0.2), run using the observed hit rates in my study (20,000 trials). **Left:** Perfection model. **Right:** Noise model, parameter = 0.15.

6 More Questions

I'll close by considering how this theory of rational (mis)calibration depends on the philosophical tenability of Deference (§6.1), as well as its implications for other forms of calibration studies (§6.2).

6.1 The Tenability of Deference

Deference is key. But, as discussed in §3, Deference is the strongest tenable interpersonal deference principle—and there are many reasons to be worried about it. First, if epistemic rationality is *permissive*, then you may have different epistemic standards than Calvin (White 2005; Schoenfield 2014, 2019). If so, the fact that *his* standards rationalize being 80%-confident in q doesn't imply that *your* standards do. Perhaps your standards warrant being systematically more cautious (less extreme) in your opinions than Calvin's do (Greco and Hedden 2016). If so, Deference will fail. Open question: can permissivists justify the rational-to-right inference, even in the best-case scenarios?

Similarly, if *epistemic modesty* can be rational—it can be rational to be unsure of what opinions are rational—then Deference must sometimes fail (Christensen 2010b; Elga 2013; Dorst 2020a). The only deference principle I know of that is both tenable in the case of modesty, and would warrant a variant of the above reasoning from §3 is a version of the "Trust" principle in Dorst 2020a, formulated in Dorst et al. 2021.[30] Other approaches to modesty (Elga 2013; Pettigrew and Titelbaum 2014; Lasonen-Aarnio 2015; Williamson 2019; Gallow 2021) allow large deviations from Deference. Open question: can they justify the rational-to-right inference, even in ideal scenarios?

[30]The variant reasoning requires pooling people's opinions into categories like "*at least* 60%" confident (instead of "exactly 60%") and seeing whether at least 60% (rather than exactly 60%) of them are true.

6.2 Dunning-Kruger and Overprecision

In addition to the binary-question tests I've focused on, there are a variety of other ways of measuring calibration. My arguments raise two salient questions about them: (1) given that tests can be hard even for rational people, to what degree should we expect rational people to be calibrated? And (2) to the extent that such tests involve *guessing*, what should we expect (rational) people to guess?

(1) First, the fact that we should often expect rational people to be over-calibrated has implications for both the Dunning-Kruger effect (Kruger and Dunning 1999) and the finding of "over-precision" (Moore et al. 2015b).

The Dunning-Kruger effect is the finding that the gap between a person's relative performance on a test and their *estimate* of this number grows as their performance decreases. For example, those in the 50th percentile estimate that they're in the 60th, while those in the 20th estimate they're in the 50th. This is often thought to indicate irrationality (Dunning 2012). But we've seen that for any set of (rational) opinions, some tests will be hard (have low hit rates)—and that as the test gets harder, the gap between hit rate and estimated hit-rate grows (§5). This supports the rational models of the Dunning-Kruger effect proposed by Moore and Healy 2008 and Jansen et al. 2018.

Meanwhile, over-precision is found on *interval-estimation tests*: ask people to give confidence intervals for the true value of some unknown parameter, like the length of the Amazon. People tend to be quite "over-precise" in the sense that their 90% confidence intervals standardly miss the true value as much as 50% of the time. Some take this to be better evidence for overconfidence than the over-calibration found in binary-question tests.[31] Although this may be correct, translating between interval- and binary-question tests (Tversky and Kahneman 1974) gives me pause: Calvin's 90% confidence interval for the length of the Amazon is "1000–5000 miles" iff he's 95% confident in both "It's at *L*east 1000" (L) and "It's at *M*ost 5000". Thus we should expect a rational miss-rate of 50% iff, given our background information, $P(L\&M) \approx 0.5$. This does not seem implausible. According to our simulations (Figures 5, 6, and 7), on hard tests we should expect no more than (and often much less than) 75% of people's 95%-opinions to be true, i.e. $P(L) \leq 0.75$ and $P(M) \leq 0.75$. Thus *if L and M were independent*, we should expect a miss-rate of at least 44% ($1 - 0.75 \cdot 0.75 \approx 0.44$). Yet by definition L and M are *not* independent: if L is false, M must be true; thus L being true makes M less likely, meaning we should expect even higher miss-rates—50% is not surprising.[32]

(2) Finally, my analysis of binary-question tests has relied on the (standard) assumption that people will guess the answer they think is most likely. But for questions with more than two complete answers, the relationship between (rational) credences and guessing turns out to be much more complicated than this (Kahneman et al. 1982; Holguín 2020; Dorst and Mandelkern 2021; cf. Horowitz 2017)—sometimes it makes sense to guess an answer that's improbable so long as it's specific (informative) enough (cf. Levi 1967). I don't know how exactly this will affect the analysis of other cases,

[31] Moore and Healy 2008; Glaser and Weber 2010; Ortoleva and Snowberg 2015; Moore et al. 2015a,b.

[32] Since $P(M|\neg L) = 1 > P(M)$, $P(M|L) < P(M)$. Thus if L lowers the probability of M by 10% (so $P(M|L) \leq 0.65$), then $P(L \wedge M) \leq 0.75 \cdot 0.65 \approx 0.49$.

but it does raise to the fore the question of what we should expect *rational* people to guess in contexts—like interval-estimation—in which the question under discussion is not a binary one. In particular, it shows that in order to perform the rational-to-right inference, we must either provide evidence that people do *not* guess the way they would if they were rational, or else control for expected rational deviations from calibration due to rational variations in hit rates, as I've done here.

7 The Upshot

Many have taken the results of calibration studies to demonstrate that people tend to be systematically overconfident in a way that is both dire and preventable. I've argued that the theoretical foundations of this inference are shaky (§2), but that we can secure them by articulating a probabilistic connection between being rational and being right (§3). But these foundations reveal a methodological flaw: no matter how well-designed the study or how rational people form their opinions, they should still be expected to be miscalibrated in systematic ways (§4). I used this result to propose an amended methodology: we must use information about our study (including hit rates and potential failures of Independence) to predict the rational *deviations* from calibration, and then compare people's performance to those predictions. I illustrated how this can be done, and argued that it has the potential to change the standard interpretation of robust empirical effects (§§5–6).

If even a portion of this discussion is correct, it suggests that certain debates in philosophy and psychology are much closer than has been realized. Psychologists have had a spirited debate about the bearing of empirical results (like calibration studies) on human rationality.[33] Yet most contemporary philosophical debates about rationality have been relatively isolated from these issues (but see footnote 7). As we've seen, these debates needn't—and arguably *shouldn't*—be isolated. Whether to what extent we have empirical evidence for overconfidence depends on the connection between being rational and being right, which in turn is directly dependent on philosophical debates about the nature of evidence, deference principles, permissivism, and epistemic modesty (§3, §6.1). Conversely, the methods of calibration studies bring to the philosophical literature the idea that there is an *evidential* connection between being rational and being right (§3), and that certain simulation-style methods can be used to make precise predictions about this connection in a variety of settings (§5).

In short: the questions and methods from both philosophical and psychological investigations of rationality can be tied together in surprising and fruitful ways. That raises an exciting question: If we bring these investigations closer together, what other

[33]For classic statements of the "irrationalist" approach, see Tversky and Kahneman 1974, 1983; Kahneman et al. 1982; Kahneman and Tversky 1996; Fine 2005; Ariely 2008; Hastie and Dawes 2009; Kahneman 2011b; Thaler 2015. For defenses of "rational" approaches see Anderson (1990); Gigerenzer (1991); Oaksford and Chater (1994, 2007); Tenenbaum and Griffiths (2006); Hahn and Oaksford (2007); Hahn and Harris (2014); Harris and Hahn (2011); Tenenbaum et al. (2011); Griffiths et al. (2012); Cushman (2018).

ties might be find?[34]

A 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 I focus on Calvin's 80%-opinions; obviously the reasoning generalizes.)

Assuming that R is a probability function, Deference follows from two 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 : $P(q_i|R = \delta) = \delta(q_i)$.[35]

> **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 plausible in the situations we're considering (all you know about the q_i is that 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, 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 $\mathbb{E}[\mathbb{1}_{q_i}] = P(q_i)$. 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 $\overline{R} = x \Leftrightarrow R \in D_x$.

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

$$\mathbb{E}[\textstyle\sum \frac{\mathbb{1}_{q_i}}{n} \mid \overline{R} = x] \;=\; \sum_{\delta \in D_x} P(R = \delta \mid \overline{R} = x) \cdot \mathbb{E}[\textstyle\sum \frac{\mathbb{1}_{q_i}}{n} \mid R = \delta]$$

[34] Thanks to [redacted].

[35] 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 (see Schervish et al. 2004; Dorst 2019).

By linearity of expectations, this equals

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

Therefore $\mathbb{E}[\sum \frac{\mathbb{1}_{q_i}}{n} \mid \overline{R} = x] = x$, 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, establishing Deference.

A.2 The Rational-to-Right 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 i.i.d. 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: $P(\overline{q} = sn | \overline{R} = t_i) = \binom{n}{sn} t_i^{sn} (1 - t_i)^{n - sn}$.

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.

Angner, Erik, 2006. 'Economists as experts: Overconfidence in theory and practice'. *Journal of Economic Methodology*, 13(1):1–24.

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.

Benoît, Jean-Pierre and Dubra, Juan, 2011. 'Apparent Overconfidence'. *Econometrica*, 79(5):1591–1625.

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, 2000. 'Should Observed Overconfidence Be Dismissed as a Statistical Artifact? Critique of Erev , Wallsten , and Budescu (1994)'. 107(4):943–946.

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.

Brenner, Lyle A, Griffin, Dale W, and Koehler, Derek J, 2012. 'A Case-Based Model of Probability and Pricing Judgments : Biases in Buying and Selling Uncertainty'. 58(1):159–178.

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 Rose, 2020. 'Imprecise Evidence without Imprecise Credences'. *Philosophical Studies*, 177(9):2735–2758.

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*, 13(4):397–422.

Comesaña, Juan, 2020. *Being Rational and Being Right*. Oxford University Press.

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

Crupi, Vincenzo, Tentori, Katya, and Lombardi, Luigi, 2009. 'Pseudodiagnosticity Revisited'. *Psychological Review*, 116(4):971–985.

Cushman, Fiery, 2018. 'Rationalization is rational'. 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.

Doody, Ryan, 2020. 'The Sunk Cost Fallacy Is Not a Fallacy'. *Ergo*, 6(40):1153–1190.

Dorst, Kevin, 2019. 'Higher-Order Uncertainty'. In Mattias Skipper Rasmussen and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 35–61. Oxford University Press.

———, 2020a. 'Evidence: A Guide for the Uncertain'. *Philosophy and Phenomenological Research*, 100(3):586–632.

———, 2020b. 'Higher-Order Evidence'. In Maria Lasonen-Aarnio and Clayton Littlejohn, eds., *The Routledge Handbook for the Philosophy of Evidence*. Routledge.

Dorst, Kevin, Levinstein, Benjamin, Salow, Bernhard, Husic, Brooke E., and Fitelson, Branden, 2021. 'Deference Done Better'. *Philosophical Perspectives*, To appear.

Dorst, Kevin and Mandelkern, Matthew, 2021. 'Good Guesses'. *Philosophy and Phenomenological Research*, To appear.

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.

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.

Gallow, J. Dmitri, 2021. 'Updating for Externalists'. *Noûs*, 55(3):487–516.

Gibbons, John, 2013. *The Norm of Belief*. Oxford University Press.

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 Harris, Adam J.L., 2014. 'What Does It Mean to be Biased. Motivated Reasoning and Rationality.' In *Psychology of Learning and Motivation - Advances in Research and Theory*, volume 61, 41–102.

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.

Harvey, Nigel, 1997. 'Confidence in judgment'. *Trends in cognitive sciences*, 1(2):78–82.

Hastie, Reid and Dawes, Robyn M, 2009. *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.

Hedden, Brian, 2019. 'Hindsight Bias is not a Bias'. *Analysis*, 79(1):43–52.

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

Holguín, Ben, 2020. 'Thinking, Guessing, and Believing'.

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

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

———, 2017. 'Accuracy and Educated Guesses'. In *Oxford Studies in Epistemology*. Oxford University Press.

———, 2019a. 'Predictably Misleading Evidence'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 105–123. Oxford University Press.

———, 2019b. '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.

Jansen, Rachel, Rafferty, Anna N, and Griffiths, Tom, 2018. 'Modeling the Dunning-Kruger Effect: A Rational Account of Inaccurate Self-Assessment.' In *CogSci*.

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, Olsson, Henrik, and Björkman, Mats, 1997. 'Brunswikian and thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment'. *Journal of Behavioral Decision Making*, 10(3):189–209.

Juslin, Peter, Wennerholm, Pia, and Olsson, Henrik, 1999. 'Format Dependence in Subjective Probability Calibration'. *Journal of Experimental Psychology: Learning Memory and Cognition*, 25(4):1038–1052.

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, 2011a. 'Don't Blink! The Hazards of Confidence'.

———, 2011b. *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, 1996. 'On the reality of cognitive illusions.'

Karlan, Brett, 2021. 'Reasoning with heuristics'. *Ratio*, 34(2):100–108.

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.

Kinney, David and Bright, Liam, 2021. 'Elite Group Ignorance'. *Philosophy and Phenomenological Research*, To appear.

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.

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.

Lam, Barry, 2011. 'On the rationality of belief-invariance in light of peer disagreement'. *Philosophical Review*, 120(2):207–245.

———, 2013. 'Calibrated probabilities and the epistemology of disagreement'. *Synthese*, 190(6):1079–1098.

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*, 144–171.

Levi, Isaac, 1967. *Gambling with Truth*. The MIT Press.

Lewis, David, 1980. 'A subjectivist's guide to objective chance'. In Richard C Jeffrey, ed., *Studies in Inductive Logic and Probability*, volume 2, 263–293. University of California Press.
———, 1994. 'Humean Supervenience Debugged'. *Mind*, 103(412):473–490.
Lewis, Michael, 2016. *The undoing project: A friendship that changed the world*. Penguin UK.
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.
Littlejohn, Clayton, 2012. *Justification and the truth-connection*. Cambridge University Press.
———, 2018. 'Stop making sense? On a puzzle about rationality'. *Philosophy and Phenomenological Research*, 96(2):257–272.
Lord, Errol, 2018. *The importance of being rational*. Oxford University Press.
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.
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, Carter, Ashli B, and Yang, Heather H J, 2015a. 'Organizational Behavior and Human Decision Processes Wide of the mark : Evidence on the underlying causes of overprecision in judgment'. 131:110–120.
Moore, Don A and Healy, Paul J, 2008. 'The trouble with overconfidence.' *Psychological review*, 115(2):502.
Moore, Don A, Tenney, Elizabeth R, and Haran, Uriel, 2015b. 'Overprecision in judgment'. *The Wiley Blackwell handbook of judgment and decision making*, 2:182–209.
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.
———, 2007. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.
O'Connor, Cailin and Weatherall, James Owen, 2018. 'Scientific Polarization'. *European Journal for Philosophy of Science*, 8(3):855–875.
Odean, Terrance, 1999. 'Do investor trade too much American economic review.1999.pdf'.
Ortoleva, Pietro and Snowberg, Erik, 2015. 'Overconfidence in political behavior'. *American Economic Review*, 105(2):504–535.
Pettigrew, Richard, 2016. 'JAMESIAN EPISTEMOLOGY FORMALISED: AN EXPLICATION OF THE WILL TO BELIEVE''. *Episteme*, 13(03):253–268.
Pettigrew, Richard and Titelbaum, Michael G, 2014. 'Deference Done Right'. *Philosopher's Imprint*, 14(35):1–19.
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.
Plous, Scott, 1993. *The psychology of judgment and decision making*. Mcgraw-Hill Book Company.
Quilty-Dunn, Jake, 2020. 'Unconscious Rationalization, or: How (Not) To Think About Awfulness and Death'.
Rinard, Susanna, 2019. 'Believing for practical reasons'. *Noûs*, 53(4):763–784.
Roush, Sherrilyn, 2009. 'Second Guessing: A Self-Help Manual'. *Episteme*, 251–268.
———, 2016. 'Knowledge of Our Own Beliefs'. *Philosophy and Phenomenological Research*, 93(3):45—-69.
———, 2017. 'Epistemic Self-Doubt'.
Salow, Bernhard, 2018. 'The Externalist's Guide to Fishing for Compliments'. *Mind*, 127(507):691–728.
Schervish, M. J., Seidenfeld, T., and Kadane, J.B., 2004. 'Stopping to Reflect'. *The Journal of Philosophy*, 101(6):315–322.
Schoenfield, Miriam, 2012. 'Chilling out on epistemic rationality'. *Philosophical Studies*, 158(2):197–219.
———, 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. 'Bridging Rationality and Accuracy'. *Journal of Philosophy*, 112(12):633–657.
———, 2018. 'An Accuracy Based Approach to Higher Order Evidence'. *Philosophy and Phenomenological Research*, 96(3):690–715.
———, 2019. 'Permissivism and the Value of Rationality: A Challenge to the Uniqueness Thesis'. *Philosophy and phenomenological research*, 99(2):286–297.
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.
Shariatmadari, David, 2015. 'Daniel Kahneman: What would I eliminate if I had a magic wand? Overconfidence''.
Singer, Daniel J, Bramson, Aaron, Grim, Patrick, Holman, Bennett, Jung, Jiin, Kovaka, Karen, Ranginani, Anika, and Berger, William J, 2019. 'Rational social and political polarization'. *Philosophical Studies*, 176(9):2243–2267.
Sliwa, Paulina and Horowitz, Sophie, 2015. 'Respecting *All* the Evidence'. *Philosophical Studies*, 172(11):2835–2858.
Staffel, Julia, 2020. *Unsettled thoughts: A theory of degrees of rationality*. Oxford University Press, USA.
Tenenbaum, Joshua B and Griffiths, Thomas L, 2006. 'Optimal Predictions in Everyday Cognition'. *Psychological Science*, 17(9):767–773.

Tenenbaum, Joshua B, Kemp, Charles, Griffiths, Thomas L, and Goodman, Noah D, 2011. 'How to grow a mind: Statistics, structure, and abstraction'. *science*, 331(6022):1279–1285.

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

Thaler, Richard H., 2015. *Misbehaving: The Making of Behavioural Economics*. Penguin.

Thorstad, David, 2021. 'The accuracy-coherence tradeoff in cognition'. *British Journal for the Philosophy of Science*, To appear.

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

———, 1983. 'Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.' *Psychological review*, 90(4):293.

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.

Wedgwood, Ralph, 2017. *The value of rationality*. Oxford University Press.

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

———, 2009a. 'Evidential Symmetry and Mushy Credence'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 3, 161–186. Oxford University Press.

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

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

———, 2019. 'Evidence of Evidence in Epistemic Logic'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 265–297. Oxford University Press.