

Calibrated Probabilities and the Epistemology of Disagreement

Barry Lam
Department of Philosophy
Vassar College

January 6, 2011

1 Comparative Reliability in the Epistemology of Disagreement

The epistemology of disagreement concerns the normative question of how you ought to revise your beliefs in a very specific epistemic context. Imagine that you and a peer form an opinion in isolation about whether P in response to mutually shared evidence, and you take your peer to be just as reliable as you about matters of this kind. How are you to respond should you subsequently discover that your peer disagrees with you? Advocates of the *Equal Weight* view state that in light of known peer disagreement about whether P , you ought to revise your opinion in such a way as to “split the difference” between your own view and that of your peer’s.¹ Let us call this the Equal Weight rule, or EWR. At another extreme, a *Stay the Course* view advocates a rival rule, which we will call STC, where in light of known peer disagreement alone, you ought not revise your previously held opinion about whether P .² Since STC only advocates maintaining an existing opinion in light of new

¹ [Feldman, 2006], [Christensen, 2007], [Elga, 2007]

² [Kelly, 2005] A more moderate view in the same spirit as the Stay the Course view, is defended in [Kelly, 2009]. Kelly’s concern is not primarily the advocacy of a rule like STC that applies in full generality to all cases of known peer disagreement. Rather for Kelly, considerations of who was the most rational in forming the original opinion about the matter can permit Staying the Course or something very close in *some* cases of known peer disagreement.

information, it is not technically a “rule” for belief-revision. Nonetheless, it is clear that EWR and STC are incompatible, and will be called competing “rules” for our purposes. These two rules in no way exhaust the range of theoretical possibilities. However, they do represent two endpoints on a spectrum of rules, and examining them in isolation will allow us to see where we might look on that spectrum for the correct rule.

In this paper, I evaluate the status of EWR and STC from a different perspective from the existing literature on the epistemology of disagreement. Instead of first asking the normative question, “which one of these rules is a rational requirement of belief-revision?”, I will begin by considering a related series of factual questions: will an agent, whose belief-revision methods in some domain are already to a certain degree reliable, improve in reliability if she follows EWR? Will such an agent methods be, at the very least, reliability-preserving? In effect, I will be concerned with what is *in fact* the most reliable way for an agent to revise her beliefs when she is *in fact* equally as reliable as a disagreeing peer, as opposed to how we *ought* to revise our beliefs when we *deem* a disagreeing peer to be just as reliable. Although these questions have not received much attention, they are relevant to the normative status of these belief-revision rules. If we are concerned with the evidential value of peer opinion, we might want to determine how well-correlated, or reliably correlated, such an opinion is with the truth about whether P. Similarly, if we are determining whether to replace one method for forming opinions in a domain with another, the comparative reliability of these methods are surely relevant. The comparative reliability of EWR and STC is a subject that deserves some attention, and it is the primary subject of this paper.

1.1 Calibration

The Equal Weight and Stay the Course rules are best understood as rules for revision of *partial* belief, or doxastic credences, rather than *all-or-nothing* beliefs. We therefore need a measure of reliability for partial belief. Reliability with respect to rules or methods that generate all-or-nothing beliefs can be measured by frequencies of true and false contents.³ Such measures

³Alternatively, we can measure reliability in terms of propensities or dispositions of such rules to generate true or false beliefs in a restricted range of counterfactual circumstances [Alston, 1995]. Though this is extensionally a quite different measure of reliability, what I argue below is just as applicable to an account of reliability in terms of actual frequencies

of reliability are not adequate for measuring the reliability of revision-rules for partial beliefs. The first reason is that such measures require us to determine the “beliefs” of an agent from the set of propositions in which she has some credence, and this is a notoriously difficult philosophical problem.⁴ Secondly, a measure of reliability for partial belief must take into account both the truth-value of the content of belief and also the strength to which the proposition is held. For example, compare two judgments, one of 70% confidence in rain on Friday, and the other of 75% confidence in rain on Friday. Now it rains on Friday. How do I use this fact to assess the reliability of these judgments? In the long run, a judgment of 20% confidence in rain whenever it does not rain seems much better than a judgment of 30% confidence in rain under those same circumstances. At the other extreme, if it always rained whenever you were exactly 80% confident that it would, your judgments would be less accurate than if you were fully confident in rain, since in those evidential conditions, it always rained. If it rained 20% of the time in which you were 20% confident of rain, then your judgments would be much better than judgments of no confidence at all in rain in those same circumstances. Both the truth value of a particular judgment and the confidence level are relevant to the assessment of reliability.

One intuitive extension of our naive measure of all-or-nothing beliefs is *Calibration*.⁵ Whenever your degrees of belief match up well with the ratio of true propositions to total propositions in which you have that degree of belief, then you are *well-calibrated*. Your partial-belief forming methods in a particular domain are *perfectly calibrated* just in case, for all x between 0 and 100, $x\%$ of the propositions in which you are $x\%$ confident are true. Thus, a weather forecaster is perfectly-calibrated if it happens to rain $n\%$ of the days in which he is $n\%$ confident that it will rain, for all n . Calibration is a measure of reliability for credences that captures both accuracy of content and accuracy of probabilistic judgment. Relative to a domain like weather forecasting, your methods of assessing evidence and forming probabilistic judgments on the basis of such assessments are reliable to the degree to which those methods lead you to well-calibrated judgments.

We now require a definition of two agents being *equally well-calibrated* when they are imperfectly calibrated. If it turns out that it rains 85% of the

of true beliefs, or in terms of propensities or dispositions to generate such true beliefs in restricted ranges of counterfactual circumstances.

⁴[Christensen, 2005] and [Kaplan, 1996] contain extensive discussions of this issue.

⁵ [Lichtenstein et al., 1982]

time when both A and B predict a 90% chance of rain, then A and B are imperfectly calibrated but also equally well-calibrated. But if this is true of an equal error rate with respect to predictions of 90% chance of rain, then one cannot take A and B to be differently well-calibrated if it so happens that A's 5% error rate occurs with his 30% predictions, while B's occur with her 90% predictions. It also should not make a difference to A and B's comparative reliability if the error is due to overconfidence for A and underconfidence for B, so long as the error-rates are the same. For instance, when there are 85% rainy days on the days in which A predicts a 90% chance, and there are 95% rainy days among the days in which B predicts a 90% chance of rain, both exhibit a 5% error rate, where A is to that degree overconfident, and B is to that degree underconfident. Generalizing from these considerations, we will say that A and B are imperfectly but equally well-calibrated just in case they have the same overall *averaged* error-rate. A's averaged error-rate is arrived at by averaging the error rates for each class of propositions in which A has a particular degree of belief, where a class is a set $\{P : pr_A(P) = n\}$, ignoring the fact as to whether the error rates are positive or negative. Whenever the averaged error-rate is the same between A and B, the two are equally well-calibrated. Trivially, when two subjects are perfectly calibrated, they are equally well-calibrated. We now define two agents to be equally *reliable* just in case they are equally well-calibrated.⁶

1.2 Closeness to Truth and Brier Scoring

Calibration is not the only nor necessarily the best measure of reliability available for partial belief. To evaluate a belief-revision method, we might want to assess overall accuracy, or *closeness to the truth*, rather than how well a set of probabilities match frequencies. Let the truth value of a certain proposition P be given a numerical value of one for true, or zero for false. Now A and B can make some judgment about whether P as a probability between zero and one. According to a measure of reliability in terms of closeness to truth, a subject's reliability is a function of the distance of a subject's

⁶There are many other ways to measure the degree of calibration of an agent that are more discriminating than the mere absolute value measure I have just defined. For instance, if one subject is always wrong in her 90% judgments, but makes only one of these, while another is always wrong in her 90% judgment but makes seventeen of these, one can make the latter subject less reliable than the former, by weighting not only proportion of error, but frequency of error, in the measure of degree of reliability.

probabilistic judgment from the truth-value of the proposition judged. So, for instance, if A has probability p_A which is distance α from the truth, and B has probability p_B which is distance β from the truth, A is closer than B to the truth with respect to P if and only if $\alpha < \beta$. The Brier Score [Brier, 1950] takes the squared deviation from the truth as the measure of reliability with respect to a proposition (α^2 as the score of A with respect to P, β^2 as the score of B with respect to P). A subject's average Brier Score with respect to all of the propositions in which she has judgments in a domain is the measure of that subject's reliability with respect to that domain. Formally, if $P_i \in \{P_1 \dots P_n\}$ iff $pr_A(P_i) = m$ for some $0 \leq m \leq 1$ and $T(P_i) = 1$ iff P_i is true and $T(P_i) = 0$ iff P_i is false, then the Brier Score for A is:

$$\frac{1}{n} \sum_{i=1}^n (pr_A(P_i) - T(P_i))^2$$

The lower a subject's Brier Score, the closer she is (on average) to the truth. We can define two agents to be equally reliable just in case they have the same Brier Score. ⁷

2 Using Both Measures

Since the Equal Weight and Stay the Course rules are best understood as rules for partial belief-revision, the relevant measures of reliability of these rules ought to be something like Calibration or Brier Scoring. Both measures make use of the degree of probabilistic judgment in addition to the truth-value of its content in evaluating reliability. Both generate two precise definitions of reliability to assess the comparative reliability of EWR and STC, and both generate precise definitions of epistemic peerhood. But which measure should we choose as the correct measure?

Both measures can be well-motivated. Perfectly calibrated subjects break-even in the long run when taking all and only fair bets on all of the propositions in which they have judgments. This is not necessarily true of an imperfectly calibrated subject who has a better Brier Score than some perfectly calibrated subject. Calibration also makes sense of the fact that we are

⁷Brier Scoring is one among many types of accuracy measures called *scoring rules* (See [Joyce, 2008]). For obvious reasons, I cannot provide an exhaustive analysis of scoring-rule evaluations for the Stay the Course and Equal Weight rules here, but the tasks would be very interesting for future research.

rational to respond to relative frequency information in the way we manage our degrees of confidence by setting them equal to the frequencies. If our total evidence indicates that it has rained 6 out of 10 times given the current atmospheric conditions, we should count as being perfectly reliable if we predict a 60% chance of rain and turn out perfectly calibrated. Calibration makes it possible for subjects who have all and only probabilistic evidence to respond probabilistically without sacrificing perfect reliability.

The Brier Score measure, on the other hand, makes only the omniscient perfectly reliable, and therefore has built into the measure a presumption that having a partial belief is a sign of epistemic imperfection. Yet, this feature might make Brier Scoring well-motivated for some. Brier Scoring captures some people’s intuition that perfect omniscience, or a subject with only binary beliefs, ones and zeros, who has all true beliefs and no false beliefs, is a kind of epistemic ideal. For those who hold that the aim of belief-revision is truth, Brier Scoring measures a subject’s closeness to the truth, capturing the intuition that the closer a method takes you to the truth, the better it is. Perhaps Calibration is a good measure of reliability, and Brier Scoring is a good measure of omniscience, where these are two distinct epistemic goods.

Much can be said about this debate. But rather than choose now between one or another measure, I will use both to compare the reliability of Stay the Coursers and Equal Weighters. As a matter of fact, the two measures are closely related, one subsuming the other [Murphy, 1973], but I will leave these issues for the end of the paper. It is interesting in its own right to investigate the comparative reliability of the two rules using both measures of reliability.

3 On the Calibration of Equal Weighters

Suppose that A and B are equally well-calibrated subjects and hence, epistemic peers. Suppose that C splits the difference in all cases in which A and B disagree, and otherwise has credences identical to A and B’s. A and B simply stay the course.⁸ We need only compare the calibration of A and

⁸Fitelson and Jehle [Fitelson and Jehle, 2009] prove that “splitting the difference” construed as straight averaging, is in fact incompatible with Bayesianism under certain minimal assumptions. In place of splitting the difference, Fitelson and Jehle formulate an alternative rule to EWR, “approximate” splitting the difference, which they take to be in the spirit of the Equal Weight View. For those who are Bayesians, in order to dodge

B with C to compare the reliability of STC to EWR. Is it true that C is necessarily, or even generally as well or better-calibrated than A and B? It turns out that C is more often less calibrated than A and B, and only equally or better calibrated in a small minority of special cases.

3.1 Comparative Calibration for Perfectly Calibrated Subjects

Let us begin with an examination of the limiting case in which A and B are perfectly-calibrated. In this circumstance, how well-calibrated is C? The first observation is that two *perfectly calibrated* subjects can disagree, and disagree in a lot of cases, while still being perfectly-calibrated. Figure 1 below illustrates this point with a specific example of two perfectly calibrated subjects A and B who disagree on precisely those propositions in which B takes to be 60% and 40% likely, and A takes to be 100% and 0% likely, where this turns out to involve 10 total propositions. Call this set of propositions S. Let $pr_A(-)$ be A's credence function, let $pr_A(P)$ be A's probability of P, and let S_A^n be the subset of S such that $\{P : pr_A(P) = n\}$. Assuming that A and B agree on everything else and are perfectly calibrated elsewhere, they are also perfectly calibrated with respect to the propositions in which they disagree, as Figure 1 shows.

the problems that Fitelson and Jehle pose for the incoherence of EWR with Bayesianism about partial belief, all of the propositions I discuss below in which A and B disagree must be assumed to be atomic, such as the propositions that it will rain next Monday, next Tuesday, next Thursday, and so forth, and comparative reliability must be construed as limited to the reliability of methods for forming partial-beliefs on atomic propositions.

Figure 1

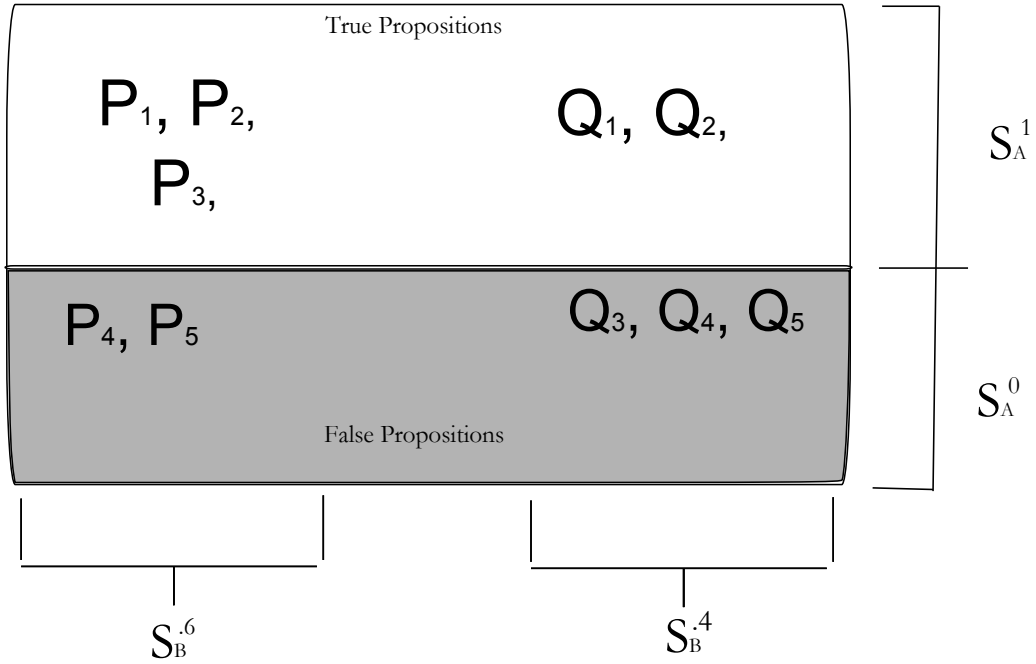


Table for Figure 1

A and B's Judgments	Ratio of True to Total Propositions	Error Rate
S_A^1	1	0
S_A^0	0	0
S_B^6	.6	0
S_B^4	.4	0

Simply counting the true and false propositions in such a diagram allow us to recover the comparative calibration rates between subjects. Let us generalize the case to show that for any set of propositions with at least one truth and one falsity, there are many different, perfectly calibrated subjects who disagree on at least one such proposition.

Where S is the set of all propositions in which A has some credence, so long as A 's credences constitute a function, $S_A^n \subseteq S$ for every n , and the set Υ of all subsets S_A^n is a partition of S . Notice that Υ is by definition a partition of S such that no two elements S_A^n and S_A^m have the same ratio

of true to total propositions. Any partition of S which has this property we will call a J-I Partition (“J-I” for “Judgment-Individuated”). Now A is perfectly calibrated just in case for all n, the ratio of true propositions to total propositions in $S_A^n = n$. For any S containing at least one true and at least one false proposition, there will be at least two J-I Partitions. For some J-I Partition $\Pi \neq \Upsilon$, each $S_X^n \in \Pi$ will have some ratio r of true propositions to total propositions. For all r, any agent B whose $pr_B(P) = r$ whenever $P \in S_B^n$ will be perfectly calibrated. Since $\Pi \neq \Upsilon$, A and B disagree on some propositions. Hence disagreement between A and B is consistent with the perfect-calibration of A and B. For any subject A where S is the set of all propositions in which A has a credence, the number of possible J-I Partitions of S equals the number of possible perfectly-calibrated credence functions which are exactly as opinionated as A, but which disagree with A on some propositions.

Given a certain set of propositions in which you have some credence, how many perfectly calibrated subjects can disagree with you? First let us see how many ways the world can be, consistent with my set of credences. Suppose I have credences in only three propositions $\{P_1, P_2, P_3\}$, one is true and two are false. In this case, there are three possible distributions of truth values consistent with this fact, (three possible ways to draw the true/false divide in a diagram). For any subject A, let $S = \{P_n : pr_A(P_n) = p, \text{ for some } 0 \leq p \leq 1\}$. For any number n, if the total number of propositions in S is n, then there are $n + 1$ possible ratios of true to total propositions in S, represented by the sequence $seq_n(S) = \langle 0/n, 1/n \dots, \frac{n-1}{n}, n/n \rangle$, where the m th element in $seq_n(S)$ is $\frac{m-1}{n}$. Let us call this the “ m th ratio” of S. Pascal’s numbers represent the possible truth-value combinations (states of the world) for the propositions in S which satisfies the ratio given by an arbitrary ratio in $seq_n(S)$ for all n. The n th row in the array represents S with n number of propositions, and the m th element in the n th row represents the number of possible truth-value combinations for S consistent with the fact that the ratio of true to total propositions in S is the m th ratio of true to total propositions in S.

Array 1: Pascal’s Numbers

0.	1									
1.	1	1								
2.	1	2	1							
3.	1	3	3	1						
4.	1	4	6	4	1					
5.	1	5	10	10	5	1				
6.	1	6	15	20	15	6	1			
7.	1	7	21	35	35	21	7	1		
8.	1	8	28	56	70	56	28	8	1	
9.	1	9	36	84	126	126	84	36	9	1
⋮										

Now to answer our original question; how many perfectly calibrated subjects can disagree with me? Recall that in our example, I have only three propositions in my credence space, $S = \{P_1, P_2, P_3\}$. Let's say that P_1 is the true one. To be perfectly calibrated, I can have 1) $pr(P_1) = 1$ and the rest have probability 0, or 2) $pr(P_1) = pr(P_2) = pr(P_3) = 1/3$, or 3) $pr(P_1) = pr(P_2) = 1/2$ and $pr(P_3) = 0$, or 4) $pr(P_1) = pr(P_3) = 1/2$ and $pr(P_2) = 0$. In this case, there are four J-I partitions of S corresponding to four distinct perfectly calibrated credence functions consistent with P_1 being true and 1 out of 3 true propositions in S. Parity of argument also gives the same number, 4, in all situations in which exactly one proposition is false, i.e., when 2 out of 3 are true.

The number of perfectly calibrated subjects who can disagree with me, i.e., the number of distinct J-I partitions of S, is therefore a function of the number of elements in S, and the ratio of true to total propositions in S. Array 2 below gives the number of perfectly calibrated J-I partitions for any given set of propositions S, where the m th element in the n th row of the array gives the number of distinct, perfectly calibrated J-I partitions for a set S with n members, with the m th ratio of true propositions in S, for some arbitrary $m - 1$ propositions designated as true, and the rest false.⁹

Array 2

⁹I designed a partitioning program to make the relevant computations for an arbitrary input of n propositions, m of which are true. Thanks to Dave Johannsen of Core Concepts for coding the program, which is available at <http://faculty.vassar.edu/balam/papers.htm>

0.	1									
1.	1	1								
2.	1	2	1							
3.	1	4	4	1						
4.	1	8	10	8	1					
5.	1	16	28	28	16	1				
6.	1	32	78	95	78	32	1			
7.	1	64	224	316	316	224	64	1		
8.	1	128	652	1058	1298	1058	652	128	1	
9.	1	256	1922	3769	5036	5036	3769	1922	256	1
⋮										

Together, multiplying the m th element in the n th row in Array 1 with the same in Array 2 gives the total possible number of perfectly calibrated subjects consistent with the fact that S contains n propositions and the m th ratio gives the ratio of true propositions in S.

Here is an example to cut through the abstraction. If a subject A has some credence in 8 total propositions, and 3 of them are true (a ratio of 3/8), then we look to the 4th element in the 8th row in Array 1 to see that there are 56 possible ways for the world to be consistent with such facts. On the assumption that, say, $P_1 - P_3$ are the true ones, and the rest false, we look to the 4th element in the 8th row of Array 2, showing that there are 1058 possible credence functions that would be perfectly calibrated, on such assumptions, but which disagree with each other on at least one proposition. Taken together, there are 59248 possible states of opinions that are potentially perfectly calibrated given the fact that 8 propositions have some probability, and 3/8 of them are true. It is quite apparent that the number of potentially perfectly calibrated credences that disagree with each other begins to skyrocket the more opinionated we get.

It is clear that many perfectly calibrated subjects can disagree. C, who employs the Equal Weight rule, will end up with a new set of opinions with regards to only those propositions about which A and B disagree. We are now in a position to show that EWR is not perfect-calibration-preserving whereas trivially, the Stay the Course rule must be.

To take the simplest case, assume that A and B are perfectly calibrated and disagree on one and only one proposition. Since $pr_C(P) = pr_A(P) = pr_B(P)$ whenever $pr_A(P) = pr_B(P)$, and $pr_C(P) = 1/2(pr_A(P) + pr_B(P))$ wherever $pr_A(P) \neq pr_B(P)$, if A and B disagree on one and only one case,

then for some P , $pr_A(P) \neq pr_B(P)$. Let $pr_A(P) = m$, and $pr_B(P) = p$. By the perfect calibration of A and B, S_A^m contains $100m\%$ true propositions and S_B^p contains $100p\%$ true propositions. By the definition of $pr_C(-)$, $pr_C(P) = 1/2(m + p)$. Now $S_C^m = S_A^m - \{P\}$, and $S_C^p = S_B^p - \{P\}$. Since S_A^m contains $100m\%$ true propositions, S_C^m does not contain $100m\%$ true propositions, and since S_B^p contains $100p\%$ true propositions, S_C^p does not contain $100p\%$ true propositions. So C is not perfectly calibrated.

Disagreement about only one case is not special. C's rule of splitting the difference also leads to imperfect calibration in cases where A and B disagree on many propositions, as in the case of Figure 1 where A and B disagree about 10. Figure 2 below illustrates how C, in splitting the difference between A and B as construed in Figure 1, is less calibrated than A and B.

Figure 2

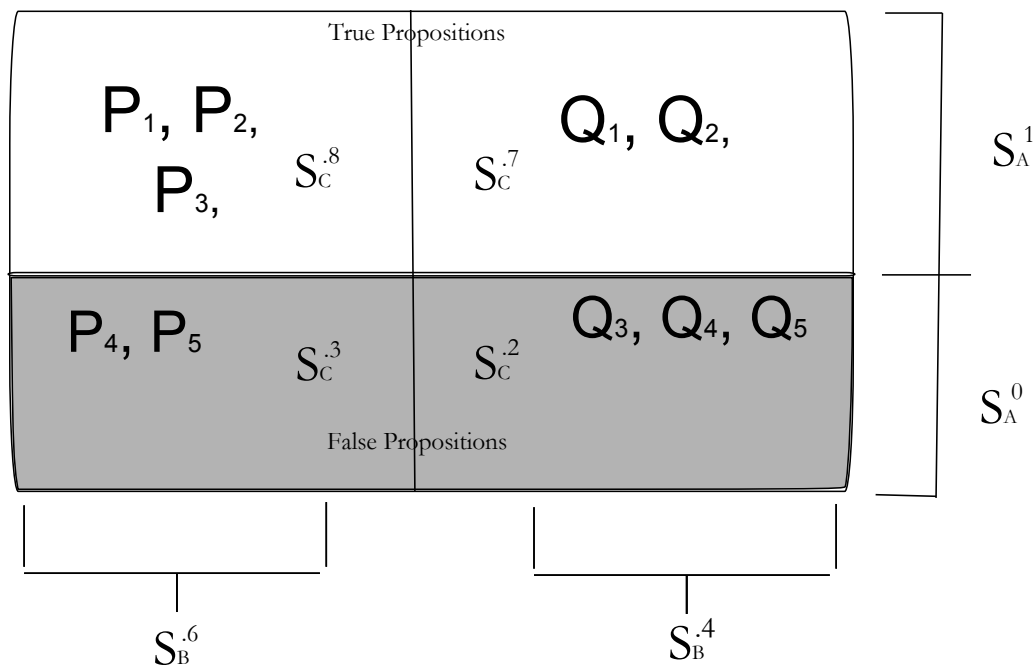


Table for Figure 2

C's Judgments	Ratio of True to Total Propositions	Error Rate
S_C^8	1	.2
S_C^3	0	.3
S_C^7	1	.3
S_C^2	0	.2
Average error rate for C		.25

Clearly, following EWR can lead you to be less well-calibrated. In fact, it might lead to such situations in all cases. I had a partitioning program compare all pairs of perfectly calibrated subjects up to $n = 30$ and $m = 1/2$. At 30 propositions, 15 of which are true, there are approximately 1.55×10^{21} perfectly calibrated subjects who disagree, and no two pairs of them are such that an application of EWR to all disagreements is calibration preserving. This is true for all $n < 30$ and all values of m .¹⁰

3.2 Comparative Calibration for Imperfectly Calibrated Subjects

I have been focusing on the case where A and B are perfectly calibrated because this is the simplest to illustrate. However, everything I have stated is also true when A and B are imperfectly, but equally-well calibrated. If A and B share average error rate m , the Equal Weight rule can lead C to have an error rate $n > m$, rendering C less calibrated than A and B. I will illustrate this point less formally. Figure 3 illustrates a situation in which A and B are equally imperfectly calibrated with error rate .2. Figure 4 shows C's degrees of belief after splitting the difference between A and B in all cases of disagreement. The table for Figure 4 shows C's error rates and average error rate.

¹⁰The program ran for 50 hours to obtain the answer for $n = 30$. Any higher n would probably be computationally unfeasible.

Figure 3

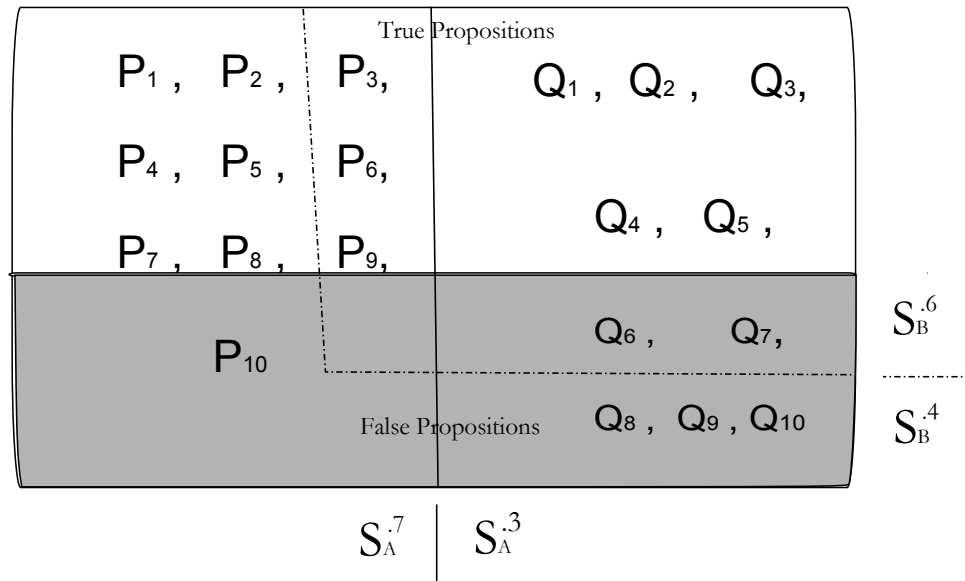


Table for Figure 3

A and B's Judgments	Ratio of True to Total Propositions	Error Rate
$S_A^{.7}$.9	.2
$S_A^{.3}$.5	.2
$S_B^{.6}$.8	.2
$S_B^{.4}$.6	.2
Average error rate for A		.2
Average error rate for B		.2

Figure 4

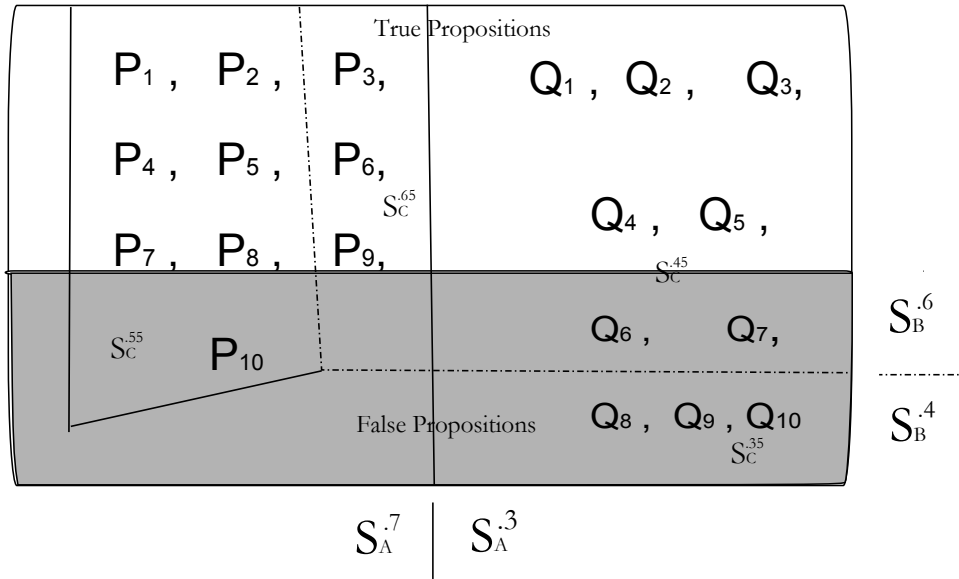


Table for Figure 4

C's Judgments	Ratio of True to Total Propositions	Error Rate
$S_C^{.65}$	1	.35
$S_C^{.55}$	6/7	.3
$S_C^{.45}$	5/7	.26
$S_C^{.35}$	0	.35
Average error rate for C		.315

In Figures 3 and 4, both A and B have a negative error rate, that is, they are to the same degree *underconfident* in all of the propositions in which they disagree. This is an inessential feature of the case. A simple series of calculations will show that C is less calibrated than A and B when A is overconfident whenever B is underconfident, vice versa, or if there is some mixture of over and underconfidence. Changing the numbers slightly will also show that C is less calibrated than A and B when A and B are to the same degree overconfident.

Figures 1, 2, 3, and 4 all indicate areas of total disagreement between A and B with respect to two sets of probabilistic judgments n and m . In other words, there is some n, m such that $pr_A(P) = n$ if and only if $pr_B(P) = m$. This is also an inessential feature of the case. Figures 5 and 6 exhibit a case without this feature for imperfectly but equally well-calibrated A and B and for less well-calibrated C. The table for Figure 6 shows that C has a higher averaged error-rate than A and B.

Figure 5

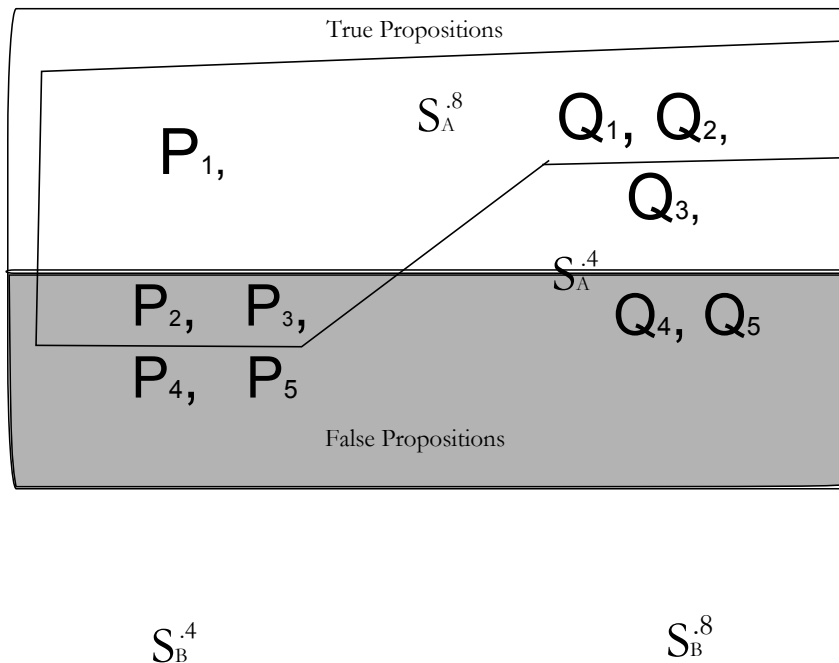


Table for Figure 5

A and B's Judgments	Ratio of True to Total Propositions	Error Rate
S_A^8	.6	.2
S_A^4	.2	.2
S_B^8	.6	.2
S_B^4	.2	.2
Average error rate for A		.2
Average error rate for B		.2

Figure 6

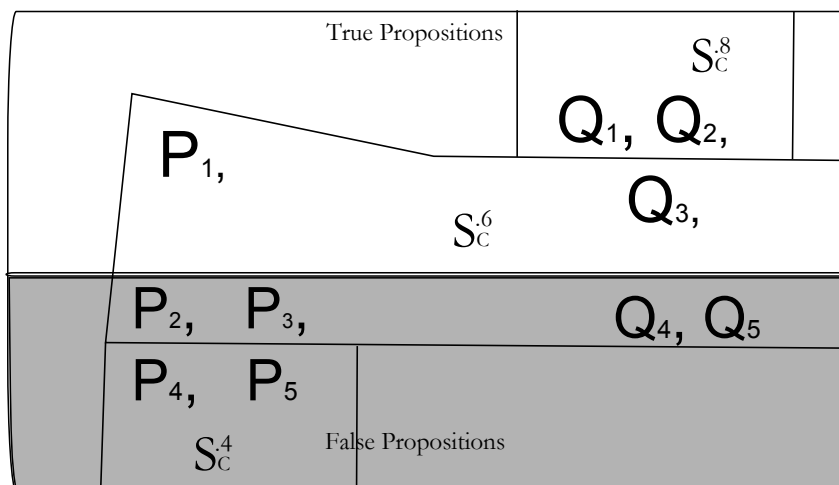


Table for Figure 6

C's Judgments	Ratio of True to Total Propositions	Error Rate
S_C^8	1	.2
S_C^6	1/3	.267
S_C^4	0	.4
Average error rate for C		.284

Similarly as in the case of Figures 3 and 4, the assumption in Figure 5 that A and B both have a positive error rate, i.e, that they are to the same degree overconfident in cases in which they disagree, is also inessential. Simple calculations will show that when A is overconfident whenever B is underconfident, vice versa, or if both are underconfident whenever they disagree, C will still be less well-calibrated than A and B. The particular numbers chosen make for ease of presentation, but they are arbitrary. From this, we can conclude that generally, where imperfectly calibrated A and B have the same average error rate, regardless of whether any particular one

is positive or negative, regardless of whether A or B disagree systematically with respect to some particular judgment or not, C will have a higher average error rate and hence be less well-calibrated than A and B.

3.3 The Equal Weight Rule Does Better in Some Circumstances

There are certain kinds of situation in which the Equal Weight rule preserves or increases upon the calibration rates of imperfectly, but equally well-calibrated subjects. Consider A and B who believe to degree one precisely what the other believes to degree zero and vice versa, and they agree everywhere else. Suppose that, with respect to the set of all such propositions disagreed about, the ratio of true to total propositions is $1/2$.¹¹ Clearly, for all propositions P disagreed about, $pr_A(P) + pr_B(P) = 1$, so $pr_C(P) = .5$, so C is perfectly calibrated while A and B have an error rate of $.5$.¹²

This kind of case illustrates the condition under which the use of EWR can improve on calibration. An application of the Equal Weight rule is *fully merging* of a set of disagreements just in case there is a σ such that $\sigma = pr_A(P) + pr_B(P)$ for all P such that $pr_A(P) \neq pr_B(P)$. A fully merging application of EWR has the effect of making C have one degree of belief for all propositions disagreed about between A and B. Let the *base rate* be the ratio of true to total propositions among all of the propositions in which A and B disagree. C can improve upon calibration rates if it so happens that the result of averaging A and B's disagreements is assigning the same probability over all of the disagreements that is close to the base rate. This can happen, not only of fully merging applications of EWR, but of *almost fully merging* cases where many propositions, and an equal number of true and false ones, in the class of disagreements, are such that A and B's credence sum to exactly twice the base rate.¹³ Under these conditions, an application

¹¹This is one possible interpretation of Christensen's case involving two subjects who disagree about how much one of them owes on a restaurant meal, where each subject is wrong 50% of the time with respect to such beliefs. Christensen's case seems to many to be the strongest intuitive case for the Equal Weight rule.

¹²The probability given for full belief does not need to be 1 for this case to work, I am simply using it for illustrative purposes.

¹³The ability to improve on calibration rates in some situations is not unique to EWR. Any rule under some circumstances might improve on calibration under specific enough conditions. For instance, the "1/3 rule" of believing a third of the sum of A and B's

of EWR can preserve or improve upon average calibration rates.

How prevalent are such cases? There are three variables involved in computing the answer. We first obtain the number n of propositions in the space in which A and B have credences but some disagreements, a number m of true propositions in that space, and the average calibration rate r of A and B. I designed another partitioning program to solve for arbitrary n , m , and r .¹⁴ To illustrate, for $n = 10$ propositions, the following chart gives the percentages of cases. On the top row are possible values for m , or total true propositions (the number stops at 5 because the numbers are symmetric past 5). On the side column are possible values for r (or average calibration rates) up to $1/2$. To check the percentage for 3 out of 10 true propositions with a calibration rate of .2, go to the column under 3, down to the row for $1/5$, and we see that EWR does better than STC in 5% of such cases.

Chart 1:
Percentages of Cases where EWR Better than STC for $n = 10$

	1	2	3	4	5
1/10	18%	1%	.4%	.2%	.2%
1/9	16%	1%	.6%	.4%	.3%
1/8	37%	3%	1%	.7%	.6%
1/7	29%	4%	1%	.9%	1%
1/6	47%	9%	3%	2%	3%
1/5	37%	12%	5%	4%	3%
1/4	50%	21%	13%	13%	13%
1/3	40%	25%	22%	27%	23%
1/2	43%	33%	43%	44%	55%

Finally, to give some numbers as to the prevalence of cases, for 1 out of 10 true propositions, for average calibration rate .1, there are about 500,000 pairs of possible cases 18% of which the EWR does better than Staying the Course. That number, 500,000, drops steadily to about 100,000 possible pairs as the calibration rate goes to .5. As the number of true propositions increases to 5, the number of pairs increases to 1×10^9 for calibration rate .1, and steadily decreases to 1.4×10^8 as the calibration rate goes to .5. In

credences will be perfectly calibrated when it fully merges a set of disagreements where A and B's credences all sum to three times the base rate.

¹⁴This program is also available at <http://faculty.vassar.edu/balam/papers.htm>. Again, thanks to Dave Johannsen at Core Concepts for coding the program.

total, the table represents approximately 1.4×10^{10} total pairs of equally calibrated subjects, and EWR preserves or improves upon calibration rates in approximately 15% of these cases.¹⁵ Generally, STC does better than EWR in matters of calibration.

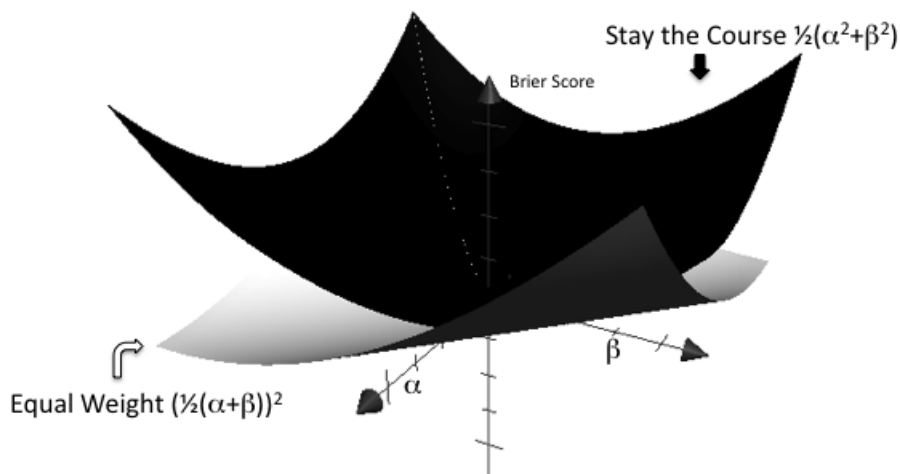
4 Comparative Brier Scoring

Now that we have looked at comparative calibration for EWR and STC, we must now look at comparative closeness to truth, or Brier Scoring. Unlike in the case of Calibration, it is clearly impossible for two agents to have perfect Brier Scores (scores of 0), and also disagree. This is because perfect Brier Scoring requires having degree 1 (and only 1) in every truth in which you have some judgment, and 0 (and only 0) for every such falsehood. Assuming that there are no inconsistencies in the world, two subjects with such beliefs will agree on everything. This fact about Brier Scoring really does capture the intuition that omniscience is a kind of epistemic ideal. Nonetheless, it is possible for two subjects to disagree while having the same Brier Scores, so long as they are imperfectly reliable. One such situation is when P is false, A and B have different, non-zero credences in P, where A is distance α from the truth with respect to P, and B is distance β from the truth with respect to P. Simultaneously, there is some other proposition Q, whereas B is distance α from the truth with respect to Q, and A is distance β from the truth with respect to Q. Assuming that A and B agree on every other proposition, A and B will disagree about P and Q, but will have the same Brier Score, and thus, will be epistemic peers. For simplicity, if we assume that A and B have Brier Score 0 for every other proposition except P and Q, then their Brier Scores will be $1/2(\alpha^2 + \beta^2)$. Equal reliability is consistent with disagreement on two cases.

In such a situation, it is easy to prove that C will in fact have a lower Brier Score than both A and B, and thus, be closer to the truth. For $pr_C(P) = 1/2(\alpha + \beta)$, which will be $1/2(\alpha + \beta)$ distant from the truth with respect to P, since P is false. Therefore, C will be $1/2(\alpha + \beta)$ distant from the truth with respect to Q. C's Brier score will therefore be $(1/2(\alpha + \beta))^2$. For any α, β between 0 and 1, $1/2(\alpha^2 + \beta^2) > (1/2(\alpha + \beta))^2$, so C necessarily has a lower Brier Score than A and B, assuming that $\alpha \neq \beta$, which is true since

¹⁵ $n = 10$ is not a special case. I invite the reader to run the partitioning program to check the numbers for $n > 10$

A and B disagree. The following figure, with Brier Scores on the z -axis and possible values of α and β on the x and y axes, illustrates this fact.



The case in which A and B disagree on two propositions is not special. On the assumption that A and B have the same Brier Score, and A and B disagree on n cases, C will have a better Brier Score than A and B.¹⁶ Here is the proof of the general case.

Proof. Assumption For all P_i

$$\frac{1}{n} \sum_{i=1}^n (pr_A(P_i) - T(P_i))^2 = \frac{1}{n} \sum_{i=1}^n (pr_B(P_i) - T(P_i))^2$$

¹⁶It is interesting to note that for both Calibration and Brier Scoring, epistemic peerhood is consistent with disagreement on arbitrarily large numbers of propositions. The inference from disagreement to epistemic inferiority or superiority is deductively invalid assuming either measure.

(Epistemic Peerhood Assumption)

Definition Let $\alpha_i = pr_A(P_i) - T(P_i)$, $\beta_i = pr_B(P_i) - T(P_i)$, and let $\delta_i = \beta_i - \alpha_i$.

Thus, the Epistemic Peerhood Assumption is

$$\frac{1}{n} \sum_{i=1}^n (\alpha_i)^2 = \frac{1}{n} \sum_{i=1}^n (\beta_i)^2$$

and implies that

$$\frac{1}{n} \sum_{i=1}^n \beta_i^2 - \alpha_i^2 = \frac{1}{n} \sum_{i=1}^n (\beta_i + \alpha_i)(\beta_i - \alpha_i) = 0.$$

Now $\beta_i + \alpha_i = (\beta_i - \alpha_i) + 2\alpha_i = 2\alpha_i + \delta_i$. Thus, by substitution, $\frac{1}{n} \sum_i (\beta_i + \alpha_i)(\beta_i - \alpha_i) = \frac{1}{n} \sum_i (2\alpha_i + \delta_i)\delta_i = \frac{1}{n} \sum_i 2\alpha_i\delta_i + \delta_i^2 = 0$. Call this *Important Fact*.

The Brier Score of C=

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2}(\beta_i + \alpha_i)\right)^2 \\ &= \frac{1}{n} \sum_i \left(\frac{1}{2}(2\alpha_i + \delta_i)\right)^2 = \frac{1}{n} \sum_i \left(\alpha_i + \frac{\delta_i}{2}\right)^2 = \frac{1}{n} \sum_i \alpha_i^2 + \alpha_i\delta_i + \frac{\delta_i^2}{4}. \text{ Now } \alpha_i\delta_i + \frac{\delta_i^2}{4} = \frac{2\alpha_i\delta_i}{2} + \frac{\delta_i^2}{2} - \frac{\delta_i^2}{4}. \text{ Therefore } \frac{1}{n} \sum_i \alpha_i^2 + \alpha_i\delta_i + \frac{\delta_i^2}{4} = \frac{1}{n} \sum_i \alpha_i^2 + \frac{2\alpha_i + \delta_i}{2} - \frac{\delta_i^2}{4} \\ &= \frac{1}{n} \sum_i \alpha_i^2 + \frac{1}{2n} \sum_i 2\alpha_i\delta_i + \delta_i^2 - \frac{1}{4n} \sum_i \delta_i^2. \end{aligned}$$

By the Important Fact, the middle term = 0, so Brier Score of C=

$$\frac{1}{n} \sum_i \alpha_i^2 - \frac{1}{4n} \sum_i \delta_i^2.$$

Since Brier Score of A and B = $\frac{1}{n} \sum_i \alpha_i$, Brier Score of C \leq Brier Score of A and B. If A and B disagree on anything, $\frac{1}{4n} \sum_i \delta_i^2 > 0$, so Brier of C $<$ Brier Score of A and B. \square

It is thus provable that if we measure reliability in terms of Brier Scoring, or the average squared distance to the truth, then so long as A and B are equally, but imperfectly reliable, and disagree on some propositions, then C, who splits the difference in all cases of disagreement will have a better Brier Score than A and B. Staying the Course may be calibration-preserving, and Equal Weight not, but Staying the Course is merely Brier-Score preserving, while the Equal Weight View is Brier-Score improving in a large majority of cases.¹⁷

5 Normative Import

Using Calibration as the definition of reliability for partial-belief forming methods, the Equal Weight rule renders a subject generally less reliable than she would be were she to Stay the Course. There are some situations in which the Equal Weight rule increases reliability. However, in general, if calibration is the measure of epistemic peerhood and reliability, then sticking to all and only the methods we use to make probabilistic judgements about whether P, whatever they are, makes us generally more reliable than first using them and then adding the method of splitting the difference with our disagreeing peers.

On the other hand, using Brier Scoring as the measure of epistemic reliability, when two subjects are equally reliable, the Equal Weight rule necessarily preserves or increases reliability whereas the Stay the Course rule only preserves it. There is no situation in which subjects start out equally close to the truth and Staying the Course does better than the Equal Weight rule in terms of moving degrees of belief closer to the truth.

¹⁷It may occur to the discerning reader to wonder whether it is essential to the proof in favor of EWR that Brier Scoring makes use of the squared deviation from the truth as opposed to the absolute deviation. It is. The use of the absolute deviation renders EWR and STC equally reliable in all cases.

Given these formal, factual results about reliability, what normative conclusions should we draw about the rationality of these inference-rules? These results can be interpreted as reducing the problem of belief-revision in light of peer disagreement precisely to the issue of whether we ought to use a measure like Calibration to define epistemic peerhood and partial-belief reliability, or a measure like Brier Scoring. The reduction of a difficult normative question into a different normative question should itself be considered progress in this dispute. We should now be concerned with the questions of which measure is the better one to adopt, if either. Moreover, how should we argue about this?

I give my answers to many of these questions more fully in another paper [Lam, forthcoming]. For now, I would like to suggest that we distinguish between two different considerations relevant to answering these questions. One way of arguing for a belief-revision rule is to argue that it is the optimal rule to use when aiming for a certain good, and that this good is *the* aim of belief-revision. Someone might, for instance, advocate that belief-revision ought to aim at the truth, so that when other things like evidence or long-term practical interests point away from the truth, epistemic rationality dictates that we adopt opinions that are closest to the truth. On the other hand, one might advocate that belief-revision aim only at best reflecting the evidence, so that when the evidence takes you farther from the truth, you ought to revise your beliefs in accordance with evidence rather than truth. All such views presuppose that there is such a thing as *the aim* of belief-revision, understood as the central epistemic good that generates the norms for belief-change.

On the other hand, we might simply be looking for a measure that is the best formally precise rational reconstruction of the informal and imprecise property we are attributing to our fellow epistemic agents when we consider ourselves to be in a very specific epistemic predicament, the predicament of epistemic peer disagreement. The parties to the debate about the epistemology of disagreement seem to consider this predicament to be prevalent and common, and one measure might better capture this prevalence and commonality than another. Considerations of the aim of belief-revision, and considerations regarding rational reconstruction, might lead to competing verdicts on the proper measures of epistemic reliability. This in turn might lead to competing verdicts on the status of EWR and STC.

Brier Scoring is the method most used in assessing accuracy in weather forecasting, and better captures our intuitions than calibration about how

good we consider certain epistemic agents in situations of decisions under uncertainty. For instance, if we consider Figure 1 above, it is clear to many people that one of the subjects, namely A, is far better epistemically than subject B. They are both perfectly calibrated, and thus, the sense in which they A is better is not captured by Calibration. A has a perfect Brier Score, and B does not. So Brier Scoring captures the sense in which A is better than B.

One interesting thing to note about the Brier Score is that it is well understood in meteorology since the work of Murphy [Murphy, 1973] that a subject’s Brier Score is an aggregate of measures one of which is a subject’s degree of calibration. Put another way, Brier Scoring is an aggregate of many features of epistemic value one component of which is calibration.¹⁸ As such, Brier Scoring is also a far more discriminating measure of epistemic value; subjects need to have far more features in common to be epistemic peers in terms of Brier Scoring than in terms of Calibration. We have already seen this: many subjects can be equally calibrated, even perfectly calibrated, and disagree, whereas this is not as much the case with Brier Scoring. This feature of Brier Scoring makes the measure far better at capturing which agents are better than others, but it also renders epistemic peerhood far more difficult

¹⁸The other two features are what Murphy calls “resolution”, which for our purposes is a measure of the frequency of very high and very low (close to zero and one) probabilistic judgments a subject makes, and “uncertainty” which is a measure of the relative frequency of truths in a domain. Formally, let us number each of the k subsets S_A^n of S_A from 1 to k , while labeling S_A as σ . Let $Freq_\Phi =$ the ratio of true to total propositions in the set labelled Φ , let $pr_k = n$ where n is A’s probability of the propositions in the set labelled k , let $m_k =$ the number of propositions in the set labelled k . Then, if we measure Calibration as a weighted average of squared difference rather than a straight average of the absolute difference;

$$\frac{1}{q} \sum_1^k m_k (pr_k - Freq_k)^2$$

where q is the total number of propositions in S_A , we can measure Resolution as;

$$\frac{1}{q} \sum_1^k m_k (Freq_k - Freq_\sigma)^2$$

and Uncertainty as;

$$Freq_\sigma(1 - Freq_\sigma)$$

Murphy showed that the Brier Score of A turns out to be Calibration–Resolution+Uncertainty, defined in this way.

to come by. The conditions under which someone has the same Brier Score as you are very restrictive. They are far more restrictive than the conditions under which someone is just as calibrated as you.

On the other hand, one might think that epistemic peerhood is a feature of subjects that can be quite easy to come by. In fact, this could be an underlying motivation for taking the issue of disagreement between epistemic peers to be epistemologically significant; it is significant because it is so widespread! On the assumption that peerhood is prevalent, we would expect any formally precise, rational reconstruction of our ordinary attributions of epistemic peerhood to rationalize our ordinary, rather widespread attributions of epistemic peerhood. The fact that we are generally warranted in attributing, or assuming, that a peer is just as reliable as us in certain domains suggests that some less discriminating property of reliability is what we are attributing. Here, equal Calibration is a better candidate than equal Brier Score.

Choosing a precise reliability measure to settle the dispute over EWR and STC will involve a tradeoff between our intuitions of comparative epistemic excellence, and our intuitions of warranted attributions of epistemic peerhood. Our choice of a precise definition of reliability depends on how many features we are warranted in believing we share with another epistemic agent in a situation in which we are assessing their status as an epistemic peer. When we are warranted only in believing of an epistemic agent that they are a peer in some less discriminating sense, then something closer to Calibration is the better measure, and Staying the Course is the generally better rule. When we are warranted in believing that an agent is a peer in some more discriminating sense, then something closer to Brier Scoring is the better measure, and we ought to employ the Equal Weight rule. In a scientific or philosophical domain, where standards of precision are far higher than in ordinary life, peerhood appears to be harder to come by, suggesting that the notion of reliability is far more discriminating. In ordinary life, peerhood seems widely attributable, suggesting that the notion of reliability is far less discriminating. Disputants in the debate over the epistemology of disagreement may be conceiving of epistemic peerhood in different ways, one as a quite inclusive club arising from a less discriminating measure of reliability, the other as a quite exclusive club arising from a quite discriminating measure of reliability. If this is true, then both sides may have latched onto an element of truth.

At the end of the day, one might end up advocating a pluralistic view

regarding both of these issues. Perhaps there are many aims of belief-revision, or multiple epistemic values none of which take priority over any other in determining how we ought to change our beliefs. There may also be no univocal property of epistemic peerhood, but a range of different ways one can be a peer. If this were true, there would be no normatively homogenous class of epistemic predicaments worthy of a single category, and thus no single rule necessarily required in all cases. The formal results are consistent with such a position, and might end up showing that that being in different predicaments of “known peer disagreement” warrant different responses.

References

- [Alston, 1995] Alston, W. (1995). How to think about reliability. *Philosophical Topics*, pages 1–29.
- [Brier, 1950] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.
- [Christensen, 2005] Christensen, D. (2005). *Putting Logic in its Place*. Oxford UP.
- [Christensen, 2007] Christensen, D. (2007). Epistemology of disagreement: the good news. *Philosophical Review*, (116).
- [Elga, 2007] Elga, A. (2007). Reflection and disagreement. *Nous*, 41(3):478–502.
- [Feldman, 2006] Feldman, R. (2006). Epistemological puzzles about disagreement. In Hetherington, S., editor, *Epistemology Futures*. Oxford UP.
- [Fitelson and Jehle, 2009] Fitelson, B. and Jehle, D. (2009). ‘what is the ‘equal weight view?’’. *Episteme*, 6(3):280–293.
- [Joyce, 2008] Joyce, J. (2008). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In Huber, F. and Schmidt-Petri, C., editors, *Degrees of Belief*. Springer Netherlands.
- [Kaplan, 1996] Kaplan, M. (1996). *Decision Theory as Philosophy*. Cambridge UP.

- [Kelly, 2005] Kelly, T. (2005). The epistemic significance of disagreement. In Hawthorne, J. and Gendler-Szabo, T., editors, *Oxford Studies in Epistemology: Volume 1*, pages 167–196. Oxford UP.
- [Kelly, 2009] Kelly, T. (2009). Peer disagreement and higher-order evidence. In Feldman, R. and Warfield, F., editors, *Disagreement*. Oxford UP.
- [Lam, forthcoming] Lam, B. (forthcoming). One the Rationality of Belief Invariance in Light of Peer Disagreement. *The Philosophical Review*
- [Lichtenstein et al., 1982] Lichtenstein, S., Fischhoff, B., and Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In Kahnemann, D. and Tversky, A., editors, *Judgments under Uncertainty: Heuristics and Biases*. Cambridge UP.
- [Murphy, 1973] Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12:595–600.