5

A Closer Look at Scoring

We have encountered the concept of scoring in relation to measuring inaccuracy, a couple of times, but have skipped formal details so far. After making the concept precise, this chapter starts with an observation that may at first seem only tangentially related to our current discussion but that actually raises serious doubts about two claims central to the inaccuracy-minimization argument, to wit, that this argument is of a strictly epistemic nature, and that Bayes's rule minimizes expected next-step inaccuracy.

5.1 Standard Scoring Rules and Truthlikeness

Scoring can easily appear a technical and purely formal concept, but the underlying idea has been familiar since elementary school. Indeed, those of us who are teachers engage in scoring quite frequently when we grade our students' exams or essays. In general, grading is not particularly hard. Suppose that a teacher gives an exam containing ten questions, each of which has a straightforward yes / no answer and each of which contributes equally to the final score. Then if a student gets nine questions right, the reasonable grade for this student appears to be 90 percent or whatever that translates to in the teacher's grading system (e.g., an A in the US public school system, or a 3.6 in a 4.0-based system).

Naturally, when questions do not have straightforward yes / no answers, grading can be more complicated. A student may get none of the questions completely right and yet show a good understanding of the relevant subject matter. For example, a student may turn in an exam that gives clear evidence of her mathematical acumen, on one hand, and of a certain inattentiveness in carrying out simple algebraic operations, on the other. Most mathematics

teachers would not want to fail such a student but would rather encourage her to address her sloppiness. What mark to give in such a case may not be entirely straightforward and may be to some degree subjective.

Grading can be more complicated still. Think of grading essays, for which there are usually no more-or-less right answers. Grading also becomes more complicated when we are dealing with questions that ask for a probability estimate of some future event. Such questions may not occur so frequently on a school exam (though see Bickel, 2010), but they are the bread and butter for a variety of professionals, ranging from financial analyst to physician, from football coach to engineer, and from foreign policy advisor to weather forecaster. Suppose that one weather forecaster predicts rain for tomorrow with a probability of .25, whereas a second predicts rain for tomorrow with a probability of .75. If tomorrow stays dry, neither forecaster can be said to have been wrong, for neither predicted rain with certainty. Yet from a pretheoretic perspective, it appears that the former forecaster did a better job than the latter; that forecaster was, we would like to say, the more accurate of the two.

How to turn this intuition into a difference in grades (or scores) for the two forecasters—supposing that our only concern is their predictions of rain tomorrow—might also be deemed a partly subjective matter. But researchers from various quarters think otherwise and have sought to answer the foregoing question by providing fully objective formal rules. More specifically, proposals for how to assess the forecasters, and for how to assess the quality of probability estimates generally, have come in the form of scoring rules. So far, I have mentioned only the Brier score, which was the historically first, dating to the 1950s, but by now a bewildering variety of scoring rules exists (Brier, 1950; Rosenkrantz, 1981, ch. 2; Cooke, 1991, chs. 8 and 9). Given that these rules may lead to different qualitative verdicts (e.g., on which weather forecaster did best), and given that, as mentioned, this kind of scoring is supposed to be objective, the question has arisen of which of the many available scoring rules today we should rely on in practice.

Theorists have almost invariably tried to answer this question by arguing that this or that scoring rule uniquely satisfies certain standards of "goodness" (Winkler & Murphy, 1968). There is debate about this issue because there is no generally shared conception of goodness in the relevant sense and hence no agreement on which criteria to impose on scoring rules. Nonetheless, it is fair to say that most theorists are, or have been, in favor of one of two scoring

rules, either the Brier rule[1] (e.g., Brier, 1950; Rosenkrantz, 1981; Joyce, 1998; Selten, 1998; Greaves & Wallace, 2006; Leitgeb & Pettigrew, 2010a) or the logarithmic (or "log") rule (e.g., Good, 1952; Bernardo, 1979; Bernardo & Smith, 2000; Bickel, 2007, 2010; Levinstein, 2012; McCutcheon, 2019).

To state these rules formally, let $\{H_i\}_{i=1}^{n}$ be a *hypothesis partition*, that is, a set of mutually exclusive and collectively exhaustive hypotheses, and let δ_{ij} (for $i,j \in \{1,\ldots,n\}$) be the Kronecker delta, which equals 1 if $i = j$ and equals 0 otherwise. Suppose that a person assigns probabilities $\mathbf{p} = (p_1,\ldots,p_n)$ to the elements of $\{H_i\}_{i=1}^{n}$, with p_i her probability that H_i is true. Then, assuming that the true hypothesis is H_j, the Brier score for this person equals $\mathcal{B}_j(\mathbf{p}) := 1/n \sum_{i=1}^{n}(\delta_{ij} - p_i)^2$, whereas her log score equals $\mathcal{L}_j(\mathbf{p}) := -\ln(p_j)$.[2] As previously mentioned, we tend to conceive of scoring rules as assigning *penalties*, as measuring how *bad* a probability distribution is. But this is just a convention; if we wish, we can take the negative of any given rule's scores and think of them as *rewards*, as measuring how *good* a probability distribution is. Note that on the conventional conception of scores, lower scores are better; on the alternative, lower scores would be worse.

Given a scoring rule $\mathcal{S}$ and a probability distribution $\mathbf{p}$ on a hypothesis partition, the $\mathbf{p}$-*expectation* of the $\mathcal{S}$-score (or $\mathbf{p}$-expected $\mathcal{S}$-score, for short) for a second probability distribution $\mathbf{p}^*$ on the same partition equals $\mathbb{E}_{\mathbf{p}}[\mathcal{S}(\mathbf{p}^*)] := \sum_{i=1}^{n} p_i \mathcal{S}_i(\mathbf{p}^*)$. A scoring rule $\mathcal{S}$ is said to be *proper* if and only if for all $\mathbf{p}$, $\arg\min_{\mathbf{p}^*} \mathbb{E}_{\mathbf{p}}[\mathcal{S}(\mathbf{p}^*)] = \mathbf{p}$, and *strictly proper* if and only if each of these minima is unique. Scoring rules were originally proposed for eliciting probabilities (Cooke, 1991, p. 121), and when used for that purpose, they should not give a person an incentive to announce probabilities that she does not actually hold. This is why many theorists regard strict propriety as an important requirement for scoring rules.

Both the Brier and the log score are known to be strictly proper. They also both achieve their minimum of 0 when probability 1 is assigned to the truth, which is generally regarded as another important desideratum. Nevertheless,

1. Or the quadratic scoring rule, which is a generalization of the Brier score; see the subsequent discussion.

2. With regard to the Brier score, note that as long as one makes comparisons given a fixed number of hypotheses, division by that number is not strictly necessary and should be considered optional. It thus appears unproblematic to equate, for most practical purposes, the Brier score with what Rosenkrantz (1992, ch. 2) calls the "least squares" score (which is, for instance, also what Leitgeb & Pettigrew, 2010a, and Pettigrew, 2016, do).

Table 5.1: Probability assignments to hypotheses A, B, and C.

	David	Emma	Frank
Alice wins (A)	.1	.5	.3
Bob wins (B)	.5	.1	.3
Charlotte wins (C)	.4	.4	.4

the following example brings out that the rules can lead to very different verdicts.

Suppose that there will soon be an election for a new president of your university. There are three candidates in the running: Alice, Bob, and Charlotte. Your colleagues David, Emma, and Frank hold different views on which of the candidates is most likely to become the new president. Specifically, their relevant probabilities are as given in table 5.1. Suppose that Charlotte wins the election. Then David and Emma have the same Brier score: $(.1^2+.5^2+.6^2)/3 \approx .21$; Frank's Brier score is lower and thus better: $(.3^2 + .3^2 + .6^2)/3 = .18$. Nonetheless, all of them have the same log score, to wit, $-\ln(.4) \approx 0.92$.

That David and Emma do equally well on either scoring rule seems as it should be: David and Emma both assign a probability of .1 to one of the false hypotheses and a probability of .5 to the other, and it is difficult to see (at least given the information provided here) how it might matter that they do not assign these probabilities to the same hypotheses.

That Frank's Brier score is lower than David's and Emma's illustrates the general fact that, for any hypothesis partition, given a particular probability assigned to the true hypothesis, one minimizes one's Brier score by assigning equal probabilities to the remaining hypotheses. How reasonable is this?

According to advocates of the log rule, this is not reasonable at all. Why should we care—they ask—what probabilities a person assigns to any hypothesis other than the truth? In their view, given that David's, Emma's, and Frank's probabilities for C are the same, they should be assigned the same score (Winkler, 1969; Bernardo & Smith, 2000, p. 72; Bickel, 2010, pp. 347–348). The log rule is in fact known to be the only strictly proper scoring rule that guarantees this outcome; it is, in more official terminology, the only strictly proper scoring rule that is *local* (McCarthy, 1956).

In the debate about scoring, various other intuitions in favor of or against either of these rules have been called upon. For instance, Eric Bickel (2010,

p. 348) notes that the log rule, but not the Brier rule, ensures that *higher* probability assignments to the truth will result in *lower* penalties.[3] He regards this as a compelling reason to prefer the log rule over the Brier rule. By contrast, Reinhard Selten (1998, pp. 49–50) prefers the Brier rule because he thinks that in some situations the log rule is more sensitive to small differences between probability assignments than is warranted by intuition and in other situations, not sensitive enough.[4]

That different authors have given different weights to intuitions regarding scoring has partly to do with the fact that they have focused on different examples, and the problem is that the features of the examples that have been used to make one or the other rule look appealing are not always generalizable.

To see how different examples may steer our intuitions in different directions, note that in our own example there is no sense in which either of the false hypotheses (whether A or B) is closer to the truth than the other, again given the information provided. That does not make this hypothesis partition special. The designated feature is not altogether general, however, and by assuming otherwise one might be implicitly favoring some scoring rules over others. More generally put, our hypotheses pertain to a so-called discrete random variable whose possible outcomes lack an intrinsic ordering, making it a *nominal* variable. But discrete random variables also come in another variety: next to nominal variables there are also *ordinal* variables, whose possible outcomes can be naturally ordered. Furthermore, sometimes our hypotheses concern *continuous* variables, which as the name suggests, can be measured on a continuum (for example, temperature or weight). These distinctions are directly relevant to scoring.

3. That the Brier rule cannot guarantee this is a direct consequence of the general fact mentioned two paragraphs back.

4. Selten instead prefers the Brier rule, mainly because, as he proves, it is the only scoring rule (up to positive linear transformations) that satisfies each of what he considers to be four important desiderata for such rules, which Selten presents as axioms. According to the first axiom, the ordering of the hypotheses should not influence the score. According to the second, the score should not be affected by the introduction of an additional hypothesis that receives zero probability. The third axiom is the requirement of strict propriety. Finally, the fourth axiom concerns a type of situation that we do not consider, namely, when a probability assignment is scored in light of another probability assignment rather than in light of the truth of one hypothesis; the axiom requires that, in this situation, the score should be the same regardless of which probability assignment is considered to be the "true" one.

Suppose that David, Emma, and Frank were wondering how well a student of theirs is going to do on an exam, and the probabilities given in table 5.1 are the probabilities that the student will receive an A, a B, or a C, possibilities that we may take to be described by hypotheses A, B, and C, respectively. These hypotheses now refer to possible outcomes of an ordinal variable: it is better to get an A than a B, which in turn is better than a C. From this, there plausibly emanate relations of truthlikeness among our hypotheses. If B turns out true, then the other hypotheses would seem equally far from the truth. But if either A or C turns out true, then, although false, B would be closer to the truth than whichever is the other false hypothesis.

To illustrate the connection with scoring, suppose that C is true. Would we still want to agree with the Brier score that David and Emma do equally well, given their probabilities for the hypotheses at issue, or even agree with the log score that all three colleagues do equally well? Although the three colleagues assign the same probability to the truth, we are tempted to say that David still is closer to the mark, given that he assigns a higher probability than the others to the false hypothesis that is closest to the truth (viz., B) and a lower probability than the others to the false hypothesis that is most distant from the truth (A).

An entire program in the philosophy of science is devoted to making the notion of truthlikeness (or "verisimilitude") formally precise, and numerous measures of truthlikeness exist.[5] Here, we will not commit ourselves to any particular such measure and just note that all measures of truthlikeness currently advocated in the literature will do for the purposes of this chapter.

The idea that truthlikeness may matter to scoring has been mentioned in the literature, but usually only to be set aside as a problem that can easily be dealt with by using the so-called quadratic scoring rule (also known as "weighted least squares metric"). If H_j is the true element of hypothesis partition $H = \{H_i\}_{i=1}^{n}$, this rule assigns a penalty of $\mathcal{Q}_j(\mathbf{p}) := \sum_{i=1}^{n} w_i(\delta_{ij} - p_i)^2$ to someone whose probabilities for the elements of H are given by $\mathbf{p}$. The only general constraints on the weights w_i are that $w_i > 0$ for all i and that $\sum_{i=1}^{n} w_i = 1$, so that we actually have a *schema* here, with the Brier score as the special case in which all hypotheses are weighted equally.

5. See, for instance, Niiniluoto (1987, 1998, 1999b, 2020), Schurz (1987, 1991, 2011, 2014), Kuipers (2000, 2001, 2014, 2019, 2020), Zwart (2001), and Cevolani, Festa, and Kuipers (2013).

For instance, Roger Rosenkrantz (1981, ch. 2) calls the previously mentioned property of the Brier score to penalize more heavily when the probabilities assigned to the false hypotheses are unequal (ceteris paribus) "attractive when all false alternatives are regarded as 'equidistant' from the truth." He further states, however, that "[w]here false answers are not equally far from the truth and we wish to weight them differently, we can use the weighted least squares metric." In the same vein, Hilary Greaves and David Wallace (2006, p. 628) claim that the quadratic scoring rule "can take account of the value of verisimilitude . . . by a judicious choice of the [weights]"; specifically, their proposal is to assign weight to a proposition depending on the extent to which it represents "a set of 'close' states" (Greaves & Wallace, 2006, p. 628).

As it stands, however, this proposal will not work. Which set or sets of states are close depends on which state is *actual*; equivalently, how far from the truth a false alternative is depends on which hypothesis is *true*. And the weights of the quadratic scoring rule lack this dependence. To be sure, if we know which hypothesis is true, the dependence can be built in by hand. For instance, given that (we said) the student will receive a C, so that A is more distant from the truth than B, we can weight A more heavily than B. That might give the desired result, and probably this is the kind of use of the quadratic scoring rule that the aforementioned authors had in mind.

But which weights are we to assign when we want to use the rule *not* knowing the truth, as when we would like to calculate our *expected* score? To calculate expected scores, we consider all possibilities of truth, calculate for each individual possibility the penalty that we would incur were that possibility to be actual, and then take a weighted average of those penalties, the weights being our probabilities for the possibilities. In using the quadratic scoring rule, however, there is a second set of weights involved. And the problem is that while truthlikeness relations shift from one possibility to the next—for instance, in the possibility in which the student receives a C, hypothesis C is closer to the truth than hypothesis A, but in the possibility in which the student receives a B, hypotheses A and C are equidistant from the truth—the weights attributed by the quadratic scoring rule stay the same whichever possibility is considered. Consequently, a set of weights that adequately reflects truthlikeness relations under one supposition of where the truth lies may well fail to do so under another such supposition.

To avoid this problem, we may adapt the quadratic scoring rule by *doubly*, instead of *singly*, indexing each weight, where the additional index then refers

to the true hypothesis, thereby obtaining what we shall call a *verisimilitude-sensitive scoring rule*, or VS rule for short. Where H_j is the true element of hypothesis partition H and w_{ij} is the distance from H_i to the truth, this rule imposes a penalty of $\mathcal{V}_j(\mathbf{p}) := \sum_{i=1}^{n} w_{ij}(\delta_{ij} - p_i)^2$ on someone assigning probabilities $\mathbf{p}$ to the elements of H. (Again, this is really a schema, yielding different rules for different weighting functions.)

To make the difference between the rules vivid, suppose that David, not knowing which grade his student will get, wishes to calculate his expected score, given his probabilities $\mathbf{d} = (.1, .5, .4)$ for hypotheses A, B, and C. First assume that he uses an instance of the quadratic scoring rule. Without loss of generality, let $w_A = .1$, $w_B = .3$, and $w_C = .6$. David will then find that[6]

$$\begin{aligned}\mathbb{E}_{\mathbf{d}}[\mathcal{Q}(\mathbf{d})] \;=\; & .1\big((.1)(.9^2) + (.3)(.5^2) + (.6)(.4^2)\big) \\ & + .5\big((.1)(.1^2) + (.3)(.5^2) + (.6)(.4^2)\big) \\ & + .4\big((.1)(.1^2) + (.3)(.5^2) + (.6)(.6^2)\big) \;=\; 0.228.\end{aligned}$$

Now suppose that David instead assumes a VS rule, for instance, that assigns a weight of .1 to the truth and that weights the other hypotheses proportionally to their distance from the truth. Supposing also that he uses a weighting function that reflects the ordering of the hypotheses—so that if A is true, then B is closer to the truth than C, and so on—but that is otherwise arbitrary, David's expected score is

$$\begin{aligned}\mathbb{E}_{\mathbf{d}}[\mathcal{V}(\mathbf{d})] \;=\; & .1\big((.1)(.9^2) + (.3)(.5^2) + (.6)(.4^2)\big) \\ & + .5\big((.45)(.1^2) + (.1)(.5^2) + (.45)(.4^2)\big) \\ & + .4\big((.6)(.1^2) + (.3)(.5^2) + (.1)(.6^2)\big) \;\approx\; 0.123.\end{aligned}$$

We see how the doubly indexed truthlikeness weights of the VS rule vary per possibility—as they must do, for the reason previously mentioned—while the singly indexed weights of the quadratic scoring rule stick to their propositions across all three possibilities.

6. To be entirely precise, instances of the quadratic scoring rule and the VS rule would have to be embellished with a superscript or subscript to indicate the weighting function that is being assumed. We will not be so fussy, however.

5.2 Truthlikeness and Inaccuracy Minimization

In the previous chapter, we criticized the Bayesian inaccuracy-minimization argument for focusing on a quite limited notion of expected inaccuracy, to wit, expected next-step inaccuracy. Suppose that this *is* the only notion of inaccuracy that we should care about, and also that, in adjudicating among update rules, we should care *only* about inaccuracy in this sense. Then a crucial question to be asked about our new rule is whether it still supports the Bayesian argument. To show that it does not, we suppose that David is about to receive some new piece of information E about the student from the previous example. How E is probabilistically related to hypotheses A, B, and C is specified by the following probability distribution over the set of possible worlds relevant to the present problem:

$$\begin{array}{lll} \Pr(\{w_{AE}\}) = .01 & \Pr(\{w_{BE}\}) = .25 & \Pr(\{w_{CE}\}) = .1 \\ \Pr(\{w_{A\overline{E}}\}) = .09 & \Pr(\{w_{B\overline{E}}\}) = .25 & \Pr(\{w_{C\overline{E}}\}) = .3 \end{array}$$

Here I am using the notation $\overline{E}$ to designate the negation of E, and w_{AE} is the possible world in which both A and E are true, and so on. Probabilities of hypotheses can be derived from this specification by summing the probabilities of the worlds in which they hold true; for instance, according to this specification, David's probability for A equals $\Pr(\{w_{AE}\}) + \Pr(\{w_{A\overline{E}}\}) = .01 + .09 = .1$, which is as it should be.

Once David has actually received the new information E, he should, according to Bayesians, conditionalize on it, which would result in the following new probability distribution over the same set of possible worlds:

$$\begin{array}{lll} \Pr_E(\{w_{AE}\}) \approx .028 & \Pr_E(\{w_{BE}\}) \approx .694 & \Pr_E(\{w_{CE}\}) \approx .278 \\ \Pr_E(\{w_{A\overline{E}}\}) = .0 & \Pr_E(\{w_{B\overline{E}}\}) = .0 & \Pr_E(\{w_{C\overline{E}}\}) = .0 \end{array}$$

To calculate David's expected VS score, assuming the weights we earlier assumed and otherwise following Greaves and Wallace (2006, pp. 459–460), or Leitgeb and Pettigrew (2010b, pp. 249–250), we set $x = .028 \approx \Pr_E(\{w_{AE}\})$, $y = .694 \approx \Pr_E(\{w_{BE}\})$, and $z = .278 \approx \Pr_E(\{w_{CE}\})$ in the following:

$$\begin{aligned} (5.1) \quad & .01\big(.1(1-x)^2 + .3y^2 + .6z^2\big) \\ & + .25\big(.45x^2 + .1(1-y)^2 + .45z^2\big) \\ & + .1\big(.6x^2 + .3y^2 + .1(1-z)^2\big), \end{aligned}$$

which yields an expected score of 0.034. The all-important question now is whether this is the minimum VS penalty David could achieve by updating (somehow) on E. And the answer is that it is not: minimizing the preceding function subject to $x + y + z = 1$ yields 0.032, where this minimum is reached at (.097, .703, .201).

To be sure, there may be no "natural" update rule via which David could have arrived at these probabilities, given his initial probabilities and the evidence that he received. But that is beside the point. Greaves and Wallace (2006) compare a Bayesian update with an update via some unspecified other rule, and they explicitly refrain from offering an "intuitive rationale" (p. 614) for the latter. Their aim is to show that Bayes's rule does better, in terms of expected (next-step) inaccuracy minimization, than that other rule, so that "considerations of intuitive plausibility need not be invoked in order to outlaw [that rule]" (Greaves & Wallace, 2006, p. 614).

It is also worth looking at natural alternatives to Bayes's rule and see how, on the present score, Bayes's rule compares to those. Of course, there are natural alternatives to Bayes's rule in which we are *particularly* interested, to wit, probabilistic versions of abduction, one of which—EXPL—we already encountered in previous chapters. Does Bayes's rule do at least better than EXPL in terms of expected VS penalty minimization? We saw that EXPL is much like Bayes's rule except that it assigns a bonus to the best-explaining hypothesis, after which it renormalizes the probabilities (so that they are probabilities properly so called). Given that the bonus gets assigned after the Bayesian update, which is the first step in EXPL updating, it makes sense to divide it over the possible worlds consistent with the evidence, which in our case is only one world.[7] So, suppose that David adheres to EXPL, at least in the present context, and the explanation bonus that he assumes is .1. Suppose further that, for whatever exact reason, he deems hypothesis A worthy of this bonus once he receives the new information E. Then, after assigning the bonus and renormalizing, his doxastic state is specified by

$$\begin{array}{lll} \Pr_E(\{w_{AE}\}) \approx .239 & \Pr_E(\{w_{BE}\}) \approx .543 & \Pr_E(\{w_{CE}\}) \approx .217 \\ \Pr_E(\{w_{A\overline{E}}\}) = .0 & \Pr_E(\{w_{B\overline{E}}\}) = .0 & \Pr_E(\{w_{C\overline{E}}\}) = .0 \end{array}$$

7. If there were more than one, it would seem reasonable to divide the bonus in proportion to the worlds' probabilities.

To calculate David's current expectation (i.e., before his receipt of E) of the VS penalty incurred by an EXPL update on E, we substitute again the preceding values into (5.1) in the way that we did previously (so setting $x = .239$, etc.). This yields an expected VS penalty of 0.037, which is actually higher than the VS penalty that David expects to incur if he updates via Bayes's rule. But was this bound to happen? Or did David just make an unfortunate choice of truthlikeness weights? What if he had chosen a different value for c, the bonus to be assigned to best explanations? What if he had deemed hypothesis B the best explanation instead of hypothesis A?

The intuition behind VS rules is that although it is bad in general to assign a positive probability to a false hypothesis, it is worse the further the hypothesis is from the truth. So, suppose that a VS rule's weights *reflect truthlikeness in a minimally adequate sense* precisely if hypotheses are weighted as a function of their distance from the truth, with hypotheses further from the truth weighted more heavily than hypotheses closer to the truth. We can then ask generally whether by updating via EXPL one is guaranteed to incur a higher VS penalty than by updating via Bayes's rule, provided the weights associated with the rule reflect truthlikeness in a minimally adequate sense.

Keeping all probabilities as they are above, I had the computer search for an answer by letting it go through combinations of weights that reflect the truthlikeness relations among our hypotheses A, B, and C in a minimally adequate sense. I found that looping over 1,000 such combinations is typically enough to output a set of weights that fulfil our requirement while yielding a VS penalty for EXPL that is lower than that for Bayes's rule. For instance, if we change the weights in (5.1) as follows:

$$\begin{aligned} .01\big(&.0198(1-x)^2 + .4835y^2 + .4967z^2\big) \\ &+ .25\big(.9798x^2 + .0202(1-y)^2 + .9798z^2\big) \\ &+ .1\big(.4967x^2 + .4835y^2 + .0198(1-z)^2\big), \end{aligned}$$

then we obtain an expected score of 0.0469 for the Bayesian update versus an expected score of 0.0467 for the EXPL update. We do not know, of course, whether these weights reflect David's truthlikeness judgments for the hypotheses at issue, but they *might*.[8]

8. Appendix E shows how one can obtain such weights and also how one can, more systematically, find weights that will maximize the extent to which the EXPL update does better than the Bayesian update.

Alternatively, suppose that we keep David's weights as they are in the example, but instead of following van Fraassen in setting $c = 0.1$ we set $c = 0.05$. In that case, David's post-EXPL-update probabilities for $\Pr_E(\{w_{AE}\})$, $\Pr_E(\{w_{BE}\})$, and $\Pr_E(\{w_{CE}\})$ would have been .146, .610, and .244, respectively. It is easily verified that with these probabilities, David's VS penalty equals 0.033, which is better than the penalty he would incur if he updated by Bayes's rule (which equaled 0.034, as previously shown).

Finally, keeping the weights as they are and also keeping $c = 0.1$, one easily calculates that if David had assigned the bonus to hypothesis B, his relevant probabilities would have been $\Pr_E(\{w_{AE}\}) \approx .022$, $\Pr_E(\{w_{BE}\}) \approx .761$, and $\Pr_E(\{w_{CE}\}) \approx .033$, and those would have yielded a score of 0.033, which is again lower than his Bayesian-update score.

It has thus come to appear that unless Bayesians have some good reason for dismissing the new rule, their inaccuracy-minimization argument, which was limited in value to begin with (as argued in the previous chapter), unravels entirely. It is simply not true that Bayesian updating is guaranteed to minimize your expected inaccuracy. Nor does EXPL come with such a guarantee, but nobody has ever claimed as much. At least if we want to be serious about truthlikeness relations among the hypotheses of interest and hold that VS rules can legitimately account for such relations, then there may be *no* updating procedure that is guaranteed to minimize inaccuracy, always and everywhere. That would in effect dovetail nicely with much that follows in later chapters.

5.3 Truthlikeness and Impropriety

But perhaps there *is* a good reason for rejecting VS rules. Bayesians might say that while VS rules offer a seemingly straightforward way to take into account truthlikeness relations, they face a devastating objection. We can start to see what Bayesians might find so objectionable by considering that relative to his current probabilities, David minimizes his expected quadratic score by having those very probabilities; formally, $\arg\min_{\mathbf{p}} \mathbb{E}_{\mathbf{d}}[\mathcal{Q}(\mathbf{p})] = \mathbf{d}$. This is no coincidence. It is known that not only the Brier score but *all* instances of the quadratic scoring rule schema are proper, and even strictly proper (Rosenkrantz, 1981, ch. 2). On the other hand, relative to David's current

probabilities, 0.123 is *not* the minimum VS score he can incur, for minimizing

$$.1\left(.1(1-x)^2 + .3y^2 + .6z^2\right) + .5\left(.45x^2 + .1(1-y)^2 + .45z^2\right) + .4\left(.6x^2 + .3y^2 + .1(1-z)^2\right),$$

subject to $x + y + z = 1$, yields 0.118. Hence, the VS rule that David assumes is improper. This might not be a matter of great concern. Perhaps it is again due to David's just having made an unfortunate choice of truthlikeness weights. Perhaps we should again let the computer search for better weights. But this time the problem is more fundamental, because we have

Theorem 5.1 *Every VS rule whose weights reflect truthlikeness in a minimally adequate sense is improper.*

For a proof, see appendix B. In other words, it is not just that David was unlucky in the weighting function that he picked; he could have picked none that would not have led to the same problem of impropriety, or at least none that is minimally adequate.

How damning is this result for the class of VS rules? Whereas the mainstream holds that impropriety is *totally* damning, I would like to argue that the answer should be the following: it all depends on the purpose of our scoring.

As mentioned, scoring rules were initially meant for *eliciting* probabilities. To serve that purpose, they better not encourage the subject whose probabilities one would like to elicit to lie about those probabilities. But if we are assessed by means of an improper scoring rule, then by revealing our actual probabilities we may expect to incur a larger penalty than if we pretend to have different probabilities, as was shown in the case of David. That can make it disadvantageous to be truthful.

It is generally recognized that scoring rules may also be used for the purpose of *self-assessment*. If David adheres to the VS rule considered previously, and he wonders about the accuracy of his current probabilities, he may conclude that he should replace these probabilities with those found to minimize his expected penalty. That appears problematic; as Sarah Moss (2011, p. 1057) notes, it implies that self-assessment by means of an improper scoring rule "could motivate you to raise or lower your credences *ex nihilo*, in the absence of any new evidence whatsoever."

In response, it could be argued that finding out that our probabilities do not minimize our expected penalty is new information, so that when we thereupon adapt those probabilities we are *not* acting in the absence of new evidence. This would in effect be an instance of what Lombrozo (2020) calls "learning by thinking," occasioned by what she calls an "observation generated inside the head." But some might want to reply that the evidence should, in an intuitively clear sense, bear on the hypotheses to which our probabilities are assigned. The intuitively clear sense may be difficult to pin down formally, but let that pass.[9] Because using a VS rule, or any other improper scoring rule because self-assessment seems to face a threat more severe than the one toward which Moss has pointed.

Previously we found that given David's probabilities in table 5.1, he would actually minimize his expected penalty by raising his probability for *A* to .146, raising his probability for *B* to .548, and lowering his probability for *C* to .306. Ironically, if he shifts his probabilities accordingly, he will find that he actually minimizes his expected penalty by setting his probabilities for *A*, *B*, and *C* to .164, .593, and .243, respectively. It does not end there, for relative to *these* probabilities, David minimizes his expected penalty by setting his probabilities for the hypotheses to .166, .630, and .204. And it goes on. And on? That would seem to prevent David from ever having stable probabilities for *A*, *B*, and *C* unless he decides to arbitrarily stop self-assessing at some point.

However, the iterative minimization process, should David engage in it, turns out to reach a fixed point. To be entirely precise, David would eventually arrive at probability assignment $\mathbf{d}^{+} = (.083, .833, .083)$,[10] a probability

9. In this connection, it is also worth mentioning that, at least according to some influential Bayesian statisticians (Gelman & Hill, 2009; Gelman & Shalizi, 2012, 2013; Kruschke, 2013b), raising or lowering probabilities in the absence of the kind of evidence with "direct bearing" is accepted as a legitimate practice, most notably, as resulting from a so-called posterior predictive check in which a statistical model may be rejected because it is found unsatisfactory (according to informal criteria) in light of simulated data. (More exactly, in a posterior predictive check we use a fitted model to generate synthetic data and then check whether those data are "similar enough" to the data that were used for fitting the model; see, e.g., Gelman & Hill, 2007, p. 158.) If rejected, the model is to be replaced by a new one, which requires, among other things, a specification of new prior probabilities. The simulated data that can motivate this kind of model revision—including probability revision—is presumably not the kind of new evidence that Moss has in mind.

10. Depending on the software that one uses to calculate the fixed point and the different floating-point formats the software uses, the number of steps that David needs to reach this fixed point can differ. Using the FixedPointList function from *Mathematica*, the number

assignment that is "strongly self-recommending" (Greaves & Wallace, 2006, p. 619) in the sense that $\arg\min_{\mathbf{p}} \mathbb{E}_{\mathbf{d}^+}[\mathcal{Q}(\mathbf{p})] = \mathbf{d}^+$. (For completeness, we note that David's expected penalty at the fixed point equals 0.057, whereas for his initial probabilities it was 0.118, as shown.)

We get a first sense of what may be the real problem here if we repeat our calculations for Emma and Frank and find that they arrive at exactly the same fixed point as David! Not only that; varying the weighting function a little, we find that the three colleagues now arrive at a different fixed point but that this fixed point is again the same for all three of them. What is more, this is what we find no matter the colleagues' initial probabilities: if they iteratively adapt their probabilities on the basis of VS score minimization, they all arrive at a fixed point, and if they use the same weighting function, they arrive at the *same* fixed point!

We can illustrate this graphically by noting that vectors in the standard unit simplex (or probability simplex) of dimensionality $n-1$ can be interpreted as probability distributions on an n-element hypothesis partition, with the ith vector component representing the probability of the ith hypothesis (de Finetti, 1962).[11] Figure 5.1 shows three two-dimensional simplexes, in each of which the initial probability assignments of David, Emma, and Frank given in table 5.1 are represented by medium-sized dots. The smaller dots represent their consecutive probability assignments while they go through the VS-rule-governed minimization process, ending with the final assignments, which are represented by big dots. The difference among the minimization processes represented in the three simplexes is only that the VS rules that are assumed in them assign different truthlikeness weights: the left simplex represents the process that relies on the VS rule that was assumed in the previous example;

found was 442; using a custom-built function in Julia, it was 428. However, that the process would reach a fixed point (if perhaps not after 428 or even 442 steps) is guaranteed by theorem 5.2, to be stated.

11. Interpreting such simplexes requires a bit of practice. Some points have a straightforward interpretation. For instance, in our example, the vertex labeled (1, 0, 0) corresponds to the point where all probability is assigned to hypothesis A, and similarly for the other vertices. Also, the geometric center of the simplex corresponds to the flat distribution. Furthermore, any point on an edge connecting two vertices represents a distribution in which the hypothesis associated with the unconnected vertex receives zero probability. More generally, given a point in the simplex, the probability assigned to the hypothesis associated with a vertex is the shortest distance from the point to the edge opposite the vertex, divided by the sum of the lengths of the shortest distances from the point to the various edges. (This generalizes to simplexes of any dimensionality.)

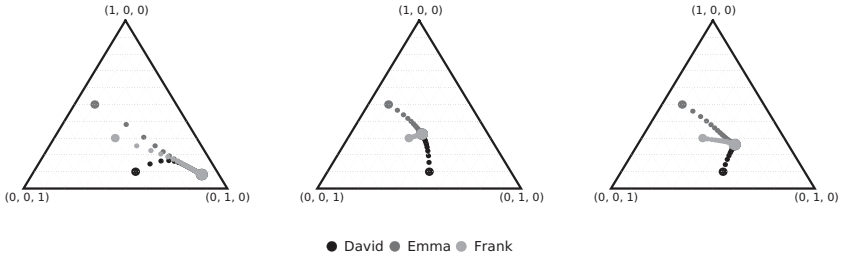

Figure 5.1: Three probability simplexes show that David, Emma, and Frank all reach the same fixed points (large dots) by iteratively adapting their probabilities on the basis of VS score minimization and illustrate that these points depend on the weights assumed by the given VS rule. Initial assignments are marked by medium-sized dots, intermediate assignments by small dots, and final assignments by big dots.

the processes shown in the other two simplexes used VS rules with slightly different weighting functions (the details are unimportant here).

To see that this finding is again no coincidence, consider the following:

Theorem 5.2 *Let S be the standard unit $(n-1)$-simplex, let $\mathbf{p}$ and $\mathbf{p}^*$ range over vectors in S, and let $m\colon S \longrightarrow S$ be defined as follows:*

$$m(\mathbf{p}) \;:=\; \arg\min_{\mathbf{p}^*} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i w_{ij} (\delta_{ij} - p_j^*)^2$$

with δ_{ij} the Kronecker delta, and with $w_{ij} > 0$ for all i, j, and $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = 1$. Then (1) there is a $\mathbf{p}^+ \in S$ such that $m(\mathbf{p}^+) = \mathbf{p}^+$; and furthermore (2) there is only one such $\mathbf{p}^+$, and (3) it depends only on the w_{ij}.

See appendix C for a proof. Further observe that for any VS rule $\mathcal{U}$

$$\arg\min_{\mathbf{p}^*} \mathbb{E}_{\mathbf{p}}[\mathcal{U}(\mathbf{p}^*)] \;=\; \arg\min_{\mathbf{p}^*} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i w_{ij} (\delta_{ij} - p_j^*)^2$$

with the w_{ij} being provided by the weighting function of the given rule. In other words, the function by which David, Emma, and Frank iteratively

update their probabilities in the represented processes is an instance of the function m defined in theorem 5.2. Thus, in light of this theorem it is unsurprising that their probability assignments reach a fixed point and that despite their different initial probability assignments, this fixed point is the same for the three of them, given that, as we assume, they use the same VS rule.

Surely we have hit upon an absurd result here, which offers a compelling reason to refrain from self-assessment by means of a VS rule. As Robert Winkler (1996) notes, however, scoring rules not only have (what he calls) an ex ante use; they can also be used ex post for evaluating performance. Potential problems with elicitation or self-assessment are beside the point if one is going to use an improper scoring rule for assessing people's performance without disclosing that rule to them. And often there will be no requirement for disclosure. An investment company might hope to identify job applicants with special capacities to predict the stock market by scoring their answers to test questions via some custom-built scoring rule without informing the applicants about the rule being used. Similarly, if a television network wants to hire a new weather forecaster and is now retroactively analyzing, by means of an improper scoring rule, the performance of various candidates being considered for the job, it will make no difference whether or not the network makes publicly known how it is conducting the analysis.[12]

As an aside, I mention an interesting real-life example of this kind of usage, which is found in recent work on forecasting carried out by a group of psychologists from various American universities (Mellers et al., 2015; Tetlock & Gardner, 2015). These researchers have organized, over a period of several years, a number of prediction tournaments, mostly concerning geopolitical questions. They found that some otherwise ordinary people were much more accurate forecasters than even professional intelligence analysts. A key objective of the research was to determine what distinguishes the most accurate forecasters from the rest of the population. The researchers used a number of different scoring rules for evaluating their participants' performance, including the Brier score but also the so-called AUROC, which is known to be an improper scoring rule (for details see, e.g., Agresti, 2007, ch. 5, or Hastie,

12. It is entirely consistent with everything said here that there are still other purposes that scoring rules can serve that may again require propriety. For instance, Roche and Shogenji's (2018) claim that we should measure informativeness in terms of inaccuracy reduction but that inaccuracy should then be measured by a proper scoring rule is not in conflict with the claim that scoring rules can serve purposes that do *not* require propriety.

Tibshirani, & Friedman, 2009, ch. 9). Given that the participants were never told what the evaluation process consisted of, the use of an improper scoring rule in that process will not have affected their responses.[13]

In short, depending on the purpose of our scoring, improper scoring rules appear to be perfectly admissible. In particular, when we are evaluating a probability distribution that is *given*—we are not eliciting the probabilities—from an *objective* standpoint—we are not assessing our own probabilities—I can see no impediment to using an improper scoring rule. Arguably, it is this objective evaluation of given probabilities that is at play in the Bayesian inaccuracy-minimization argument. The argument is not meant to apply only to *elicited* probabilities nor does it assume that we frequently, or even ever, assess our own probabilities in terms of inaccuracy minimization—indeed, we do not. It thus appears that the conclusion from the previous section—that the claim that Bayes's rule minimizes expected next-step inaccuracy is not generally tenable—still stands.

To end this chapter, it is worth pointing out that verisimilitude sensitivity *need* not come at the expense of propriety, for there is a scoring rule that is both verisimilitude sensitive *and* proper. This is the so-called ranked probability score (RPS), first proposed by Edward Epstein (1969) and shown to be strictly proper by Allan Murphy (1969). This rule has received hardly any attention from philosophers nor is it widely discussed outside philosophy (O'Hagan et al., 2006, p. 169).[14] Rather than comparing a probability distribution **p** on a partition of hypotheses with the vector **v** of truth-values of those hypotheses, it compares the cumulative distribution function of **p** with the cumulative distribution function of **v**. Given a partition $\{H_i\}_{i=1}^{n}$ and a probability distribution **p** on this partition, and supposing H_j to be true, the

13. Although in this research both proper and improper scoring rules were used for the purposes of selection, one could also use an improper scoring rule to select participants while at the same time scoring them via a proper scoring rule to determine their compensation in the experiment. Letting participants know how they will be compensated will then encourage them to post their true probabilities, while the improper scoring rule—the use of which is *not* disclosed to the participants—may still yield more useful information.

14. To the best of my knowledge, Konek (2016) and Vassend (in press) contain the only references to the rule (actually, the continuous version of the RPS rule) in the entire philosophical literature. I am not aware of any mention of the rule in the psychological literature.

ranked probability score associated with **p** is defined

$$\mathcal{R}_{\cdot j}(\mathbf{p}) \;:=\; \frac{\sum_{k=1}^{n}\left(\sum_{i=1}^{k} p_i - \gamma_{kj}\right)^2}{n-1},$$

where $\gamma_{kj} = 1$ if $k \geqslant j$, and $\gamma_{kj} = 0$ otherwise.

To illustrate, given David's probabilities **d** and still supposing C to be the true hypothesis (whence $j = 3$), David's ranked probability score is

$$\mathcal{R}_{\cdot 3}(\mathbf{d}) \;=\; \frac{(.1-0)^2 + (.1+.5-0)^2 + (.1+.5+.4-1)^2}{2} \;=\; 0.185.$$

In the same way, we find that Emma's ranked probability score equals 0.305, and Frank's, 0.225.[15] In the example in which the three colleagues' probabilities are about what grade their student will receive, these outcomes make perfect sense: as previously stated, David appears to do best in this case because he assigns a higher probability to the false hypothesis that is closer to the truth than to the one further from the truth. By the same token, Emma would seem to do worst, given that she does exactly the opposite. Frank steers a sort of middle course between his colleagues in assigning the two false hypotheses equal probability.

Bayesians might particularly like the RPS rule, for it is easy to verify that, at least in the example from the previous section (on p. 144), their favored rule does minimize expected RPS penalty (see appendix E). It may well do so generally, but I do not have a proof of that. Naturally, though, supposing

15. Because, as noted, the RPS rule is strictly proper, it satisfies Selten's third axiom (see footnote 4 in this chapter). To see that it also satisfies his fourth axiom, note that for comparing a probability assignment $(p_1, \ldots, p_n)$ with a "true" probability distribution $(p_1^*, \ldots, p_n^*)$, the RPS rule takes this form:

$$\frac{(p_1 - p_1^*)^2 + \left((p_1 + p_2) - (p_1^* + p_2^*)\right)^2 + \cdots + \left((p_1 + \cdots + p_n) - (p_1^* + \cdots + p_n^*)\right)^2}{n-1}.$$

The symmetry required by the fourth axiom then follows from the fact that the addends in the numerator are all squared. Furthermore, that David's and Emma's ranked probability scores are different, as seen in the main text, is enough to show that the rule does *not* satisfy Selten's first axiom. Finally, to show that neither does it satisfy the second axiom, we can add to the partition consisting of hypotheses A, B, and C the hypothesis that the student will receive a C–, where this has zero probability for David. Keeping his probabilities for A, B, and C as they were, David's ranked probability score then becomes (approximately) 0.243, and hence the addition of the zero-probability alternative did affect the score.

Bayes's rule does minimize expected inaccuracy vis-à-vis the RPS rule, that does not militate against the family of VS rules. Indeed, it is not even true that in the presence of the RPS rule, we have no use for VS rules. VS rules are still valuable because they are flexible in a way that the RPS rule is not. In particular, given the latter rule, relations of truthlikeness are completely fixed by the *ordering* of the hypotheses in the partition, and such orderings will not always reflect intuitive judgments of truthcloseness.[16]

Consider this example: Two football teams that are about equally strong and whose past five encounters have all been draws just played another match, in which neither team managed to score. Then the prediction that the match would end in a 0 : 4 away win appears to be quite a bit further from the truth than the prediction that the match would end in a 0 : 1 win. After all, the former much more than the latter would suggest a strongly dominant away team, which would have been a surprise in view of the relative strengths of the teams and the outcomes of their past encounters. On the other hand, a 0 : 24 away win prediction would hardly be further from the truth than a 0 : 21 away win prediction: we would find these end results about equally stunning and might say that both are "about as far from the truth as can be."[17]

Or consider any value representable in a conceptual space (in the manner of Gärdenfors, 2000), for example, a color shade. We might want to order different hypotheses about this value on the basis of distances in color space (CIELAB space or CIELUV space; see Fairchild, 2013). It is known, however, that although such distances correlate well with human similarity judgments

16. Thanks to Ilkka Niiniluoto for bringing this to my attention; also to William Roche for particularly helpful comments here.

17. Some might want to hold that we are to measure distance from the truth here in terms of the difference in goals scored. In my opinion, it is more reasonable to look at how different the various mentioned non-actual worlds (the world in which the match ends in a 0 : 1 win, the world in which the match ends in a 0 : 4 win, and so on) are from the actual world. And given what we know about the teams, our world would, as mentioned, have to be rather different from the actual world for the match to end 0 : 4 while it would not have to be very different for the match to end 0 : 1. By contrast, for the match to have ended 0 : 21, something entirely out of the ordinary would have had to occur, and whatever that would have been, it would have been about equally compatible with a 0 : 24 end result. For instance, if all players who normally play for the home team had been suspended, and the coach of that team had to line up their most inexperienced players, then a devastating loss would be explainable—but the explanation would be about as good in the case of a 0 : 21 end result as it would be in the case of a 0 : 24 end result. (Thanks to Theo Kuipers and Ilkka Niiniluoto for helpful discussion here.)

at a short range, they stop doing so at a longer range (Shepard, 1987). Thus, if the actual value of the color that we are looking for is some shade of red, then the hypothesis that it is a particular shade of orange may be intuitively closer to the truth than the hypothesis that it is a particular shade of yellow, and that will be reflected by the ordering of hypotheses. By contrast, the hypothesis that the actual color is a particular shade of blue would appear as far from the truth as the hypothesis that it is a particular shade of green, and that judgment is quite compatible with the particular shade of blue being closer in color space to the actual shade than the particular shade of green.

It is hard to see how one could accommodate the intuitions at play in these examples only in terms of an ordering of hypotheses. However, one can find refined measures of truthlikeness in the literature that do allow one to go beyond merely ordering hypotheses in terms of truthlikeness and to express truthlikeness relations among the aforementioned hypotheses in a way that does justice to the aforementioned intuitions (see, e.g., Niiniluoto, 1984, ch. 7, 1987, ch. 12; Kuipers, 1992, 2000, ch. 12). And while the RPS rule cannot take such refined measurements into account, VS rules *can*.

Roger Cooke (1991, p. 121) remarks that one of the main questions concerning scoring rules is whether "the score reward[s] those features that we would like subjective probability assessments to display." Much of the debate on scoring has been premised on the assumption that the desirable features are fixed across contexts and are independent of which goal or goals a scoring rule is to help achieve. I have argued that this assumption is false, and that different scoring rules may be called for in different contexts and also for different goals.[18] This undercuts the Bayesian claim that, in contrast to the dynamic Dutch book argument, the inaccuracy-minimization argument only involves a notion of epistemic rationality.

I have in particular emphasized the importance of taking truthlikeness relations into account in our scoring, if such relations are present. Following suggestions from Greaves and Wallace led us to the family of VS rules, which *are* sensitive to truthlikeness, but which also spelled trouble for the Bayesian inaccuracy-minimization argument, in that it is not generally the case that Bayesian updates minimize expected inaccuracy when inaccuracy is measured by a VS rule. So, lest Bayesians have a decisive reason for dismissing VS rules,

18. To my knowledge, the only other authors explicitly open to this possibility are Levinstein (2017) and Schurz (2019).

nothing is left of what most of them now consider to be the main argument in favor of Bayes's rule.