Response to RSS discussion

Glenn Shafer
Rutgers University, Newark, New Jersey, USA
E-mail: gshafer@business.rutgers.edu

I am gratified by this bounty of thought about my paper and my talk. I am also grateful to the organizers of the meeting, who worked so hard to make it work in this age of remote communication, and to Philip Dawid and Frank Coolen for initiating the vote of thanks.

I offer special thanks to the discussants who had already contributed directly to the paper's ideas: Dawid, Vladimir (Volodya) Vovk, and Peter Grünwald. Phil's work on probability forecasting in the 1980s and 1990s, especially his prequential principle, was seminal, and he has been a constant presence in the further development of game-theoretic probability by Volodya and myself. Peter's work is a more recent influence, but the writing of the paper was spurred by his and my efforts to understand each other, and his comments on the paper mark the success of those efforts.

Thanks also to all the other discussants: old friends, new friends, and others I hope to meet. My greatest delight is seeing a new generation of researchers who have been putting the ideas of my paper, sometimes differently named and motivated, to work.

Testing pundits and weather forecasters. As Tze Leung Lai and Anna Choi point out, betting scores are particularly relevant today, when so much attention is paid to time-varying probability forecasts for weather, elections, and sports. Our weather forecasts change hourly. My paper assumes that only one probability forecast is made for each outcome. But I began my talk by pointing out that we can also test by betting when the forecaster changes their forecast repeatedly before an outcome is realized or revealed.

On 5 November 2019, the widely followed statistician and pundit Nate Silver announced that he would post and regularly update probabilities for who would become the Democratic nominee for the United States presidency. How could we have tested Silver's successive probability distributions? By betting. On Day 1, using play money because Silver had not offered to bet, we could have bought a nonnegative payoff to which Silver's Day 1 distribution assigned expected value 1. On each following day we could have "rebalanced our portfolio", selling back to Silver at his new prices the payoff we had bought on the previous day and buying a new one with the proceeds. Had we multiplied our initial 1-unit stake by a large factor when the contest was over, we could claim to have discredited Silver's forecasting.

Shafer and Vovk (2001) suggested using this method to test the efficiency of financial markets. How else can you do it? The academic community in finance has not found any alternative. They once tried to address the issue by testing models that account for the observed prices. But this led nowhere; it merely created what financial economists have called "the joint-hypothesis problem". Eugene Fama (1991) hoisted a white flag with these words:

> Ambiguity about information and trading costs is not, however, the main obstacle to inferences about market efficiency. The joint-hypothesis problem is more serious. Thus, market efficiency per se is not testable. It must be tested jointly with some model of equilibrium, an asset-pricing model.

Twenty years ago, Leo Breiman warned us that the culture of prediction was overtaking the culture of stochastic data models (Breiman, 2001). We now see, in business, government, and our daily life, a flood of numerical predictions, many produced by algorithms (neural networks and physical models, for example) that bear only a tortured resemblance to stochastic models. Testing by betting provides common ground, a starting point for teaching and understanding that connects significance tests, estimation, and likelihood with other algorithms of data science.

Perhaps this broader perspective can be advanced by replacing the terms "hypothesis testing" and "significance testing" with "prediction testing". Even Nick Longford, who wants us to abandon the "ritual of hypothesis testing" in favor of a calculus that sums up expectations of costs and benefits, may sometimes want to check how well his expectations are predicting what happens.

Choosing among bets. Philip Dawid asks what we should do when we have no particular alternative Q in mind but think of several bets we might like to make. Can we make them all, find the resulting maximum, and make some sort of multiplicity adjustment? Yes. Any nonnegative increasing function of the maximum that has expected value 1 will provide the needed adjustment. But an easier and usually better way of combining bets is to average them; any weighted average of a collection of bets is itself a bet. As Ruodu Wang points out, this requires no assumptions about the dependence structure and essentially dominates any other method of combination.

Averaging bets two bets against P, say S_1 and S_2, is equivalent to averaging the two implied alternatives S_1P and S_2P. The idea of averaging possible alternative distributions is very familiar. When we when average alternative distributions with weights proportional to prior probabilities, the resulting betting score is identical with what is often called the *Bayes factor* for testing the simple null hypothesis P. (See for example, (Kass and Raftery, 1995, page 776). The name is more often used for the inverse of this quantity, which is Harold Jeffreys's K.) We should not, however, equate betting scores with Bayes factors. Some betting scores are not Bayes factors. These include any betting score obtained by multiplying betting scores against successive predictions when the theory, forecaster, or pundit making the predictions did not give a joint distribution in advance and hence the bettor

did not make a joint bet in advance. As Vladimir Vovk notes, some Bayes factors (namely, many for testing composite null hypotheses) are not betting scores.

Aaditya Ramdas, Vovk, and Wang report numerous important examples of the effectiveness of averaging bets. Ramdas notes that he has shown that such averaging can lead to inferences that are, in some practical sense, universal. The idea of averaging all bets that we can compute in order to obtain a universal bet goes back to Jean Ville, who noted that the universal bet thus obtained is not itself computable (Bienvenu et al., 2009). For Ville, this impossibility of a bet that it is both truly universal and implementable was a reason to use the axioms of measure rather than game theory as the mathematical foundation of probability. But from a more practical point of view, it may mean only that what is universal for one purpose may not be universal for another.

Judith ter Schure provides important insights about the choice of bets from the perspective of an applied statistician doing meta-analysis. If I understand correctly, she proposes to consider not an alternative to P that she considers most possibly true but rather one that departs from P in ways most damaging if P is used as an input for decisions. This can be achieved most readily by choosing a test statistic T that measures the unwelcome departures, perhaps a statistic already widely used or a simple modification of such a statistic. If T is nonnegative and bounded, then $T/\mathbf{E}_P(T)$ may serve as an appropriate bet. Averaging bets may also have a role to play here. Suppose, for example, that $Y = (W, Z)$. Suppose the statistician is most concerned about P's predictions of Z and worries that that these predictions may err in certain ways depending on W. In this case, she might begin by choosing bets S_w against P's conditional distribution for Z given w. Then she can obtain a bet S by averaging the S_w using P's marginal for W. It will be interesting to see these ideas put into practice.

Betting scores and e-values. My paper's notions of bet and betting score are very old, and a variety of names have been given to these objects. In 1993, Vladimir Vovk and Vladimir V'yugin wrote $p(\omega)$ instead of $S(y)$ and called p an *impossibility measure* (Vovk and V'yugin, 1993, p. 257). Testing by betting is the central idea of Shafer and Vovk (2001) and Shafer and Vovk (2019), but these books did not settle on a single name and symbol for the factor by which Sceptic's capital is multiplied. After the 2019 book was completed, Vovk and I introduced new names for this factor. I introduced *bet* and *betting score* in an early version of the paper under discussion (Shafer, 2019a). Vovk (2019) introduced *e-value* for what I am calling a betting score, and Vovk and Ruodu Wang subsequently settled on *e-variable* for what I am calling a bet (Vovk and Wang, 2019).

The name *e-value* has proven attractive, because it evokes only the standard notion of expected value, with no explicit reference to betting and no hint of any game-theoretic excursion outside familiar territory. In the discussion, we see *e-value* used by Peter Grünwald and Aaditya Ramdas, as well as by Vovk and Wang. The name *e-variable* has been used less. Vovk and Grünwald both distinguish systematically between *e-variable* and *e-value*. But Wang uses *e-value* for both concepts, and Ramdas first distinguishes *e-value* from *e-variable* but then calls a martingale

an e-value. These variations are not surprising; variable/value distinctions generally tend to be unstable in mathematical conversation. I am hoping that the distinction *bet/betting score*, being less abstract, will be easier to bring to mind in settings where it is helpful. A bet is an action; the betting score is the result of the action.

In one respect, the notion of a *betting score* is more general than Vovk's notion of an *e-value*. In my paper, I emphasized the case where one scientist makes a bet S_1 on the outcome of her experiment and then, perhaps because $S_1(y_1)$ is large but only moderately so, the same or a different scientist does a newly devised experiment to test the same hypothesis, making the bet S_2. This, I proposed, yields a betting score $S_1(y_1)S_2(y_2)$, even though no corresponding bet or strategy for betting had been declared at the outset. No one had a joint distribution for Y_1 and Y_2 at the outset, because no one had a model for whether and how some someone might react to the first result and devise and perform the new experiment that created Y_2. So we have a betting score with no single bet, whereas an e-value must always be the value of an e-variable.

Some readers may prefer to eliminate this divergence by positing that some demigod, who somehow had a joint distribution for what experiments would be performed but did not know how they would come out, made a bet S whose payoff turned out to be $S_1(y_1)S_2(y_2)$. I am not tempted by this fantasy, and I question its usefulness in scientific communication. I propose that we instead rely on the simple principle that a forecaster or theory is discredited when and to the extent that a person or community multiplies an initial stake by a large factor by betting against its predictions, no matter whether there was only one bet or many in succession, and no matter how each successive bet was chosen.

When S is a bet at prices given by P, Markov's inequality tells us that

$$P\left(S \geq \frac{1}{\alpha}\right) \leq \alpha. \tag{1}$$

The intuition that $S(y)$ should not be large if P is correct is then supported by the principle that an event of small probability should not happen. This principle, sometimes called *Cournot's principle* (Shafer and Vovk, 2006), underlies all "frequentist" statistical methods. But when spoken aloud, it evokes debate from students and philosophers. Isn't what happens always an event of small probability? A good answer is that we consider only simple events that are declared in advance. If our students remain unsatisfied, we hasten to change the subject with words like "significance" and "confidence".

Implicit in my paper is the proposal that we turn our thinking about (1) upside down, taking as basic not the contested principle that an event of small probability should not happen but the principle that multiplying my money by a large factor discredits P's probabilistic predictions. This has multiple advantages as a strategy for communication. The public can understand it. It makes clear the need for a declaration in advance, for I cannot make a bet without declaring it. It puts just the right burden on the argument that failure to multiply my capital would be evidence for the hypothesis or theory being tested. (Am I so knowledgeable and effective, and the data so extensive, that no one will be able to refute the hypothesis if I cannot?)

It requires no discussion of what probability really means and no fantasies when we think about $S_1(y_1)S_2(y_2)$ as evidence against a theory even though it is not the realized value of a single bet.

When there is not a single bet, we are outside the framework of a single probability distribution. In this sense, we are outside measure-theoretic probability. As Vovk expains, the natural habitat for "betting scores" is game-theoretic, the natural habitat for "e-values" measure-theoretic. Shafer and Vovk (2019) show that game-theoretic probability has the same rigor as measure-theoretic probability and that the two are equivalent in essential respects. But working with betting scores does not require a study of game-theoretic probability, any more than working with significance tests requires a study of measure-theoretic probability.

Stephen Senn notes the analogy between multiplying betting scores and R. A. Fisher's multiplication of p-values. The analogy is only partial, because the product p_1p_2 of two p-values p_1 and p_2 is not itself a p-value. You must treat p_1p_2 (or $-\ln p_1p_2$, as Fisher preferred) as a test statistic and calculate a new p-value from it. But the broader logic is parallel. Fisher did not give a fig about measure theory, and he did not worry about the absence of a comprehensive probability distribution, posited in advance, for the existence and outcomes of multiple experiments.

Confidence and warranty. Philip Dawid asks, "does a $1/\alpha$-warranty set have any property analogous to the coverage property of a classical confidence interval?" Yes, at least in Protocol 5, the context in which I defined $1/\alpha$-warranty in my paper. In this protocol, where each parameter value defines a joint probability distribution for all the observations, a $1/\alpha$-warranty set is the same thing as a $(1-\alpha)$-confidence set. Only the name and the interpretation are different. I should have said this in the paper.

Why is $1/\alpha$-warranty the same as $(1-\alpha)$ confidence in Protocol 5? Because once Sceptic knows θ, the protocol reduces to Protocol 2, and there Sceptic has a strategy that multiplies his money by $1/\alpha$ for all $(y_1,\ldots,y_N) \in W$ if and only if $\alpha \geq P(W)$. Paraphrasing:

> $P(W)$ is the least number α such that Sceptic can multiply his money by $1/\alpha$ whenever W happens.

This is the game-theoretic definition of probability (Shafer and Vovk, 2019, Sections 2.1,7.3). The more general statement for expected value was Christiaan Huygens's definition of the value of an expectation (Shafer, 2019c), but today's textbooks no longer treat it as a basic principle. It can be derived from today's principles using Markov's inequality and the reasoning in my paper in the paragraph after Protocol 2. To extend the reasoning to protocols where N is random, use Ville's inequality rather than Markov's (Shafer and Vovk, 2019, p. 49).

Being identical to confidence intervals, warranty sets can share all their problems, including the problem of relevant subsets mentioned by Stephen Senn. But insights provided by the picture of sequential betting may sometimes help us choose warranty strategies (= confidence procedures) that mitigate some of these problems.

These insights can be used even when the data is obtained as a batch rather than sequentially (Waudby-Smith and Ramdas, 2020).

The identity between $(1-\alpha)$-confidence and $1/\alpha$-warranty holds in any context where confidence sets can be defined. But the notion of $1/\alpha$-warranty also extends to the situation where a sequence of experiments is not initially planned or even contemplated. There we can obtain a final $1/\alpha$-warranty set by supposing that Sceptic, who knows the value of θ from the outset of the first experiment, uses his capital at the end of each experiment to begin betting in the next. The statistician's final $1/\alpha$-warranty set will then be the set of θ for which the final capital after all the experiments is less than $1/\alpha$. Equating the warranty set with a confidence set in this situation would again require a very special demigod.

The coverage property that Dawid mentions requires that a confidence set cover the true parameter value with probability at least $1-\alpha$. Some statisticians, including Antoine Augustin Cournot (Cournot, 1843, Section 108) but not including Dawid, have further explained this by saying that if someone (another demigod!) were to repeat the experiment and the calculation many times in exactly similar circumstances, the true parameter value would be in the calculated set at least $1-\alpha$ of the time. Paul Vos pins our difficulties in communicating about probability on this talk about hypothetical repetition. I agree that such talk is no longer helpful in the ways it might have been in 1843. It misleadingly privileges identical repetition, directing our attention away from the case of a diverse unplanned sequence of tests.

All warranty sets do have a game-theoretic coverage property: a $1/\alpha$-warranty set will fail to cover the true parameter value with *game-theoretic probability* at most α. This is not a deep statement, however. By the definition of game-theoretic probability, an event has game-theoretic (upper) probability α or less when there is a betting strategy that multiplies the capital risked by $1/\alpha$ or more when the event happens.

Paul Smith reports that a familiar proposal popped up in the live chat: we should stop testing subsets of parameter spaces in favour of confidence intervals. One shortcoming of this proposal is that it applies only to models that specify a class of probability distributions (parametric or nonparametric), ignoring all other situations where we want to test predictions. Another difficulty, and the reason the proposal has never taken root, is that confidence intervals inevitably give birth to p-values and their abuses. Confidence intervals were first widely used after Joseph Fourier explained how to calculate them for the difference between two proportions, using error probabilities 1/2, 1/20, 1/200, etc. But who will refrain from asking which of the intervals do and do not contain zero? Jules Gavarret argued for fixed-level significance testing for the difference between the success rates of two medical treatments. But p-values were of course more popular. By 1843, Cournot was denouncing the resulting abuses. See Shafer (2019b) for details.

Betting strategies for parametric statistics. There is a large literature in parametric statistics concerning how tests of different parameter values might cohere. How should the way we test θ_1 be related to the way we test θ_2? Should we seek or expect to find a test of θ_1 that is best both when θ_2 is true and θ_3 is true? Philip

Dawid asks whether there are general principles for answering these questions when we test by betting.

Dawid uses my Protocol 7 as an example. It was structured so as to suggest a connection between Sceptic's betting strategies for the different values of μ: the bet on the error e_n should not depend on μ, which Sceptic knows, but only on the preceding errors $e_1, \ldots, e_{n-1}$. Is this a matter of principle? I think not. I exploited the group structure to get a simple betting strategy for Sceptic, but whether the statistician chooses this strategy should depend, I think, on the statistician's hunches about how the model might err. If the statistician has different hunches for different μ, or perhaps hunches that depend on other information (signals as in my Protocol 4), then she should take this into account rather than respect the group structure.

Similarly, we will often fail to find a test of θ_1 that is best against all alternatives. In protocols such as Protocol 5, where each parameter value specifies a complete probability distribution for each parameter value, the optimal bets $P_{\theta_2}/P_{\theta_1}$ and $P_{\theta_3}/P_{\theta_1}$ will generally be different.

Xiao-Li Meng's and Peter Grünwald's comments clarify the relationship between betting scores and Bayes factors for composite null hypotheses. In the case of a composite null hypothesis Θ_0 and a simple or composite alternative Θ_1, a Bayes factor is a random variable S of the form $S(Y) = Q(Y)/P(Y)$, where P is the weighted average of the distributions in Θ_0 obtained using a prior distribution μ_0 on Θ_0, and Q is similarly obtained using a prior distribution μ_1 on Θ_1. We have $\mathbf{E}_P(S) = 1$, and so S is a bet for testing P. But we did not ask for a bet that tests P. We wanted to simultaneously test all the P_θ with $\theta \in \Theta_0$. The relation $\mathbf{E}_P(S) = 1$ does not imply that S is such a bet.

Grünwald reports that he and his colleagues have shown that for every distribution μ_1 over the alternative, we can choose a distribution μ_0 over the null such that the resulting Bayes factor does test all $\theta \in \Theta_0$. This is nice generalization of what we know about a simple null hypothesis P: for every simple alternative Q there is a bet (namely Q/P) that tests P and whose value will be a Bayes factor for testing P.

Meng reports that his research with colleagues reveals a different connection between betting scores and Bayes factors for composite null hypotheses. Suppose a parameter θ, for which we have a prior π, indexes both the null and the alternative: the null is a class $(P_\theta)_{\theta\in\Theta}$; the alternative is a class $(Q_\theta)_{\theta\in\Theta}$. Write P and Q for Y's marginal under the null and alternative, respectively. Then the Bayes factor $Q(y)/P(y)$ is the expected value under π of the betting score $Q_\theta(y)/P_\theta(y)$ that we could have calculated had we known θ.

In some problems, including Harold Jeffreys's problem of testing whether additional parameters are needed, it seems reasonably natural to index the null and alternative with a common parameter. In others, such an indexing may be rather contrived. But in any case, a subjective expected value for an unknown betting score against the predictions of a partially known hypothesis is not necessarily a betting score against them, and its value as a test result may be questioned. If we use my device of having the statistician stand outside a betting protocol in which

Sceptic does the betting, we may be unimpressed when the statistician announces a high subjective expected value for the betting score.

Betting scores and p-values. A fixed significance level α and a numerically equal p-value p have different meanings and carry different weight. Rejection at level 0.01 means that an event selected in advance and alleged to have this small probability has happened. In the case of a p-value 0.01, the event alleged to have this probability was not selected in advance; only a class of events, the tail events for a particular test statistic, was selected in advance. So the p-value carries less weight. The two numbers are on different scales. It is reasonable to ask for a convention for shrinking p to fit it onto the α scale, while acknowledging that such a convention must be largely arbitrary. Because $1/\alpha$ is the payoff when the fixed-level test is interpreted as a bet, the scale for betting scores is the same as that for $1/\alpha$. So the task is to choose a convention for translating p to the $1/\alpha$ scale.

I am delighted to learn, from Vladimir Vovk, that Harold Jeffreys's rule of thumb for comparing p-values with Bayes factors agrees closely with my suggestion, $p \mapsto p^{-1/2} - 1$, for the key values $p = 0.05$ and $p = 0.01$. It is also notable that in the same appendix where Jeffreys offers his rule of thumb, he suggests that a Bayes factor of 10^{-2} can be considered decisive. Inverting this, we obtain a betting score of 100, corresponding under the mapping $p \mapsto p^{-1/2} - 1$ to $p = 1/10{,}201$. This number would not surprise Joseph Fourier, who treated a large-sample confidence interval as conclusive when the error probability had this order of magnitude.

The main purpose of the mapping $p \mapsto p^{-1/2} - 1$ is to facilitate conversation between people using different methods of inference; it may help one person understand roughly the meaning of numbers reported by another. But if you want a betting score, there is no good reason to calculate a p-value and then map the p-value to a betting score. It may be better to work directly with the test statistic T from which a p-value might be calculated. Usually T is chosen so that it tends to be larger than P expects when P is wrong in a way that concerns us. Even when we find it difficult to formulate a particular alternative Q for Y, we may be able to specify a range of values of $T(Y)$ that we are most concerned about, and this may help us fashion a bet $S(T(Y))$.

Sander Greenland writes, quite reasonably, that other mappings from $[0, 1]$ to $[0, \infty]$ are appropriate when other study goals are pursued. When measuring information, for example, probabilities are often transformed using $p \mapsto -\log_2 p$. The study goal of my paper is testing and combining tests, not information measurement. As I wrote in my Section 3, I need a function f such that $f(p)$ has expected value 1 when p is uniformly distributed between 0 and 1. The function $p \mapsto -\log_2 p$ does not satisfy this condition. The function $p \mapsto -\ln p$ does, but $p \mapsto p^{-1/2} - 1$ and other more complicated functions in the literature referenced in (Shafer and Vovk, 2019, Section 11.5) come closer to established opinion about the comparison of p-values with likelihood ratios and Bayes factors. Here are a few comparisons, using the inverse functions $S \mapsto \exp(-S)$ and $S \mapsto 1/(S+1)^2$.

S	$S \mapsto \exp(-S)$	$S \mapsto 1/(S+1)^2$	Jeffreys's rule of thumb
$10^{1/2}$	0.04	0.06	0.05
10	5×10^{-5}	0.008	0.01
15	3×10^{-7}	0.004	
$10^{3/2}$	2×10^{-14}	0.0009	

Harold Jeffreys called values of his K between 10^{-1} and $10^{-3/2}$ "strong" evidence (Jeffreys, 1961, p. 432). R. A. Fisher never made an equally relevant comparison, but he did state that parameter values with likelihood less than 1/15 of the maximum likelihood are "open to grave suspicion" and are "definitely unlikely" (Fisher, 1956, Sections III.6 and V.7).

There is an interesting connection between the problem of shrinking a p-value and the problem of adjustment when Sceptic is supposed to announce his final capital after a sequence of bets but instead announces the maximum he attained. As it turns out, the mappings that seem acceptable for the two cases are the same. This was demonstrated in Dawid et al. (2011), and it provides a partial answer a question Dawid himself asks now: can betting scores ever be adjusted to account for not all information being reported?

Simplicity. Participants in the live chat suggested that experiments might help us decide whether testing by betting is simpler than conventional testing, and Judith ter Schure promised to think about how to design relevant experiments.

Perhaps the most crucial choice will be the selection of participants. Will they be highly trained statisticians? Poker players, as Philip Dawid playfully suggests? Scientists who use statistical tests? Students in statistics classes? Or perhaps teenagers?

I did not play poker as a teenager, but I remember that my classmates, when disputing each others' predictions, readily used betting taunts: "Wanna bet?", "Put your money where your mouth is." Imagine making these reports to teenagers:

- Prof Shafer tested your app's rainfall predictions over the past year by betting against them and turned \$1 into \$10. He concluded that the app is not doing a good job.

- Prof Shafer constructed a statistical model for your app's rainfall predictions over the past year. The app's predictions were inaccurate by an amount he would have expected only 1% of the time. He concluded that if his model is right, the app is not doing a good job.

Which report would the teenager be more likely to remember? Which would she be able to repeat to a classmate? If you teach statistics, which do you think your students would be able to repeat accurately?

Arthur Paul Pedersen asks what would be left of my paper's contribution if my claims about simplicity were jettisoned. Other participants, especially Peter Grünwald, Aaditya Ramdas, Vladimir Vovk, and Ruodu Wang, answer Pedersen's question by pointing to numerous applications where simplicity is not the only

salient advantage of testing by betting. But simplicity is always an advantage, and in some cases it is decisive.

Jorge Mateu directs our attention to the possibility that betting might be the only testing method simple enough to be implemented. Suppose a probability model P is defined in such a way that its probabilities and expected values can be obtained only by simulation; see for example the spatio-temporal models in Tamayo-Uria et al. (2014). There may be obvious test statistics, but the simulations required to estimate tail probabilities may be impractical, and it may be even more difficult to identify alternatives and make power calculations. But if we choose a bounded non-negative test statistics T, then estimating $\mathbf{E}_P(T)$ to a couple significant figures may be much easier than estimating a tail probability. This will allow us to make the bet $S := T/\mathbf{E}_P(T)$. We can obtain the implied target with just one more simulation, because $\mathbf{E}_Q(\ln S) = \mathbf{E}_P(S \ln S)$. (Here, as Bruce Levin has pointed out to me, we are simulating Q with importance sampling.) Following the example of Augustine Kong and Nancy J. Cox, as reported by Xiao-Li Meng in his comments, we might then study the parametric model defined by $P_\theta(y) := P(y)\exp(\theta \ln S)/c_\theta$.

Testing is not our only task. I just suggested that we replace the term "hypothesis testing" with "prediction testing", because predictions are the only thing we can test. But testing predictions is not the statistician's only task. For one thing, we must make predictions. When we say that we are testing a hypothesis, we are usually constructing a complex argument that involves repeatedly making and testing predictions. Calling this process testing gives it a patina of objectivity that can enhance the statistician's authority but may come back to bite her.

In this iterative process of predicting and testing we deal with an important issue that both Priyantha Wijayatunga and Sander Greenland raise: uncertainty about the assumptions on which predictions are based. This is primarily an issue about how we make predictions, not about how we test them. If we multiply our money a lot betting against the predictions, they are discredited no matter how confident we were in their assumptions. If the predictions withstand many tests by many able and well informed scientists, the assumptions may be better than we thought.

Wijayatunga also raises another important issue: how accurate we need predictions to be. When we are testing with a single bet against a probability distribution (or with a strategy for betting on a sequence of outcomes for which we have a joint probability distribution), the implied alternative gives us an opportunity to answer this question. If the predictions provided by the null and by the implied alternative do not differ enough for us to care, then the study is of no value. When the predictions we are testing are not provided by a comprehensive probability distribution, as in meta-analysis or when we are testing time-varying forecasts, we can instead bet against upper and lower probabilities obtained by expanding each point prediction to an interval that represents the precision that matters (Shafer and Vovk, 2019, Chapter 6). A related idea is to introduce transaction costs, as in Wu and Shafer (2007).

When we are making predictions, playing the role of Forecaster in my Protocol 4 for example, some of the issues raised by Stephen Senn arise.

- Forecaster must avoid making offers that open him to arbitrage. This is assured in Protocol 4 by the requirement that his offers be defined by a probability distribution.

- The signal x_n may be thought of as a vector of covariates. Forecaster must decide which of them to use and how.

The statistician must also decide how to use the covariates in choosing her bets or the strategy for betting that she prescribes for Sceptic.

Frank Coolen raises the question of how we design an experiment to test P. This is another of statistician's many tasks. The notion of implied target, like the notion of power, should help us choose between experiments, but it does not help us design one.

Coolen also mentions the problem of pooling expert opinion to form probability distributions. This is one way Forecaster can make his predictions. Methodology on pooling opinion developed outside the statistical community, such as the work on prediction with expert advice, should also be part of the statistician's toolbox. Something may be gained by posing the problem in terms of a betting protocol like Protocol 4 (Shafer and Vovk, 2019, Chapter 12). The work by Frank Hampel that Coolen cites is also about how to make probabilistic predictions, not about how to test them.

Several discussants emphasize decision problems. As I said in my paper, I consider decision theory an important chapter in statistical methodology. In particular, I believe there are many situations where costs or utilities are available and the Neyman-Pearson lemma is applicable. In these situations, we are making an accept/reject decision that will not be revisited, and so we want to maximize $Q(S \geq 1/\alpha)$ rather than $\mathbf{E}_Q(\ln S)$. I also believe that there are many situations where we have reasonable ingredients for Bayesian assessments.

Contrary to R. A. Fisher's polemical suggestion that the Neyman-Pearson theory belongs only in industrial settings, we know that it is often applied to good effect in scientific investigations where the abundance of possible choices is so great that a preliminary accept/reject screening is required. Christine P. Chai notes that such screening can also be used to discard variables from a study. Chai characterizes this as a use of p-values, but when we use a cut-off, we are doing Neyman-Pearson (fixed-level) significance testing rather than using a p-value as a means of communication.

Chloe Krakauer and Kenneth Rice propose that we think of significance testing as a decision problem and discuss Bayesian solutions. A salient aspect of their proposal is that they consider three possible decisions: decide that a parameter is negative, decide that it is positive, or make no decision. Judging from their tone, I think they would agree with my own first reaction: their proposal may be helpful in some but hardly all cases where statisticians and scientists have been using significance tests. If I am testing the predictions of a theory in which many people are interested, those who come after me to test its other predictions will be interested in what evidence has been accumulated so far, but not in what decision I made.

Krakauer and Rice conclude with a question to me: "We welcome Prof Shafer's

thoughts on quantifying the plausibility of hunches, and how any corresponding calculus differs from that of Bayes." Here "hunch" may refer to my statement that my choice of S and hence Q "may be guided by some hunch about what might work..." The simplest answer to their question is that S quantifies my hunch but not its plausibility. This is one respect in which my proposal is non-Bayesian.

The role of the implied alternative. Christian Hennig applauds the notion of implied alternative as a tool to understand tests better but finds it too sophisticated for the communication of statistical results. He foresees, no doubt correctly, that it could generate its own abuses, and he fears that the whole betting picture, because bets are made to win, may lead to yet more misleading significant results.

In reflecting about these concerns, we need to recognize that different levels of simplicity and sophistication are needed for different audiences. Betting scores are simple enough to be communicated in newspapers, where they have the advantage that they communicate not only the strength of the evidence but also its uncertainty; everyone is immediately aware that a betting score depends on the choice of bet. The notion of an alternative target is indeed much more sophisticated, but it should have its place in communications among research workers. As my simple examples demonstrate, *bet/implied target* work together in a much simpler and less confusing way than the *significance level/p-value/power* triplet we now try to teach research workers.

The research workers with whom I have interacted the most in recent decades are professors in accounting and finance and doctoral students who aspire to this role. So I am painfully aware of the misuse of the concept of p-value and the non-use of the concept of power that characterizes the "top journals" in these fields. In my paper, I cited Cready (2019), Cready et al. (2019), and Harvey (2017), which provide glimpses into this situation. Here, as in a number of other research fields, the time has surely come to try something different.

Hennig nevertheless worries that I *seem* to imply that scientists should want to win their bets, and that this *seems* to take for granted "the incentive of journals for finding significance". Perhaps his use of *seem* acknowledges the possibility that abuses might be mitigated. We do want scientists to try very hard to win their play-money bets, no matter whether the predictions they are testing are their own or others'. But the implied target could actually help us rectify incentives. Its adoption by journals would force authors to evaluate the scientific merit of a proposed study. A low implied target means the study has little merit. A plausible implied alternative and high implied target means the study will be informative *regardless of its outcome.* Journals could provisionally accept such informative studies before they are carried out, thus encouraging the preregistration of studies and lessening the incentive to get statistical significance by hook or by crook.

Morality and objectivity. My friend and colleague Harry Crane wants us to bet real money. Philip Dawid and Christian Hennig worry, in contrast, that the negative consequences of betting make it problematic as a basis for communication. Even though I did bet $2 on a horse a few years ago (at my request, Harry was showing

me around the local racetrack), my own attitude towards gambling is rather negative. When proposals to relax laws restricting betting were on the ballot in states where I lived, I always voted no. But a vice is not always curtailed by suppressing understanding of it. By forgetting 19th-century insights into betting systems, we have facilitated their replication in 20th- and 21st-century finance (Crane and Shafer, 2020).

Testing by betting is so natural an idea that its absence from statisticians' discourse for hundreds of years cannot be merely an oversight. Statisticians usually avoid translating their methods and insights back into betting language, even though we all know that probability's rules derive from betting. Is this avoidance due primarily to moral objections? I suspect that another motive is at play: the pursuit of objectivity. We see this already in the first pages of Jacob Bernoulli's *Ars Conjectandi*, the book that first sought to turn the betting calculus into a general theory of probability. Bernoulli replaces Huygens's betting arguments with reasoning about equally possible cases. The bettor, the subjective constituent of the story, is disappeared.

This motive endures. The statistician's public still wants objective results. But there are now also many ready to poke holes in any scientific claim that has social implications. It is my hope that the language of betting can help us educate the public about why and where choice is needed when testing scientific claims statistically.

Martingales. As soon as it considers a sequence of outcomes, probability theory becomes the theory of martingales. Blaise Pascal, Christiaan Huygens, and Abraham De Moivre did not use the word *martingale*, but to find the value of a payoff that depends on multiple successive outcomes, they constructed betting strategies that yield the payoff. By the end of the 19th-century, casino-goers called almost any betting strategy a *martingale*. Jean Ville (1939) used the word instead for a betting strategy's capital process—the sequence of random variables that represents the capital of a bettor following the strategy. Abraham Wald, who was familiar with Ville's work in the 1930s, learned after he came to the United States that mathematical statisticians called a martingale a sequence of likelihood ratios. Joseph Doob adopted Ville's word *martingale* but used only its measure-theoretic definition. The word has now been adopted in a number of branches of statistics, but its betting meaning and fundamental role in probability theory usually remain unspoken. It is a measure of the degree to which we have suppressed the role of betting in probability that some of our discussants can discuss the identity "martingale = capital process" almost as if it were news.

As I noted in Section 4.1 of my paper, a global bet in a stochastic process can be implemented by a strategy for Sceptic's step-by-step betting; see Protocol 2. The betting score is thus the final value of a nonnegative martingale. This idea is generalized in many directions in Shafer and Vovk (2019); there we call capital processes *nonnegative supermartingales*, because we allow Sceptic to make disadvantageous bets, and because betting offers may be one-sided and fragmentary.

Aaditya Ramdas's very rich comments emphasize the relation between martin-

gales and bets in the measure-theoretic case. It will be very interesting to see how and to what extent his results extend to nonnegative supermartingales in the game-theoretic framework.

Tze Leung Lai has made innovative contributions to the use of martingales in statistics for many decades, and I am honored that he and Anna Choi have contributed to the discussion. I am especially glad that they call attention to the work by Lai, Gross, and Shen, which builds on work that Françoise Seillier-Moisewitsch and Philip Dawid published in 1993. Seillier-Moisewitsch and Dawid's basic insight, that a martingale central limit theorem can be based on only what happens on the path taken by a stochastic process, was crucial for the development of game-theoretic probability.

Examples. As Paul Smith reports in his note on the live chat, Peter Grünwald has posted some detailed examples of testing by betting at `safestatistics.com`. The simple example I discussed in Section 2.4 of my paper, that of testing whether a normal mean is zero, is also much more than a bauble. For two centuries, beginning with Laplace and Fourier, statisticians have tested the difference between two proportions using a normal approximation. This was already happening in medicine in the 1830s. So perhaps my simple classical example was already a first step towards meeting Sander Greenland's challenge to show how testing by betting translates into "real medical applications".

The unity and diversity of probability. In the 1980s, I argued that we should understand probability and mathematical statistics in terms of betting games. These games stand at the center of a wide circle of ideas. Different statistical methods use them in different ways. Different interpretations of probability should not be understood as different ways of assigning meaning to numbers but rather as different ways of assigning reference to entire games (Shafer, 1990). Beginning in the late 1990s, I learned from Vladimir Vovk how we can formalize betting games in game theory, distinguishing the roles of different players while reproducing and extending the classical mathematics of probability. This has only strengthened my belief that betting games can be used by statisticians, scientists, and other analysts in a variety of ways. Some probability arguments merely draw analogies with betting games. Others use a betting game as a model of a physical or social process. Others put a decision maker in the position of one of the players in a betting game. Others use the results of one of more betting games as evidence for incidentally related questions.

In the paper under discussion, I develop one way of using a betting game to test predictions. My passionate exposition of it may have led a few of the discussants to worry that I was arguing against the ways they have been using probability. I hope this response has convinced them otherwise.

In particular, let me assure Kuldeep Kumar that I fully support the use of Neyman-Pearson decision theory in tasks such as classification. The work on conformal prediction to which I have contributed falls into this category (Vovk et al., 2005). Testing by betting is also not in competition with Bayesian inference, as

Barbara Osimani's comments may suggest. Bayes's theorem is used to find probabilities for hypotheses, not to test predictions. Jeffreys sought to use Bayesian intuitions in testing, but as we have seen, testing by betting re-interprets rather than rejects many of his tests.

Let me similarly assure Ryan Martin that I have not renounced my own work on Dempster-Shafer theory and that I admire his and Chunhai Liu's refinements of it. Their interpretation of p-values as plausibilities is especially revealing. I see Dempster-Shafer theory, however, as only one tool for constructing arguments. Martin and Liu's inferential models are similarly only one tool. There is no single tool that should be used in every analysis or argument that draws on the idea of a betting game.

I would also avoid trying to stuff every insight afforded by one tool into the use of a different tool. Arthur Paul Pedersen and some participants in the live chat ask whether insights about utility should be brought into testing by betting, and Osimani asks whether prior information can be used in testing by betting in the same spirit as it is used in a Bayesian analysis. The ideas advanced by Judith ter Schure may be responsive to these questions. To my mind, more formal efforts in these directions would only muddy the waters. No sense in trying to use a screw driver as a saw.

Conclusion. Having disagreed with some of the discussants on small points, let me thank them all again for their interest, their support, and the many gems of insight they have added. Some have even given me valuable feedback on drafts of this response.

I especially appreciate Jorge Mateu's succinct statement of one of the most important virtues of betting scores: the uncertainty associated with a large betting score is highlighted as much as the certainty provided by a small p-value is sometimes exaggerated.

References

Bienvenu, L., G. Shafer, and A. Shen (2009). On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics 5*(1).

Breiman, L. (2001). Statistical modeling: The two cultures (with discussion). *Statistical Science 16*(3), 199–231.

Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités.* Paris: Hachette. Reprinted in 1984 as Volume I (Bernard Bru, editor) of Cournot (2010).

Cournot, A. A. (1973–2010). *Œuvres complètes.* Paris: Vrin. The volumes are numbered I through XI, but VI and XI are double volumes.

Crane, H. and G. Shafer (2020). Risk is random: The magic of the d'Alembert. Working Paper 57, `www.probabilityandfinance.com`.

Cready, W. M. (2019, August). Complacency at the gates: A field report on the non-impact of the ASA Statement on Statistical Significance and P-Values on the broader research community. *Significance 16*(4), 18–19.

Cready, W. M., J. He, W. Lin, C. Shao, D. Wang, and Y. Zhang (2019). Is there a confidence interval for that? A critical examination of null outcome reporting in accounting research. Available at SSRN: `https://ssrn.com/abstract=3131251` or `http://dx.doi.org/10.2139/ssrn.3131251`.

Dawid, A. P., S. de Rooij, G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk (2011). Insuring against loss of evidence in game-theoretic probability. *Statistics and Probability Letters 81*(1), 157–162.

Fama, E. F. (1991). Efficient capital markets: II. *The Journal of Finance 46*(5), 1575–1617.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd. Subsequent editions appeared in 1959 and 1973.

Harvey, C. R. (2017). The scientific outlook in financial economics. *Journal of Finance 72*(4), 1399–1440.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Shafer, G. (1990). The unity and diversity of probability (with discussion). *Statistical Science 5*(4), 435–462.

Shafer, G. (2019a). The language of betting as a strategy for statistical and scientific communication. arXiv:1903.06991 [math.ST].

Shafer, G. (2019b). On the nineteenth century origins of significance testing and p-hacking. Working Paper 55, `www.probabilityandfinance.com`.

Shafer, G. (2019c). Pascal's and Huygens's game-theoretic foundations for probability. *Sartoniana 32*, 117–145.

Shafer, G. and V. Vovk (2001). *Probability and Finance: It's Only a Game!* New York: Wiley.

Shafer, G. and V. Vovk (2006). The sources of Kolmogorov's *Grundbegriffe*. *Statistical Science 21*(1), 70–98.

Shafer, G. and V. Vovk (2019). *Game-Theoretic Foundations for Probability and Finance*. Hoboken, New Jersey: Wiley.

Tamayo-Uria, I., J. Mateu, and P. J. Diggle (2014). Modelling of the spatio-temporal distribution of rat sightings in an urban environment. *Spatial Statistics 9*, 192–206.

Ville, J. (1939). *Étude critique de la notion de collectif*. Paris: Gauthier-Villars.

Vovk, V. (2019). Non-algorithmic theory of randomness. arXiv:1910.00585 [math.ST].

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic Learning in a Random World*. Springer.

Vovk, V. and V. V. V'yugin (1993). On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society, Series B 55*(1), 253–266.

Vovk, V. and R. Wang (2019). Combining e-values and p-values. arXiv:191206116v1 [math.ST], to appear in *Annals of Statistics* as "E-values: Calibration, combination, and applications".

Waudby-Smith, I. and A. Ramdas (2020). Variance-adaptive confidence sequences by betting. arXiv:2010.09686 [math.ST].

Wu, W. and G. Shafer (2007). Testing lead-lag effects under game-theoretic efficient market hypotheses. Working Paper 23, `www.probabilityandfinance.com`.