Authors' Reply to the Discussion of 'Safe Testing'

Peter Grünwald[12], Rianne de Heide[3], and Wouter Koolen[1,4]

[1]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[2]Leiden University, The Netherlands
[3]Vrije Universiteit Amsterdam, The Netherlands
[4]University of Twente, The Netherlands

January 14, 2025

We thank all discussants, and especially the proposer and the seconder of the vote of thanks, for their insightful and encouraging contributions. We are particularly delighted by the diversity of philosophical positions found among the respondents (both *nullistic* and *non-nullistic* Bayesians, frequentists, likelihoodists) — apparently we struck a chord among all these groups, which was exactly what we had hoped for with this paper. We received so many comments that it is impossible for us to respond to each of them — freely paraphrasing [Diaconis and Freedman, 1986, page 86], we hope discussants whose remarks were not or not all singled out for reply will not feel insulted, on the theory that silence is consent, and no news is good news.

We grouped major themes, brought up by several discussants, into Section 1–4. This is followed by a final section with some more specific issues and questions.

1 GRO(W) Construction: Limitations, Extensions, and Alternatives

Limitations and Extensions I: getting rid of regularity conditions Our main result, Theorem 1, requires the alternative Q and the composite null to all have densities relative to the same underlying measure, and also imposes the *finite KL condition* $D(Q\|P_W) < \infty$: the KL divergence between Q and at least some subset of the convex hull of the null has to be finite. All extensions of Theorem 1 require analogous conditions. As noted by *Vovk* and *Larsson, Ramdas and Ruf*, in nonparametric settings these conditions are quite strong. Relatedly, our basic growth optimality notion GRO can sometimes trivialize if the finite KL condition does not hold. Luckily though, over the course of the last year, both Theorem 1 and the GRO notion have been significantly extended. First, Lardy et al. [2023] replaced the KL divergence by the more general notion of *description gain* and extended GRO accordingly. Second, as *Larsson, Ramdas and Ruf* discuss, in [Larsson et al., 2024] they managed to come up with further sweeping generalizations of the GRO e-variable, the reverse information projection and the result linking them, i.e. our Theorem 1. Their generalization of the GRO *e*-variable, which they call the *numeraire e*-variable, exists in complete generality, *under no regularity conditions whatsoever*. We are of course very happy that our result can be generalized in this way, thereby showing that the basic idea underlying our Theorem 1 is really the same idea as that employed in the construction of various nonparametric e-processes such as those in [Waudby-Smith and Ramdas, 2024] and those implicitly derived by e.g. Honda and Takemura [2010], Agrawal et al. [2021]. In our view, this essentially implies full generality of the GRO/numeraire e-variable for the case that the alternative Q is (a) a singleton which (b) does not assign positive probability to any event that is null under every distribution in the convex hull of the null. But what if (a) or (b) is violated?

Limitations and Extensions II: Changing the Filtration As to (a), as demonstrated in our Example 8, in the case of composite alternatives in combination with worst-case growth optimality (GROW and relative GROW), it can be necessary to move to a coarser filtration to obtain the optimal e-variable. *Tian and Yu* ask when such coarsening can lead to more (worst-case) growth or power. Our answer is that at present we only have a partial understanding of this matter, and it is an exciting area of ongoing research; see for example [Pérez-Ortiz et al., 2024, Choe and Ramdas, 2024]. Violation of (b) often occurs if the outcome $\mathbf{Y}$ really represents an infinite sequence of data. *Vovk* considers, for example, the case that $\mathbf{Y}$ is an infinite binary sequence and the null hypothesis expresses that the sequence is exchangeable. In this case, (b) is necessarily violated [Ramdas et al., 2022], for any Q that is not in the null. Then the extended GRO/numeraire e-variable against any such Q will still exist, but it will have an expected logarithmic growth of ∞ [Larsson et al., 2024, Theorem 2.5] and this will not tell us anything about whether it is a useful choice in an e-process defined on finite subsequences of $\mathbf{Y}$. In this case, again, changing the filtration can come to our rescue: Vovk and collaborators have developed elegant methods for designing e-processes in a particular coarsened filtration. Interestingly, once one restricts e-processes to the chosen filtration, expected logarithmic growth again becomes a natural notion of optimality. Given the fact that different types of filtration-coarsenings (i.e. not just Vovk's) have been successfully used in e-value design [Ramdas et al., 2023], we speculate that currently unknown yet useful filtration-coarsening are still to be discovered, and even that designing e-variables and processes by filtration change may somehow be unified with design by growth optimality and information projection.

Appropriateness of (RE)GRO(W) It is here that we turn to the insightful comments of *Ruodu Wang*, proposer of the vote of thanks. Professor Wang notes that the interpretation of the *e-power* (i.e. expected logarithmic growth) of *individual* e-variables is not clear. We agree that there is a sore point here: in our paper we define and optimize it for individual e-variables, but it derives its meaning from the optional continuation setting, in which more and more batches of data may be added. Wang also provides some intriguing examples in which the e-power is minus infinity yet the e-variable will quite likely grow large, and cannot become zero, under the alternative. The reciprocal of this very e-variable can be used to show that the opposite can happen as well: infinite e-power yet a large probability of being very small in practice. This clearly shows that there is more to e-variable theory than just GRO and its variations! One option is to turn attention to concave functions $f(E)$ of E different from the logarithm, i.e. those that are induced if we construct E-variables by Rényi-projection instead of reverse information projection — Lardy et al. [2023], Larsson et al. [2024] recently showed that such projections give rise to E-variables as well, and it appears that the corresponding functions f also admit a monetary interpretation [Soklakov, 2020]. Also the developments in [Grünwald, 2022] suggest that when loss functions may be set in a data-dependent manner, we may sometimes want to diverge from GRO. Nevertheless, we feel that the GRO criterion and the reverse information projection will remain central to the area, and we concur with Wang that the coming years" will bring better understanding of its advantages and limitations, after which it will remain the default notion of e-power, with properly acknowledged caveats."

Hennig writes that the GRO-criterion, namely optimising $\mathbf{E}_Q[\log E]$, may not look so attractive to a gambler who is interested in a payout of E rather than $\log E$. As stated, we agree that GRO may not always be the criterion of choice, but this particular remark seems besides the point: in our testing-by-betting game, the pay-off is E, not $\log E$, and we clearly state in the paper why it is not at all a good idea to optimize for E directly if optional continuation is a possibility. We are similarly puzzled by the remark "I am not convinced that their [i.e. the author's monetary] interpretation of e-values is clearer than the interpretation of p-values. Probabilities and consequently p-values have been given monetary interpretations as well in the literature". We would like to turn this remark on its head and argue that the monetary interpretation of p-values on the contrary makes clear that they are problematic as a notion of evidence in testing. Such a monetary interpretation goes as follows: suppose a collection of lottery tickets $\{\mathcal{T}_r : 0 < r < 1\}$ is for sale. Ticket $\mathcal{T}_r$ pays out $1/r$ euro if the event $p \leq r$ happens, where p is the observed p-value and 0 otherwise. Now, under the null hypothesis, the fair price of each ticket $\mathcal{T}_r$ is 1 euro. Now suppose you observe a p-value of p. This indicates that "you have observed an event for which, *if you had bought the ticket* $\mathcal{T}_p$ *in advance*, you would now have gained $1/p$ euro. The problem is with the *in advance*: you didn't know in advance that $\mathcal{T}_p$ is the ticket you should have bought (giving you the highest pay-off among all tickets). Hence, stating p as your evidence against the null, in betting terms, amounts to a form of cheating. We refer to Shafer [2021] who emphasizes the point and interprets p-to-e calibration as *making p-values honest.* Indeed, in his discussion of our paper, *Shafer* even writes regarding this issue "...the authors concede too much. Testing with e-values, they tell us, requires more data than testing with p-values. Was this concession extracted by referees? When we make p-values *honest*, e-values do not need more data to catch up." We agree with this statement and hasten to say that, when we wrote 'requires more data', we referred to a test that outputs merely a binary accept/reject decision, without a continuous measure of evidence accompanying it.

2 GRO(W) and Power

A Concrete Example *Greenland* hopes we can explain *by example* how e-variables should be chosen in practice, and in particular asks us to consider the case of observing a k-degree-of-freedom χ^2 statistic Y taking value y with p-value p. Greenland himself notes that two possible e-variables which suggest themselves for this null hypothesis perform terribly in terms of power, and challenges us to do better. We happily take up the challenge! We first note that in many practical scenarios in which such a test is done, Y will have a k-degree noncentral χ^2-density p_δ under the alternative, with some (usually unknown) noncentrality parameter δ. This suggests to take as our e-variable a likelihood ratio $E_\delta(y) := p_\delta(y)/p_0(\delta)(y)$, or a mixture $E_W(\delta)(y) := \int E_\delta(y) dW(\delta)$; since the null is simple, all such E_δ and E_W are e-variables. Since the family of noncentral χ^2-distributions is strictly of Pólya (∞)-type [Karlin, 1956], it constitutes a monotone-likelihood ratio family and therefore, for each δ and W, the corresponding likelihood ratio is a monotone function of the p-value, facilitating comparison. Like Greenland we henceforth focus on the case with 1 degree of freedom. As we show in Section 3.1, Proposition 3 of our paper, if we want to test $\delta = 0$ (i.e. the null hypothesis of central χ^2 is correct) vs. $\delta \geq \delta^*$ for some fixed minimal relevant effect size δ^* then the results of our paper can be *directly* applied: by the monotone likelihood ratio property, the GROW criterion prescribes to use the e-value E_{δ^*}. In Figure 1 we plot $E_\delta(y)$ as function of δ for fixed data $y = 3.84$, the observation corresponding to p-value $p = 0.05$. Now, in practice often no minimally relevant δ^* is given. In this case, we may still follow the prescription of our paper and equip δ with a prior to optimize for the REGROW (relative growth) criterion of Section 4.1-4.3. Exact computation of the REGROW-optimal prior is difficult; here we merely give an indication

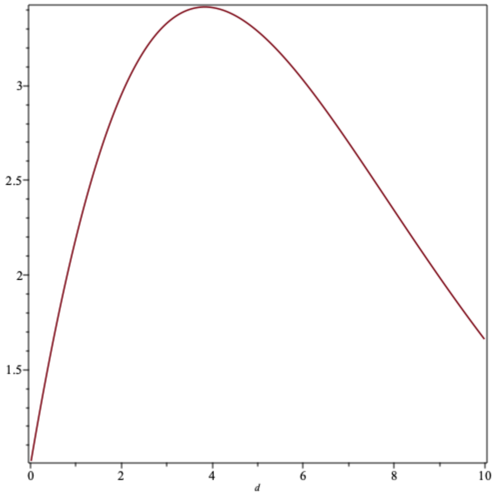

Figure 1: $E_\delta(y)$ as function of δ for fixed y corresponding to $p = 0.05$.

of what happens by taking an intuitively noninformative prior which puts heavy mass near 0 and ∞: the exponentiated Cauchy prior W_{CAUCHY} with density

$$w_{\text{CAUCHY}}(\delta) = \frac{1}{\pi} \frac{1}{\delta(1 + \log^2 \delta)}.$$

With such a prior, we get $E_{W_{\text{CAUCHY}}}(y) = 1.81$, for $y = 3.84$, i.e. $p = 0.05$, and we get $E_{W_{\text{CAUCHY}}}(y) \geq 20$ as soon as $y \geq 11.52$, corresponding to a $p \leq 0.0007$. Not very impressive, but already an enormous boost compared to Greenland's two intentionally naive proposals, the most powerful of which exceeds $1/\alpha = 20$ ('reject the null') if $p \leq 0.0000077$. However, we would like to do better. The REGROW criterion has been designed for settings in which data comes in sequentially and will perform especially well in combination with optional stopping, as in Section 6 of our paper. In the current one-shot, sample-size-1 setting (where e-values are still useful, because we may want to multiply them with other e-values later on), the REGROW criterion is not the method of choice. Instead, to obtain higher power, when α is given, then with monotone likelihood ratio families one may use the *uniformly most powerful (UMP) Bayes factor* as an e-variable [Johnson, 2013], obtained by using a point prior on δ which depends on α — it is designed to achieve, among all priors, the most power when rejecting if it reaches value $\geq 1/\alpha$ (while thus being optimized for a specific α, it keeps its Type-I error guarantee under optional continuation for any other threshold as well). In our case, this prior puts all mass on $\delta^* = 7.378..$, and we obtain $E_{\delta^*}(y) \geq 20$ if $y \geq \delta^*$, i.e. if $p \leq 0.007$ — a factor 10 better than what is obtained with the noninformative prior. This is the best that can be obtained for e-variables of the likelihood ratio form. We aim to investigate the use of UMP Bayes factors as priors to obtain powerful e-variables more generally in future work.

GRO vs UI *Martin*, seconder of the vote of thanks, together with V. Dixit, has recently [Dixit and Martin, 2023] proposed to design e-processes based on *predictive recursion* combined with *universal inference (UI)*. Dixit and Martin show that, up to first order, this leads to asymptotic growth rate comparable to our exact optimal (GRO) growth rate. Martin asks if there are practical or theoretical benefits to exact growth rate optimality compared to asymptotic growth rate optimality; a similar question was asked by *Tian and Yu*. While the precise answer depends on the situation, it turns out that exact growth rate optimality can be hugely important in some cases. For example, when testing parameters of a multivariate $d \times d$ Gaussian, it can be shown that UI and GRO have asymptotically the same expected logarithmic growth (or *e-power*, in *Wang*'s terms), up to sublinear terms, yet if n is not much larger than d the difference both in classical power and in e-power at n can be extreme. This is suggested by results from Tse and Davison [2022] (see especially the discussion by Spector et al. [2023]) and also by our own experiments; in ongoing work we make a theoretical comparison of both methods for exponential family nulls which points in the same direction.

We hasten to add that we really like Martin's idea to use predictive recursion as an alternative to e.g. Robbins' method of mixtures or the plug-in method — loss of power in the nonasymptotic setting occurs because of replacing GRO by UI, and seems unrelated to the use of predictive recursion.

Martin also notes that while classical methods (i.e. Neyman-Pearson style tests) do not automatically accommodate uncertain stopping times, they can be "manually" adjusted to handle pre-specified sets of stopping times $\mathcal{S}$. When $\mathcal{S}$ is somehow restricted, this will lead to procedures with larger classical power than our e-processes, which are designed to allow for optional stopping under *all* stopping times under the given filtration. Of course we concur with this (in fact, the classical method of α-spending is one way of doing just this), and we very much look forward to see Martin's intriguing conjecture that Choquet integration can be leveraged to efficiently handle the "some stopping times" case. On the other hand, we would like to emphasize that pre-specifying a set $\mathcal{S}$ and then adjusting classical methods deals with (restricted) optional stopping within a study, but not with optional *continuation* from one study to the next, which was our main target in this paper— directly using e-values enables this automatically. Similarly, one of us recently put forward a rather different motivation for testing with e-values: in contrast to classical p-value based methods, they can be used to obtain valid (in a certain frequentist sense) risk assessments of decisions of *post-hoc* loss functions, that are only determined after the data has been seen [Grünwald, 2022, 2023]. We doubt that this is easily realized by manual adjustment of classical methods.

3 Point- vs. Smooth- vs. No Priors for $\mathcal{H}_0$ and $\mathcal{H}_1$

Objective Bayes and Reference Priors We agree with *Bon and Robert* (who, adopting *Pawel and Held*'s terminology further below, we might call *nullistic* Bayesians) that the central proposal of our paper is relevant for the so-called *objective Bayes* approach and we strongly urge them and other interested readers to have a look at [Pérez-Ortiz et al., 2024, Section 5.2.]: the right-Haar prior is a standard choice in objective Bayesian modeling, but it has been criticized for not being uniquely defined in certain settings. This problem disappears if we consider e-variable-based "e-posteriors" [Grünwald, 2023] rather than Bayesian posteriors. They further write that it remains unclear to them (i) how to construct the least favorable prior on the null hypothesis on a general basis, especially from a computational viewpoint, and (ii) whether it degenerates into a point mass. Our reply is that the version of GRO that is most similar to reference Bayesian analysis is the REGROW criterion of Section 4.1–4.3, since, if the null is simple, it asymptotically amounts to a recommendation of Jeffreys' prior for the alternative. Nevertheless, we concede that it is not clear to us either how to construct it in general in a computationally efficient manner. In contrast to GROW, it will generally not lead to point mass priors. In fact one may think of our approach as allowing objective Bayesians who are willing to use REGROW priors on alternative and corresponding RIPr priors on the null, and frequentists preferring point priors (e.g. optimizing non-relative GROW) to collaborate — if they test the same null hypothesis on different data, they may meaningfully multiply their e-values even though these are based on different underlying inferential principles.

On a related point, *Robert and Bon* worry that, if the alternative is the Gaussian (unit variance) location family equipped with a 0-mean Gaussian prior, whereas the null is a unit variance Gaussian location family restricted to lie between $(-a, a)$, the corresponding GRO e-variable would use a prior W_0 in the denominator that places all mass on the extremities. It is not clear though whether this would actually be the case — experiments in similar (but not exactly the same) settings [Hao et al., 2023] indicate that the mass would be spread out symmetrically over many points or perhaps even form a continuum in the null (if the Gaussian prior over the alternative had a mean outside $[-a, a]$, then W_0 would indeed be concentrated on a singe point). In fact developing precise, not just asymptotic e-variables for multivariate Gaussians is a subject of ongoing research, with preliminary results appearing in [Grünwald et al., 2024].

From Empirical Bayes to Empirical E Consider a collection of e-variables $\{E_\gamma : \gamma \in \Gamma\}$ with γ representing some nuisance parameter. *Rizelli* asks if we could still get an 'approximate' e-variable (i.e. its expectation under the null should not be much larger than 1), if, akin to empirical Bayes methods, we would use the data-optimized e-variable $E_{\hat{\gamma}}$ where $\hat{\gamma}$ is determined by the data, e.g. the value which maximizes $E_\gamma = E_\gamma(\mathbf{Y})$ on the given data $\mathbf{Y}$. We hesitate to do so, since we regard having *exact, nonasymptotic* Type-I error guarantees and confidence intervals as one of the attractive features of our approach. On the other hand, one may certainly take the idea as a starting point, and turn $E_{\hat{\gamma}}$ into a proper e-variable. A simple (not necessarily optimal) means of doing this would be *renormalizing*, i.e. setting $E'(\mathbf{Y}) := E_{\hat{\gamma}}(\mathbf{Y}) / \sup_{P \in \mathcal{H}_0} \mathbf{E}_P[E_{\hat{\gamma}}(\mathbf{Y})]$. For simple nulls, this would lead to a variation of the GROW criterion in which, in information theoretic terms, redundancy is replaced by regret; a similar approach is taken by Orabona and Jun [2024]. We are currently exploring this idea further — it seems fruitful.

GROW vs. Shrinking Anytime-Valid Confidence Intervals *Ly* notices an inconvenience in cases where achieving GROW optimality necessitates the use of point priors. In these instances, the specified point of the point prior (often a minimal relevant effect size δ) never vanishes from the associated anytime-valid confidence interval. This is problematic, if — as often happens in applied work — the statistician wants to show the applied scientist at the same time the result of a null hypothesis test (with null corresponding to $\delta = 0$) and a confidence interval. The confidence interval should (a) be well-behaved, i.e. one should have the guarantee that its width shrinks to 0 as sample size increases. At the same time, (b) one wants the test and the confidence interval to be consistent with each other, i.e. the test should reject iff 0 falls outside of the interval. Ly's observation implies that for GROW optimal tests one cannot have (a) and (b) at the same time. We agree that this this is an issue, and we did not realize this to be the case when we developed the GROW criterion. Ly asks whether some adaptation of GROW ensures that the optimal solution involves a prior with full support, which avoids this issue. The short answer is: yes, the REGROW criterion we describe later on in the paper has this property - indeed with simple nulls it leads to priors that asymptotically agree with Jeffreys'. But the disadvantage of REGROW is that it does not allow us to make an informed prior guess of the effect size under the alternative. We are currently exploring the non-local priors of Johnson and Rossell [2010] which, it seems, may give us (almost) the best of both worlds.

4 Foundational Aspects

Logical Coherence and Likelihood Principles *Bickel* worries that the GRO-type e-variable constructions are not *logically coherent* in the sense of e.g. [Hansen and Rice, 2023], whereas e-variables based on *universal inference* [Wasserman et al., 2020] are, and concludes that e-variables call for a power-coherence trade-off. While we agree with this conclusion, we do not think the situation is as bleak as Bickel's quote of Royall [1997], "*We may conclude 'neither A nor B' but we may not conclude 'not-A'*" suggests. The reason is that the GRO e-variable relative to null hypothesis $\{P_\theta : \theta \in \Theta_0\}$ *is* a valid e-variable relative to any null hypothesis $\Theta'_0 \subset \Theta_0$ that is a subset of the original null hypothesis Θ_0. In practice we often start with one grand null Θ_0, for which we create the GRO e-variable. We can then later keep the same e-variable when we consider subsets of Θ_0, and whenever the data allows us to reject Θ_0 based on this e-variable, it will also allow us to reject any $\Theta'_0 \subseteq \Theta_0$. Logical incoherence only enters the fray once we start applying the GRO criterion also to design *novel* e-variables for subsets Θ'_0. Thus, two statisticians who both employ GRO, using the same alternative, but using null hypotheses indexed by Θ'_0 and Θ_0 with $\Theta'_0 \subsetneq \Theta_0$, may exhibit logical incoherence; but a single statistician who first considers Θ_0 and Θ'_0 may not.

Logical coherence is closely related to adherence to the likelihood principle. *Pace and Salvan* claim that 'Likelihood-based e-values obey the strong likelihood principle, in spite of the opinion that any frequentist desideratum is irreconcilable with such principle.' We think that the truth of this statement, just like Bickel's claim, depends on quite specific definitions and assumptions — because of this, Bickel's conclusion sounds a bit overly pessimistic to us, whereas Pace and Salvan's may be a bit overly optimistic. In any case, we are fascinated by connections to foundational notions such as logical coherence and the likelihood principle, and we plan to study these thoroughly in future work.

Testing vs. Estimation with Confidence *Pawel and Held* oppose what they call *nullistic* thinking, which lies behind much of testing methodology. We would like to emphasize that we are in fact very sympathetic to working with confidence intervals or depicting e-values as a function of the parameter, as they do — as witnessed by the similar figures appearing in the follow-up papers [Grünwald, 2022, 2023] written by one of us. But to engage in such uncertainty quantification beyond testing, one first has to get the testing part in order *mathematically*, as we do in this paper — even if later one does not aim to use it for actual testing.

Games vs. Measures *Shafer* believes that the value of our work will become even clearer once mathematical statisticians no longer distort their ideas to fit them into measure theory. — we fully agree, and we feel that optional (or 'free', as Shafer recently called it) continuation is really best given a game- rather than measure-theoretic treatment and interpretation, as in fact the example Shafer gives in his discussion clearly shows. We have to admit though that, while such a treatment would certainly be cleaner, it may also be quite hard to understand in detail by most statisticians (including us), who have been accustomed to measure-theoretic thinking since their undergraduate years.

Statistics meets Probability Theory, as it should *Wang*, proposer of the vote of thanks, enthusiastically notes that "The theory of martingales is essential for this methodology [i.e., using e-values], allowing statisticians to appreciate and contribute to this classic theory in probability. Conversely, the theory of e-values has successfully attracted probabilists and financial mathematicians to join the journey of statistics" and provides references proving this point. We fully agree, and we indeed hope and trust that this renewed joining of forces between probabilists and statisticians will turn out to become even more fruitful in the coming years. In the same vein, *Srakar* proposes extensions of the theory to more general stochastic processes — we agree this is a promising avenue.

Is our method really *safe*? *Pawel and Held* have reservations about whether 'safe' tests can make the scientific enterprise more reliable. As they write "For instance, researchers may still prefer classical non-sequential over sequential study designs, as the latter can be difficult to implement in practice. For example, interim analyses in randomised clinical trials will require unblinding and may as such threaten the integrity of the trial". We disagree: the main point of our paper is optional *continuation*, not interim analyses: we want to combine several studies with the same null, where the decision to start an entire new study may be informed by previous outcomes. The fact that you can often (not always, see Section 5 of our paper) do interim analysis within a study is a nice additional bonus feature, but not directly relevant to the main respect in which we claim 'safety'.

We do agree however, that in retrospect, 'safe' may not be such a good name, because it suggests safety *in every sense*, and this is of course impossible to achieve. This point was also made by *Hennig*. With hindsight, we should perhaps have given the paper a different (indeed, 'safer') name.

We agree with all these discussants that *every* method, including ours, should be used with utter care. For example, if outliers may occur (an issue raised by *Hennig*), we may want to safeguard against it by modifying any given e-variable E towards $(1-\gamma)E+\gamma$, which is still an e-variable. In betting terms, this corresponds to keeping a fraction γ of one's money in one's pocket. γ can then in fact be learned from the data by e.g. Robbins' method of mixtures (essentially, by equipping it with a prior). Relatedly, *Greenland*, more implicitly, warns about the lurking danger of e-hacking — we agree that this is a concern, as do Ramdas et al. [2023].

5 Miscellaneous Points and Questions

From Testing and Confidence Intervals to Estimation Suppose that E_δ is an e-variable for data $\mathbf{Y}$ relative to null hypothesis of effect size δ. *Cattaneo* suggests to use $\arg\min\{\delta : E_\delta(\mathbf{Y}) \leq 1/\alpha\}$ as an *estimator* of effect size δ, and asks how it behaves under optional continuation. This is an intriguing suggestion and question - we have not looked at it but we agree it deserves further study.

Terminology; Unfortunate over-loading of the term e-value *Dickhaus* and *Greenland* note that the term e-value is used with a different meaning in different sub-fields of statistics. We agree that this is quite unfortunate (we had no idea at the time the term was introduced), but by now 'our' use of the term e-value had become so widespread that it seems impossible to change. Alas, we do not see an easy way out.

Misspecification and Loss-Based Inference *Pawel and Held* ask whether e-values can be adjusted for potential model misspecification. One option is to resort to nonparametric e-values (see Ramdas et al. [2023] for an overview), which avoid the risk of misspecification to some extent. Sometimes though, one really wants to use some aspects of a quite specific parametric model that may still be misspecified. Within the Bayesian literature, a popular method in such cases is to either equip the likelihood with a learning rate or to use a *Gibbs posterior*, defined on a set of predictors rather than probability distributions, with the likelihood replaced by the exponentiated loss of the predictor on the observed data — methods that one of us has argued for already back in the 20th century [Grünwald, 1999]. It would be most interesting to see if our results can somehow be transplanted to this setting, since the mathematical analysis of Gibbs posteriors and e-variables bears quite some resemblance, both concepts being related to nonnegative supermartingales [Grünwald and Mehta, 2020]. And in fact, a first step in this direction has recently been taken by *Dey, Martin and Williams*. They generalize the universal inference framework by replacing the statistical model's negative log-likelihood with a multiple of the empirical risk and prove that it offers safe, anytime-valid inference on risk minimizers. We feel this is an exciting development and might call for a corresponding notion of growth optimality leading to better nonasymptotic behavior, just like standard GRO provides better growth and power than universal inference in the standard, likelihood-type setting.

Additional Questions *Pace and Salvan* note that under likelihood ratio symmetry, the developments of [Berger et al., 1994] suggests that it may be sensible to replace the threshold $1/\alpha$ by $(1-\alpha)/\alpha$, making the decision to reject based on an e-value with a Type-I error guarantee of α equivalent to the Bayesian posterior probability of H_0 being smaller than α, thereby bringing e-value and Bayes factor-based methods even closer. We agree. In fact, one of us has been intrigued by the very same question for a long time, and the 1994 paper by Berger et al. was a major source of inspiration for the present paper. A partial answer is given by [Grünwald, 2023, Section 3.2.,Eq. 31]. Essentially, rather than modifying the threshold, it is argued that under likelihood symmetry a special e-value can be designed which achieves threshold $1/\alpha$ if a standard, likelihood-ratio based e-value would reach $(1-\alpha)/\alpha$. However, it is also made clear that such an e-variable cannot easily be transferred into an e-process, and thus provides only for optional continuation and not for optional stopping.

Tian and Yu ask several additional questions, which are partially answered in other recent papers. For example, they ask about practical applications for our methods — these can be found in e.g. [Ter Schure et al., 2024, Turner et al., 2024]. Combination with regression is possible [Pérez-Ortiz et al., 2024] and by now there is a whole lot of literature on nonparametric e-processes [Ramdas et al., 2023, Waudby-Smith and Ramdas, 2024]. Their construction can often be thought of as a variation or strict generalization of the GRO criterion of our paper, as made explicit by Larsson et al. [2024]. They also ask about characterizing admissibility of an e-process — a succinct characterization is given by Ramdas et al. [2020].

Chai notices unclarity as to whether 0.8 is the Type-II error or the power in our experiments. It invariably refers to power. She also asks whether there are there any real-life examples where OC (optional continuation) or OS (optional stopping) is more appropriate? If we can choose either one, what are the guidelines to decide? The short answer is that OC automatically comes into play as soon as we deal with meta-analysis [Ter Schure and Grünwald, 2022]; whether OS is appropriate or not depends on the context. We may certainly be in situations in which they *both* apply. She also asks how the e-value compares with Held's *sceptical p-value*. A short reply is that, in contrast to e-values, the sceptical p-value has not been designed to allow continuation of studies more than two times, let alone indefinitely. A thorough reply would really require an in-depth study, for technically the ideas seem very different.

References

Shubhada Agrawal, Wouter M Koolen, and Sandeep Juneja. Optimal best-arm identification methods for tail-risk measures. In *Advances in Neural Information Processing Systems*, volume 34, pages 25578–25590, 2021.

J.O. Berger, L.D. Brown, and R.L. Wolpert. A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Annals of Statistics*, 22(4):1787–1807, 1994.

Yo Joong Choe and Aaditya Ramdas. Combining evidence across filtrations. *arXiv preprint arXiv:2402.09698*, 2024.

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14 (1):1–26, 1986.

Vaidehi Dixit and Ryan Martin. Anytime valid and asymptotically optimal statistical inference driven by predictive recursion. *arXiv preprint arXiv:2309.13441*, 2023.

P. D. Grünwald. Viewing all models as "probabilistic". In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, pages 171–182, 1999.

Peter Grünwald. Beyond Neyman-Pearson. *arXiv:2205.00901*, 2022.

Peter Grünwald. The E-posterior. *Philosophical Transactions of the Royal Society, Series A*, 2023. doi: 10.1098/rsta.2022.146.

Peter Grünwald, Tyron Lardy, Yunda Hao, Shaul K. Bar Lev, and Martijn de Jong. Optimal e-values for exponential families: the simple case. *arXiv preprint arXiv:2404.19465*, 2024.

Peter D Grünwald and Nishant A Mehta. Fast rates for general unbounded loss functions: from erm to generalized bayes. *The Journal of Machine Learning Research*, 21(1):2040–2119, 2020.

Spencer Hansen and Ken Rice. Coherent tests for interval null hypotheses. *The American Statistician*, 77(1):20–28, 2023. doi: 10.1080/00031305.2022.2050299.

Yunda Hao, Peter Grünwald, Tyron Lardy, Long Long, and Reuben Adams. E-values for k-sample tests with exponential families. *arXiv:2303.0047*, 2023.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of the Twenty-third Conference on Learning Theory (COLT 2010)*, 2010.

Valen E Johnson. Uniformly most powerful Bayesian tests. *Annals of statistics*, 41(4), 2013.

Valen E Johnson and David Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:143–170, 2010.

Samuel Karlin. Decision theory for pólya type distributions. case of two actions, i. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, pages 115–129. University of California Press, 1956.

Tyron Lardy, Peter Grünwald, and Peter Harremoës. Universal reverse information projections and optimal e-statistics. *arXiv preprint arXiv:2306.16646*, 2023.

Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *arXiv preprint arXiv:2402.18810*, 2024.

Francesco Orabona and Kwang-Sung Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 70(1), 2024.

Muriel Felipe Pérez-Ortiz, Tyron Lardy, Rianne De Heide, and Peter Grünwald. E-statistics, group invariance and anytime valid testing. *The Annals of Statistics*, 2024.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2020.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109, 2022.

Aaditya Ramdas, Peter Grünwald, Volodya Vovk, and Glenn Shafer. game-theoretic statistics and safe anytime-valid inference. *statistical science*, 2023. To appear.

Richard Royall. *Statistical evidence: a likelihood paradigm.* Chapman and Hall, 1997.

Glenn Shafer. Testing by betting: a strategy for statistical and scientific communication (with discussion and response). *Journal of the Royal Statistic Society A*, 184(2):407–478, 2021.

A.N. Soklakov. Economics of disagreement–financial intuition for the rényi divergence. *Entropy*, 8 (22), 2020.

Asher Spector, Emmanuel Candès, and Lihua Lei. A discussion of "a note on universal inference" by tse and davison. *Stat*, 12(1):e570, 2023. doi: https://doi.org/10.1002/sta4.570. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.570.

J. Ter Schure and P. Grünwald. ALL-IN meta-analysis: breathing life into living systematic reviews. *F1000Research*, 11(549), 2022.

J. Ter Schure, M.F. Perez-Ortiz, A. Ly, and P. Grünwald. The anytime-valid logrank test: Error control under continuous monitoring with unlimited horizon. *New England Journal of Statistics in Data Science*, 2024.

Timmy Tse and Anthony C Davison. A note on universal inference. *Stat*, 11(1):e501, 2022.

Rosanne Turner, Alexander Ly, and Peter Grünwald. Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *Statistical Planning and Inference*, 230, 2024.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2024. With discussion.