A Bayesian Critic for Frequentist Procedures

Isaiah Andrews, *MIT and NBER*
Simon Essig Aberg, *Harvard University*
Jesse M. Shapiro, *Harvard University and NBER*[*]

May 2026

Abstract

We propose a method for automated, probabilistic evaluation of the frequentist properties (e.g., bias, coverage) of procedures (e.g., estimators, confidence intervals) in a given setting. A Bayesian critic observes a sample of data and updates their prior belief on the underlying data-generating process (DGP). The resulting posterior belief about the DGP implies a posterior belief about the property of interest. When the critic's prior is in a low-precision Dirichlet process class, the critic's posterior can be approximated via a Bayesian bootstrap, making the method fully automated. We apply the method to several canonical settings and show that the critic shares some concerns raised in previous work and delivers new insights.

keywords: Monte Carlo, simulation, Bayesian nonparametrics
JEL codes: C11, C15, C18

[*]We thank a seminar audience at Harvard University and Kevin Chen, Guido Imbens, Sanjog Misra, Jon Roth, Bas Sanders, Frank Schorfheide, Matt Taddy, and Elie Tamer for helpful comments. We thank our dedicated research assistants for their contributions to this project, and Andrew Gelman and Guido Imbens for sharing their replication code and data. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2140743. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. E-mail: iandrews@mit.edu, sessigaberg@g.harvard.edu, jesse_shapiro@fas.harvard.edu.

> "[L]imit theorems 'as n tends to infinity' are logically devoid of content about what happens at any particular n. All they can do is suggest certain approaches whose performance must then be checked on the case at hand."
>
> — Lucien Le Cam, *Asymptotic Methods in Statistical Decision Theory*, 1986, p. iv.

1 Introduction

In the frequentist approach to statistics, an analyst selects a procedure (e.g., an estimator, a confidence procedure) with the aim of obtaining certain guarantees (e.g., unbiasedness, coverage) over a class of data-generating processes (DGPs). If the audience for the analyst's findings trust that the class is large enough to include the true DGP, such guarantees aid the audience's interpretation of the findings. As it is typically impossible to evaluate the performance of a given procedure numerically across all DGPs in a large class, such guarantees are often derived analytically. As exact analytic characterizations are intractable in many cases, analytical guarantees are often based on asymptotic approximations. Such approximations are, by their nature, imperfect, and their accuracy at a given sample size can depend on the unknown DGP.[1]

How well a given frequentist procedure performs is, then, typically an open question, whose answer depends on the setting. Econometric theorists have in recent decades devoted a great deal of attention to characterizing economically relevant situations (e.g., with heteroskedasticity, influential observations, failure of overidentifying restrictions, or weak identification) where common procedures may fail to fulfill their guarantees, and to finding alternative procedures which are more reliable. Yet as these alternative procedures are themselves grounded in (alternative) approximations, the question of how a given procedure performs in a given economic setting remains.

In this paper, we propose a method for automated, probabilistic evaluation of the frequentist properties of a given procedure in a given economic setting. We apply the method to several canonical settings in empirical economics where doubts have been raised about

[1] Pfanzagl (1994, p. ix) calls the reliance on approximations "the crucial drawback of asymptotic theory." In addition to the epigraph, see, for example, Freedman (2009, p. 211).

the performance of specific frequentist procedures. In each case, we show that the method automatically recovers the original cause for concern, and also answers additional questions not posed in the original analyses.

We consider an analyst who observes a sample of independent and identically distributed observations drawn from an unknown distribution that we call the DGP. The analyst proposes a procedure that yields some sample statistic. The interpretation of the procedure is summarized by a property. The property depends on the DGP through the distribution of the procedure and through the value of a statistical parameter that is a known functional of the DGP. For example, if the procedure is an estimator (a function of the sample), then the property could be the bias of the estimator for the corresponding parameter. The bias is equal to the estimator's expected value (a functional of the DGP), less the parameter's true value (also a functional of the DGP). Likewise, if the procedure is a confidence interval for some parameter, then the property could be the coverage of the interval for the parameter. The coverage is the probability that the interval contains the true value of the parameter, where this probability is again a functional of the DGP.

The analyst reports the result of the procedure to a scientific audience who are interested in the parameter and who hold potentially varying views about which DGPs are (most) plausible. We define a frequentist guarantee as a claim that the property is close to some desired value (e.g., zero bias, nominal coverage) for any DGP in a class of DGPs. If the class contains all DGPs deemed plausible by members of the audience, then the audience trusts the guarantee, establishing a sense in which the audience can safely rely on the output of the procedure.

How can the analyst assess the guarantee? Numerical simulation from all plausible DGPs is infeasible in many settings. Numerical simulation from a single DGP is often feasible, but does not reflect uncertainty about the DGP. Numerical simulation from multiple a priori reasonable DGPs may fail to capture that some DGPs are more consistent with the data than others.

We propose instead to leverage ideas from Bayesian data analysis. Suppose the analyst has access to a critic. The critic is a Bayesian who holds a prior belief about the DGP. The critic updates their belief based on the observed sample, yielding a posterior belief about the

true DGP. Such a posterior belief naturally implies a posterior belief about the statistical parameter of interest; this posterior belief is the basis of standard Bayesian inference. A key observation is that, because the frequentist property can also be expressed as a functional of the true DGP, any posterior belief about the DGP naturally also implies a posterior belief about the frequentist property. In many situations, this posterior belief can be elicited using standard numerical methods.

The analyst can therefore ask the critic to assess the frequentist guarantee under the critic's posterior belief. We focus for concreteness on the critic's posterior probability that the guarantee fails (e.g., that bias is too far from zero, or coverage is too far below nominal), which we call the critic's doubt. The critic may doubt the guarantee for different reasons, for instance because the sample size is too small for the guarantee to apply (e.g., as with influential observations or weak identification), or because the sample data appear inconsistent with the conditions on the DGP (e.g., homoskedastic errors, agreement with overidentifying restrictions) needed for the guarantee to apply.

If the critic doubts the guarantee, then the property fails to meet the desired tolerance for at least some DGPs. If the critic is a plausible member of the analyst's audience, or has beliefs plausibly close to such a member, then at least some audience member cannot rely on the procedure in the sense prescribed by the guarantee. For guarantees justified by large-sample approximations, we further show conditions under which the critic does not raise "false alarms": if the frequentist guarantee holds in a sufficiently large sample, then a critic who has seen a sufficiently large sample will not doubt the guarantee.

Importantly, assuming that the critic is a plausible member of the analyst's audience, or has beliefs plausibly close to such a member, is weaker than assuming that all audience members (or the analyst) accept the critic's prior as authoritative. If the critic's prior were authoritative, then the critic could validate, as well as cast doubt on, the guarantee, but in that case the analyst could dispense with frequentist statistics altogether, and simply report the critic's posterior belief about the parameter of interest to the audience.

For settings in which a compelling subjective prior is not available, we propose taking the critic's prior to be a low-precision Dirichlet process. In this case, existing results imply that the Bayesian bootstrap (Rubin 1981, Gasparini 1995) can be used to sample from a

distribution over DGPs that approximates the critic's posterior belief (see also Chamberlain and Imbens 2003, Taddy et al. 2016, Andrews and Shapiro, forthcoming). The resulting default critic is fully automated up to a choice of numerical precision.

As a warm-up, we first illustrate the use of a critic with a simple application to data from a randomized get-out-the-vote experiment. Invoking the default critic only requires defining a parameter of interest, a procedure for learning about it, and a property that relates the two. In this application, one parameter of interest is the difference in voting probabilities between the treatment and control groups. Here, the critic concludes that (up to simulation error) the plug-in estimator is unbiased and that the usual asymptotic confidence interval is likely to have correct coverage. Another parameter of interest is the cost effectiveness of the intervention, which is proportional to the inverse of the difference in voting probabilities. Here, the critic concludes that the plug-in estimator is likely somewhat biased and the delta-method confidence interval may (rarely) undercover. These conclusions are unsurprising given theoretical knowledge about these procedures. The value of the critic is in allowing the analyst to reach these conclusions automatically.

We then turn to richer applications. We focus on canonical settings in which doubts have been raised about the frequentist guarantees for common procedures. We ask the critic to assess the properties of these common procedures, and of some existing alternatives.

Our first application is to the returns to schooling. Bound et al. (1995), Staiger and Stock (1997), and others raise concerns about the bias of one of Angrist and Krueger's (1991) two-stage least squares (2SLS) estimators, and the coverage of the associated confidence interval, due to the use of many weak instruments. The critic agrees with both concerns. The critic is also able to evaluate various confidence procedures that have been proposed for use in situations with possibly weak instruments. Because the critic's prior and, hence, posterior, do not respect the instrumental variable (IV) model's overidentifying restrictions, the critic doubts the coverage of procedures whose guarantees rely on such restrictions, and has less doubt about the coverage of some other procedures.

Our second application is to inference in randomized controlled trials. Young (2019) argues that many studies use inferential procedures vulnerable to influential, high leverage observations and other problems that can cause size distortions. The critic raises similar

doubts, concluding that many of the inferential procedures in Young's (2019) applications are likely to undercover. Though the critic tends to agree with Young (2019) that leverage is associated with undercoverage, the critic also picks out examples of procedures they think cover well despite high leverage, and that cover poorly despite low leverage. In addition to exposing these nuances, the critic is able to answer novel questions about the performance of an alternative procedure that has been proposed for these same settings.

Though flexible and convenient in many respects, the Dirichlet process prior is not suitable for settings in which the analysis calls for smoothing (e.g., interpolating) between similar observations. For such situations, we propose Dirichlet process mixtures as an alternative prior class. We illustrate this proposal with a third application to the incumbency advantage in US House elections, for which Lee (2008) adopts a regression discontinuity design. Gelman and Imbens (2019) argue that estimation and inference based on global polynomials is less reliable than that based on local polynomials. The critic agrees with these conclusions, and also provides an assessment of alternative procedures that have been proposed more recently.

We emphasize two aspects of our approach. First, throughout our setup and applications, we take as given that some statistical parameter, defined as a functional of the DGP, is of interest to the audience. The critic's assessment is not generally informative about the economic interpretation of such a parameter. Second, because the performance of a given procedure can depend on the true DGP, a guarantee may hold in some settings and fail in others. The critic's assessment depends on the setting both through the choice of prior belief and because, even if a default prior is used, the critic's posterior belief depends on the data.

One may view the critic as a means of automating a probabilistic choice over simulation designs. Our work therefore relates to the literature on simulation design for the social sciences. Blair et al. (2019) recommend describing research designs in such a way as to facilitate the use of simulation to diagnose the properties of procedures. Huber et al. (2013), Lechner and Wunsch (2013), Busso et al. (2014), and Knaus et al. (2021) consider empirically calibrated simulations for evaluating methods of causal inference. Advani et al. (2019) find that a nonparametric bootstrap can yield more accurate finite-sample conclusions than other proposed designs (see also Ferman 2025). Athey et al. (2024) propose to automate the construction of synthetic data by using generative adversarial networks to ensure similarity

between the synthetic data and the sample (see also Schuler et al. 2017; Parikh et al. 2022).

Our work differs in taking a Bayesian perspective. This perspective allows the critic to assess frequentist properties probabilistically, taking into account that the true DGP is itself unknown. In this respect, our work relates to a literature on the Bayesian interpretation of frequentist procedures (e.g., Sellke et al. 2001; Müller and Norets 2016). Schorfheide (2000) uses a Bayesian mixture model to assess the loss from applying a specific economic model. Schorfheide and You (forthcoming) use a hierarchical Bayesian model to perform a meta-analysis of related studies and evaluate the probability that reported inference procedures cover a common cross-study parameter. We instead propose an abstract approach to Bayesian evaluation of a class of frequentist properties. Our approach applies standard ideas in nonparametric Bayesian inference, but in a new way, aiming at inference on frequentist properties of procedures, rather than directly on the parameter of interest.

Our work relates to a long tradition in statistical decision theory that treats frequentist desiderata as a means of ensuring that an audience can trust some implication, use, or interpretation of a reported statistic (see, e.g., discussions in Savage, 1954, Chapter 10.2; Pratt, 1965; Efron, 1986, Section 6; and Armstrong et al., forthcoming, Section 5.1). We discuss connections to this tradition in the text and appendices.

The remainder of the paper proceeds as follows. Section 2 defines the setting, explains the role of the critic, and develops its properties. Section 3 proposes a default implementation based on the Bayesian bootstrap and presents a warm-up application using data from a get-out-the-vote experiment. Section 4 presents applications to instrumental variables estimation of the returns to schooling and to inference in randomized controlled trials. Section 5 proposes a default class of smoothing priors and an application to incumbency advantage in US House elections. Section 6 concludes the main text. An Appendix presents proofs of formal statements made in the main text. An Online Appendix presents additional supporting analysis and findings.

2 The Analyst and the Critic

2.1 The Analyst and Their Procedure

An analyst observes a sample of N observations $X_1, ..., X_N$ drawn iid from an unknown **data-generating process (DGP)** P with support $\mathcal{X}$, that is $P \in \Delta(\mathcal{X})$. The analyst uses the **sample** $X = (X_1, ..., X_N) \in \mathcal{X}^N$ to implement a **procedure** $T : \mathcal{X}^N \to \mathcal{T}$ with range $\mathcal{T}$. The analyst reports the output $T(X)$ of the procedure to an audience. The audience is interested in a **parameter** which can be expressed as a known functional $\theta : \Delta(\mathcal{X}) \to \Theta$ of the DGP, where we sometimes shorthand $\theta(P)$ as θ.

The audience would like to trust that, under the true DGP P, the (random) output $T(X)$ of the procedure has some known relationship to the true value $\theta(P)$ of the parameter. To describe such a relationship, we define a **property** of the procedure, which we write as $M_P(T, \theta)$. The property depends on the procedure, which is a function of the sample; the parameter, which is a functional of the DGP; and on the DGP itself. We focus on scalar-valued properties that can be expressed as the expectation of some function $m : \mathcal{T} \times \Theta \to \mathbb{R}$ of the procedure and the parameter,

$$M_P(T, \theta) = \mathrm{E}_P[m(T(X), \theta(P))] = \mathrm{E}_{X \sim P^N}[m(T(X), \theta(P))],$$

where we take $\mathrm{E}_P = \mathrm{E}_{X \sim P^N}$ to denote the expectation when X is drawn iid according to P.

Example. (Bias.) Suppose that the procedure is an estimator for the parameter, and that both are real-valued, so that $\mathcal{T}, \Theta \subseteq \mathbb{R}$. Then the **bias** of the procedure is the property

$$M_P(T, \theta) = \mathrm{E}_P[T(X) - \theta(P)]. \qquad \triangle$$

Example. (Coverage.) Suppose that the procedure is a confidence set for the parameter, which is again real-valued, so that $\mathcal{T} \subseteq 2^{\mathbb{R}}$. Then the **coverage** of the procedure is $\Pr_P\{\theta(P) \in T(X)\} = \mathrm{E}_P[1\{\theta(P) \in T(X)\}]$, and its coverage relative to some desired

nominal level μ is the property

$$M_P(T,\theta) = \Pr_P\{\theta(P) \in T(X)\} - \mu = \mathrm{E}_P[1\{\theta(P) \in T(X)\} - \mu]. \qquad \triangle$$

In these examples, common frequentist desiderata (e.g., unbiasedness, exact coverage) require that the property take the value zero. As such exact finite-sample performance is rare outside of special cases, we instead define a frequentist **guarantee** as a claim that the property is within some tolerance $\eta \geq 0$ of the desired level, i.e., that

$$M_P(T,\theta) \in \mathcal{M}_\eta,$$

where $\mathcal{M}_\eta = [-\eta, \eta]$ for two-sided guarantees (e.g., approximate unbiasedness), and $\mathcal{M}_\eta = [-\eta, \infty)$ for one-sided guarantees (e.g., a lower bound on coverage). Any such guarantee is equivalent to a claim that the true DGP P is a member of the class

$$\mathcal{P}^\eta = \{P \in \Delta(\mathcal{X}) : M_P(T,\theta) \in \mathcal{M}_\eta\}$$

under which the property satisfies the relevant tolerance η in the sample of size N.

If the audience knew the true DGP P, the analyst could check by simulation whether $P \in \mathcal{P}^\eta$. In that case, however, there would also be no need for the procedure, as $\theta(P)$ would also be known. Instead, we suppose that each audience member believes that some particular set of DGPs is plausible, with $\mathcal{P}$ denoting the union of these sets across audience members. The audience trusts the guarantee if $\mathcal{P} \subseteq \mathcal{P}^\eta$. We suppose that it is infeasible to evaluate the guarantee by exhaustive simulation, i.e., by checking whether $M_P(T,\theta) \in \mathcal{M}_\eta$ for each DGP in $\mathcal{P}$. Instead, the analyst turns to a Bayesian critic.

Remark 1. (Bayesian foundations of frequentist guarantees.) *There are several reasons why a Bayesian audience may value trusting a frequentist guarantee. These reasons relate to some classic justifications for frequentist desiderata (e.g., Savage, 1954; Pratt, 1965). Online Appendix A.1 connects these justifications to our setup.*

Remark 2. (Extension to more general properties.) *We focus for concreteness on properties (e.g., bias and coverage) that can be expressed as the expectation of a scalar-valued function.*

The approach extends naturally to vector-valued properties (e.g., bias of an estimator for a vector-valued parameter), and to properties that can be expressed as more general functionals such as quantiles (e.g., median bias).

2.2 The Critic and Their Posterior

A Bayesian **critic** holds a **prior** belief π over possible DGPs, $\pi \in \Delta(\Delta(\mathcal{X}))$. After observing the analyst's sample, the critic updates their prior to form a **posterior** belief $\pi(\cdot|X) \in \Delta(\Delta(\mathcal{X}))$. We write the prior and posterior beliefs about the DGP itself as $\pi(P)$ and $\pi(P|X)$ respectively. The critic's posterior belief $\pi(P|X)$ about the DGP describes the critic's view of which DGPs, among the a priori plausible DGPs, are most consistent with the sample.

The critic's belief about the DGP implies beliefs about all functionals of the DGP. We write the prior and posterior beliefs about a functional $f : \Delta(\mathcal{X}) \to \mathcal{F}$ as $\pi(f(P)) \in \Delta(\mathcal{F})$ and $\pi(f(P)|X) \in \Delta(\mathcal{F})$ respectively. Standard Bayesian inference focuses on the posterior belief $\pi(\theta(P)|X)$ about the parameter. Instead, we focus on the critic's posterior belief

$$\pi(M_P(T,\theta)|X) = \pi\left(\mathrm{E}_{\tilde{X}\sim P^N}\left[m\left(T\left(\tilde{X}\right),\theta(P)\right)\right]|X\right)$$

about the property.

The critic's belief about the property implies a belief about the guarantee. The critic's **doubt** $\pi(M_P(T,\theta) \notin \mathcal{M}_\eta|X) = 1 - \pi(\mathcal{P}^\eta|X)$ describes the critic's assessment, given the sample, of the probability that the property fails to meet the desired tolerance.

If the critic is a plausible member of the analyst's audience, or has beliefs plausibly close to those of an audience member, then a doubtful critic implies that the guarantee fails in the sense that at least some audience member does not trust the guarantee. In such a case, the audience cannot rely on the guarantee for interpreting the procedure. Online Appendix A.2 sharpens these statements (see also Remark 1).

Example. (Bias.) For a guarantee of approximate unbiasedness, the doubt is

$$\pi(M_P(T,\theta) \notin \mathcal{M}_\eta|X) = \pi\left(\left|\mathrm{E}_{\tilde{X}\sim P^N}\left[T\left(\tilde{X}\right) - \theta(P)\right]\right| > \eta|X\right).$$

Suppose that the data are real-valued and the parameter is a known function of the population mean, $\theta(P) = g(\mathrm{E}_P[X])$ for known $g(\cdot)$. Suppose further that the procedure is the plug-in estimator $T(X) = g\left(\overline{X}\right)$ for $\overline{X}$ the sample mean.

If $g(\cdot)$ is linear then the bias $M_P(T, \theta) = \mathrm{E}_P[T(X) - \theta(P)]$ is zero whenever $\mathrm{E}_P[X]$ exists. Thus, so long as the critic is certain that $\mathrm{E}_P[X]$ exists, they have no doubt: $\pi(M_P(T, \theta) \notin \mathcal{M}_\eta | X) = 0$ for all $\eta \geq 0$.

If $g(\cdot)$ is instead nonlinear but continuously differentiable, then the bias is asymptotically small under regularity conditions (see, e.g., van der Vaart 1998, Section 3.5). The actual bias $M_P(T, \theta)$ in a sample of size N depends on details of the DGP P and the function $g(\cdot)$. If the critic doubts the guarantee of approximate unbiasedness, then if the audience includes members with beliefs similar to those of the critic, the audience cannot rely on the guarantee. △

If the critic doubts the guarantee, then, the audience cannot rely on it. The converse need not hold, in the sense that the critic might trust the guarantee, but some audience members might still doubt it. A special case in which the critic can validate, as well as invalidate, a guarantee is where all audience members share the critic's belief. But in such a case, the analyst could simply report the critic's posterior belief $\pi(\theta(P)|X)$ about the parameter of interest, and thus avoid the need for frequentist statistics altogether!

Remark 3. (Use of a posterior assessment.) *Because the critic's assessment is based on their posterior, we should think of the critic as representing either an audience member who is informed about the setting or, alternatively, what an uninformed audience member would believe given access to the data (again see Online Appendix A.2).*

Remark 4. (Restriction to iid data.) *We restrict attention to settings with iid observations. This restriction permits settings in which each observation is itself a cluster of multiple units, with each observation (cluster) drawn iid from some distribution. This restriction precludes settings with richer dependencies such as those often encountered in time-series, spatial, or network analysis. In principle it is possible to incorporate such dependencies by considering a suitably rich class of DGPs, provided one can specify a suitable prior on such a class (see, e.g., Hoff et al., 2002; Gelfand et al., 2005; Fox et al., 2011).*

Remark 5. (Restriction to statistical parameters.) *We restrict attention to parameters defined as functionals of the DGP. Such parameters are purely "statistical" in the sense that their values are known if the DGP is known. This restriction permits certain treatments of partially identified parameters, for example where the procedure returns a confidence set, the parameter is an identified set, and the property is coverage. This restriction precludes using the critic to assess dimensions of what Andrews et al. (forthcoming) term "econometric" misspecification, such as whether a given statistical parameter (say, a population regression coefficient) has a desired economic interpretation (say, as a causal effect). In principle, with a sufficiently rich prior (defined on the space* $\Theta \times \Delta(\mathcal{X})$*), the critic can make such assessments. In practice, we do not know of convenient default ways to specify such priors, so we require that the parameter is known if the DGP is known.*

Remark 6. (Connection to uniform asymptotic guarantees.) *Frequentist guarantees are often motivated by uniform asymptotic arguments which show that, for a class of DGPs* $\mathcal{P}^U$ *satisfying certain regularity conditions, and all* $\eta > 0$*, the guarantee holds uniformly in sufficiently large samples (i.e., that* $\mathcal{P}^U \subseteq \mathcal{P}^\eta$ *for* N *sufficiently large). If the critic doubts the guarantee, i.e., if* $\pi(\mathcal{P}^\eta|X) < 1$*, then by laws of probability it follows that either* $\pi\left(\mathcal{P}^U|X\right) < 1$*,* $\pi\left(\mathcal{P}^\eta|X, \mathcal{P}^U\right) < 1$*, or both. In the former case, the critic doubts that the regularity conditions hold at the true DGP. In the latter case, the critic doubts that the guarantee holds, in the available sample, even if the DGP is sufficiently regular, i.e., the critic doubts that the asymptotic results have "kicked in" at the given sample size.*

2.3 No Asymptotic False Alarms

Because it relies on Bayes' rule, the critic's posterior belief is the best possible assessment of any property given the critic's prior. But, as with any judgment under uncertainty, it may still be wrong. Here we ask whether, in a large-sample limit in which a given frequentist guarantee holds, the critic may incorrectly assess a failure of the guarantee. We show conditions under which this does not happen. These conditions include that the critic's posterior approaches the true DGP in the large-sample limit. Although our primary focus is on finite-sample assessments, we think the finding here is reassuring to an analyst who is inclined to

trust large-sample approximations.

We first formalize the idea of a large-sample frequentist guarantee for the analyst's procedure. We imagine a sequence of sample sizes N, which will tend to infinity. In the sample of size N the researcher will apply procedure T_N, and is interested in property $M_{N,P}(T_N, \theta) = \mathrm{E}_P[m_N(T_N, \theta)]$. In keeping with conventional frequentist practice, we suppose that the procedure sequence T_N satisfies a frequentist guarantee uniformly on a (potentially sample-size dependent) class $\mathcal{P}_N^U$ of data distributions.

Assumption 1. (Uniform asymptotic guarantee.) *The sequence $\{T_N\}_{N=1}^{\infty}$ satisfies a uniform asymptotic guarantee on a non-decreasing sequence of sets $\mathcal{P}_N^U$, in the sense that either*

$$\limsup_{N\to\infty} \sup_{P\in\mathcal{P}_N^U} |M_{N,P}(T_N, \theta)| = 0 \tag{1}$$

or

$$\liminf_{N\to\infty} \inf_{P\in\mathcal{P}_N^U} M_{N,P}(T_N, \theta) \geq 0. \tag{2}$$

Letting $\mathcal{P}_N^\eta$ be the analog of $\mathcal{P}^\eta$ in the sample of size N, Assumption 1 implies that for any $\eta > 0$, $\mathcal{P}_N^U$ is nested by $\mathcal{P}_N^\eta$ for N sufficiently large.

We next formalize the idea of a large-sample frequentist guarantee for the critic's posterior. Let $\mathcal{P}^\pi \subseteq \Delta(\mathcal{X})$ be a set which contains the support of the critic's posterior, in the sense that $\pi\left(\mathcal{P}^\pi | X^N\right) = 1$ for all N and every $X^N = (X_1, ..., X_N)$. Let τ be a topology on $\mathcal{P}^\pi$. We assume that the posterior is ($\tau-$)consistent under any fixed $P_0 \in \mathcal{P}^\pi$.

Assumption 2. (Posterior consistency.) *For any $P_0 \in \mathcal{P}^\pi$ with $X_1, ..., X_N \overset{iid}{\sim} P_0$ for all N, and any τ-open set $\mathcal{U} \subseteq \mathcal{P}^\pi$ containing P_0,*

$$\pi\left(\mathcal{U} | X^N\right) \to_p 1 \text{ as } N \to \infty.$$

Sufficient conditions for posterior consistency with respect to a given topology τ are given in, for example, Ghosal and van der Vaart (2017, Chapter 6).

Under posterior consistency and additional conditions, the following result establishes that, in the large-sample limit, the critic accepts the validity of the uniform guarantee.

Proposition 1. (No asymptotic false alarms.) *Suppose that Assumptions 1 and 2 hold. For any P_0 in the τ-interior of $\mathcal{P}_N^U \cap \mathcal{P}^\pi$ for some N, with $X_1, ..., X_N \overset{iid}{\sim} P_0$ for all N, and any $\eta > 0$, if (1) holds and $\mathcal{M}_\eta = [-\eta, \eta]$, or if (2) holds and $\mathcal{M}_\eta = [-\eta, \infty)$,*

$$\pi\left(M_{N,P}\left(T_N, \theta\right) \notin \mathcal{M}_\eta | X\right) = 1 - \pi\left(\mathcal{P}_N^\eta | X\right) \to_p 0$$

Proposition 1 shows that, asymptotically, the critic doesn't raise "false alarms." In particular, for data-generating processes P_0 such that the frequentist guarantee eventually holds on a neighborhood of P_0, the critic eventually concludes that violations of the guarantee are negligible. Put differently, if the frequentist guarantee holds in a sufficiently large sample, then a critic who has seen a sufficiently large sample will not doubt the guarantee at that sample size.

Online Appendix Proposition 1 further shows a sense in which the converse also holds: For DGPs P_0 where the guarantee is eventually violated on a neighborhood, the critic eventually concludes this as well.

2.4 Numerical Evaluation by the Bayesian Critic

Given a method of drawing from the critic's posterior distribution on DGPs, and a method of sampling data from any given DGP, it is straightforward to obtain draws from the critic's posterior belief about the value of the property under the true DGP. Specifically, given the data X, we form the posterior $\pi\left(P|X\right) \in \Delta\left(\Delta\left(\mathcal{X}\right)\right)$ by Bayes' rule. We then repeatedly draw DGPs $P_d \in \pi\left(P|X\right)$. For each DGP draw P_d, we compute the value $\theta_d = \theta\left(P_d\right)$ of the parameter. We then draw S samples of size N from P_d. Averaging over these samples gives us an estimate of $M_{P_d}\left(T, \theta\right)$, and by taking the empirical distribution of these estimates across D draws we obtain an estimate of $\pi\left(M_P\left(T, \theta\right)|X\right)$. Algorithm 1 summarizes these steps.

Algorithm 1 Numerical evaluation by the Bayesian critic

- Given a sample X and prior π, for each $d \in \{1, ..., D\}$, do:
  - Draw $P_d \sim \pi(P|X)$.
    - Calculate $\theta_d = \theta(P_d)$.
    - For each $s \in \{1, ..., S\}$, do:
      - Sample $X_{d,s}$ so that $X_{d,s,i} \overset{iid}{\sim} P_d$ for $i \in \{1, ..., N\}$.
      - Calculate $T(X_{d,s})$.
      - Calculate $m_{d,s} = m(T(X_{d,s}), \theta_d)$.
    - Calculate $\hat{M}_d = \frac{1}{S} \sum_{s=1}^{S} m_{d,s}$.
- This algorithm yields approximate draws $\left\{\hat{M}_1, ..., \hat{M}_D\right\}$ from the critic's posterior belief about the value of the property under the true DGP.

Algorithm 1 describes a procedure for approximating the critic's posterior belief about the property. The algorithm uses finite approximations along two dimensions. First, the algorithm draws only a finite number D of distributions from the critic's posterior belief. Second, for any given distribution P_d, the algorithm samples only a finite number S of datasets. Accuracy along each dimension can be controlled and measured in standard ways (e.g., Robert and Casella, 2004, Chapter 4; Rainforth et al., 2018).

Remark 7. (Changing the sample size.) *Algorithm 1 covers the leading case where the critic evaluates the property in a sample of the same size as the original,* $N = \dim(X)$. *To ask the critic to evaluate the same property in a sample of any other size* N'*, simply replace* N *with* N' *in Algorithm 1.*

3 A Critic with a Default Prior

In this section we propose the low-precision Dirichlet process as a convenient and flexible default prior class for the critic. We discuss the properties of the prior class, explain the resulting sampling algorithm, and illustrate the use of the prior class with a warm-up application to data from a get-out-the-vote experiment.

3.1 A Default Prior Class

Suppose that the critic's prior is in the Dirichlet process class, so that $\pi = \pi_{\alpha,Q} = DP(\alpha, Q)$, where the parameter $\alpha > 0$ controls the precision of the prior, and the centering measure $Q \in \Delta(\mathcal{X})$ controls its location. In this case, the critic's posterior on the DGP is

$$\pi_{\alpha,Q}(P|X) = DP\left(\alpha + N, \frac{\alpha}{\alpha + N}Q + \frac{N}{\alpha + N}\hat{P}_N\right)$$

where $\hat{P}_N$ is the empirical distribution of the sample X. In the limit as $\alpha \to \infty$, the prior becomes arbitrarily precise. In that limit, the posterior does not depend on the data, and instead concentrates on a single measure Q, making Algorithm 1 akin to a prespecified numerical simulation with a fixed distribution.

We focus instead on the limit as $\alpha \to 0$. In that limit, the prior becomes imprecise, and the posterior converges to the **Bayesian bootstrap distribution** $\pi^B(P|X) = DP\left(N, \hat{P}_N\right)$, which is centered at the empirical distribution of the sample.

Proposition 2. *If $M_P = M_P(T(\cdot), \theta(\cdot))$ is continuous in P with respect to convergence in distribution almost everywhere in the support of $DP\left(N, \hat{P}_N\right)$, then*

$$\pi_{\alpha,Q}(M_P|X) \xrightarrow{d} \pi^B(M_P|X)$$

as $\alpha \to 0$.

Proposition 2 is immediate from standard results on Dirichlet processes (e.g., Theorem 4.16 of Ghosal and van der Vaart, 2017) and the continuous mapping theorem.

In many situations, it is convenient to draw DGPs from the Bayesian bootstrap distribution. To sample $P_d \sim \pi^B(P|X)$, we draw weights for the N sample observations from a standard Dirichlet distribution; this can be done by normalizing standard exponential variates. Defining P_d as the resulting weighted empirical distribution, we can compute $\theta_d = \theta(P_d)$ by applying our estimation procedure to the weighted sample, draw samples $X_{d,s}$ by sampling (with replacement) from the weighted empirical distribution P_d, and proceed otherwise as in Algorithm 1.

Algorithm 2 Numerical evaluation by a critic with a Bayesian bootstrap posterior

- To draw $P_d \sim \pi^B(P|X)$, do:
  - Draw pseudoweights $V_d \in \mathbb{R}^N_{>0}$ as $V_{d,i} \overset{iid}{\sim} Exp(1)$ for $Exp(1)$ the standard exponential distribution.
  - Construct weights $W_d \in \Delta(\{1, ..., N\})$ as $W_{d,i} = V_{d,i} / \sum_j V_{d,j}$.
  - Define $P_d = P(X; W_d)$ as the empirical distribution of the sample X, weighted by W_d.
- To sample $X_{d,s} \sim P_d$, sample each $X_{d,s,i}$ independently from X according to sampling probabilities W_d.

In addition to being convenient to draw from, the Bayesian bootstrap distribution is also flexible in multiple respects.

Remark 8. (Posterior consistency of the Bayesian bootstrap.) *In the large sample setting of Section 2.3, the Bayesian bootstrap distribution is consistent with respect to the topology τ of convergence in distribution (see e.g. Ghosal and van der Vaart 2017, Chapter 4.7).*[2] *Hence, for frequentist procedures which are uniformly asymptotically valid on τ-open sets of distributions $\mathcal{P}^U \subseteq \Delta(\mathcal{X})$, Proposition 1 implies that a critic using the Bayesian bootstrap does not asymptotically raise false alarms.*

Remark 9. (The Dirichlet process prior has large support.) *If the centering measure Q has support equal to the sample space $\mathcal{X}$, then the weak support of a Dirichlet process prior $\pi_{\alpha,Q}$ is the set $\Delta(\mathcal{X})$ of all distributions on $\mathcal{X}$ (see, e.g., Ghosal and van der Vaart 2017, Chapter 4.3).*[3] *In this sense such a prior is (weakly) non-dogmatic.*

Remark 10. (The Dirichlet process prior does not smooth.) *Under a Dirichlet process prior $\pi_{\alpha,Q}$, the posterior distribution $\pi_{\alpha,Q}(P(A)|X)$ of $P(A)$ for any event $A \subset \mathcal{X}$ depends on the data only through the sample size N and the share of the observations that are in A,*

[2] While the topology of convergence in distribution is too weak for some purposes, e.g., ensuring validity of the central limit theorem for unbounded variables, under given P_0 the Bayesian bootstrap distribution is also consistent in the stronger topology obtained by intersecting the topology of convergence in distribution with the Euclidian topology on $\phi(P) = E_P[h(X_i)]$ for any vector-valued $h(\cdot)$ such that $E_{P_0}\left[h(X_i)\, h(X_i)'\right]$ is finite.

[3] The weak support of a probability measure is the smallest set, closed with respect to the topology of convergence in distribution, that has probability one under the given measure.

i.e., through the tuple $\left(N, \frac{1}{N}\sum_{i=1}^{N} 1\{X_i \in A\}\right)$. *The posterior therefore does not depend on how "close" the observations not in A are to those in A. The failure of the posterior to "smooth" across similar observations can be desirable when the goal is to be "agnostic" about the structure of the data, but such agnosticism can be implausible in some situations, such as when the parameter of interest inherently suggests a desire to smooth. We return to such situations in Section 5.*

3.2 A Warm-up Get-out-the-vote Example

As a warm-up example, consider a get-out-the-vote experiment reported in Nickerson et al. (2006). In the experiment, $N = 16,181$ citizens are randomized between a treatment and control group. The treatment consists of a phone call encouraging the citizen to vote. For each citizen i, the analyst observes $X_i = (T_i, V_i)$, where $T_i \in \{0, 1\}$ indicates whether i receives treatment and $V_i \in \{0, 1\}$ indicates whether i votes.

One parameter of interest is the difference in the probability of voting between the treatment and control groups, $\theta^{TC}(P) = \mathrm{E}_P[V_i|T_i = 1] - \mathrm{E}_P[V_i|T_i = 0]$. The analyst considers a plug-in estimator $T^{TC}(X)$ given by the difference in sample frequencies. The analyst would like to guarantee that the bias of the estimator, $M_P = \mathrm{E}_P\left[T^{TC}(X) - \theta^{TC}(P)\right]$, is small. Of course, this bias is zero (if we condition on the estimator being well-defined), but the analyst may not know this theoretical fact. The analyst asks the critic to assess the likely bias.

Adopting a low-precision Dirichlet process prior as suggested in Section 3.1, the assessment by the critic is fully automated. The analyst need only select the number of draws D from the critic's posterior on the DGP, and the number of samples S from each of these draws. Algorithm 3 describes how we take a single draw d from the critic's posterior belief about the bias of the estimator.

Panel (a) of Figure 1 reports the critic's assessment of the bias of the estimator. Relative to the sample point estimate of 4.77 percentage points, the critic assesses that the bias is small. The critic's posterior expected bias is -0.01 percentage points, numerically close to its theoretical value. Although numerical imprecision means that the (numerical approximation to the) critic does not conclude that bias is always exactly zero, the critic's posterior

Algorithm 3 Taking a single draw from the critic's posterior belief about the bias of the estimator

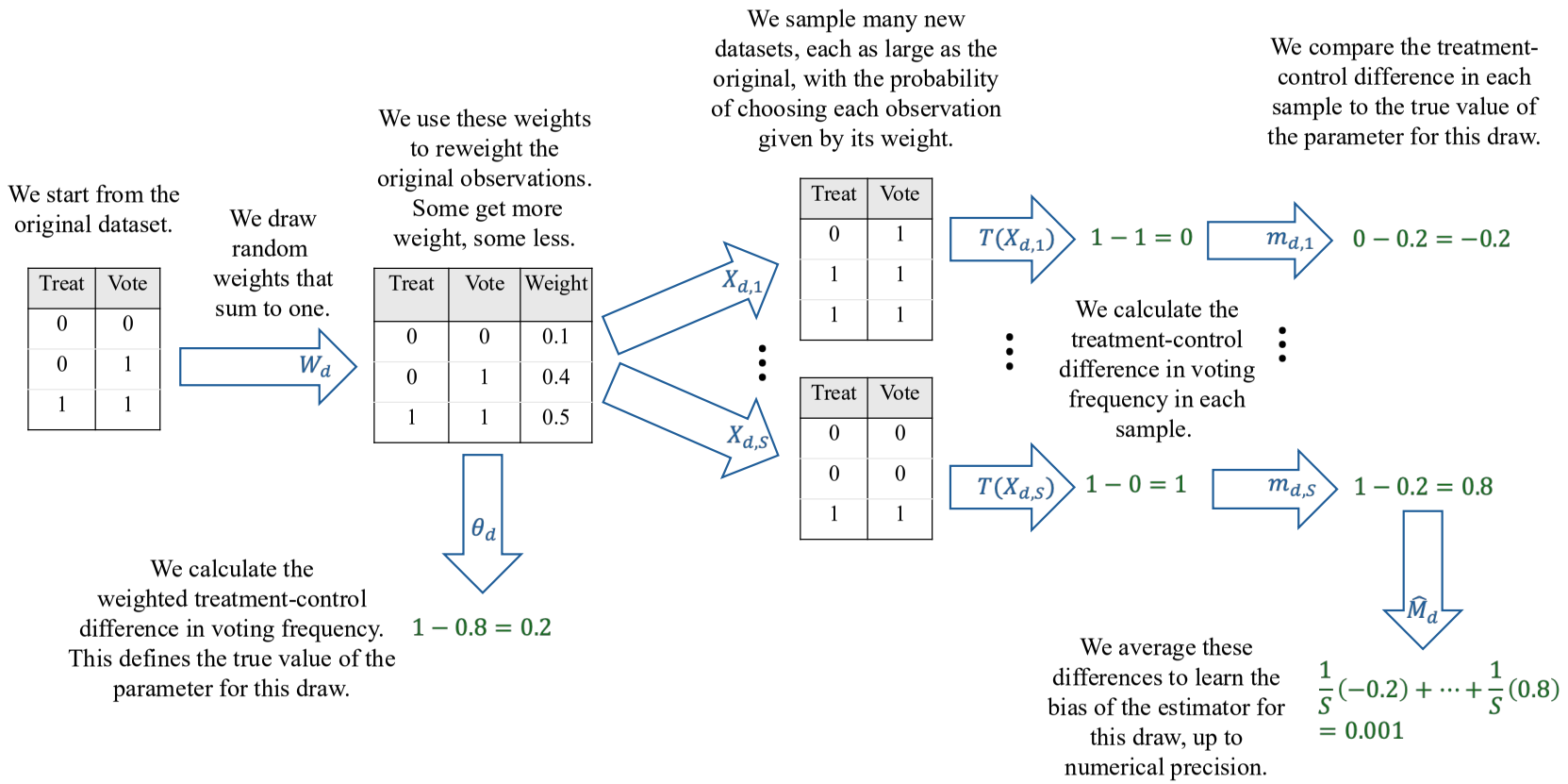
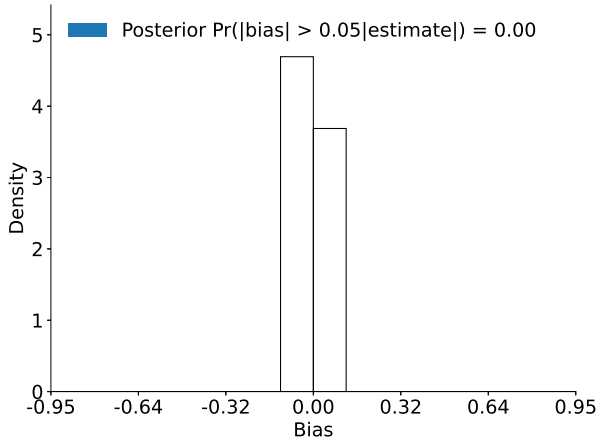
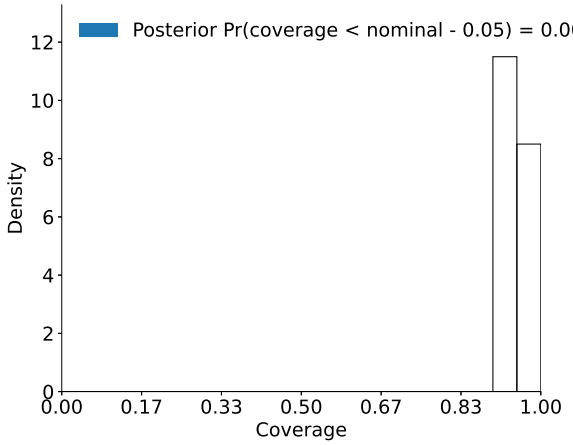


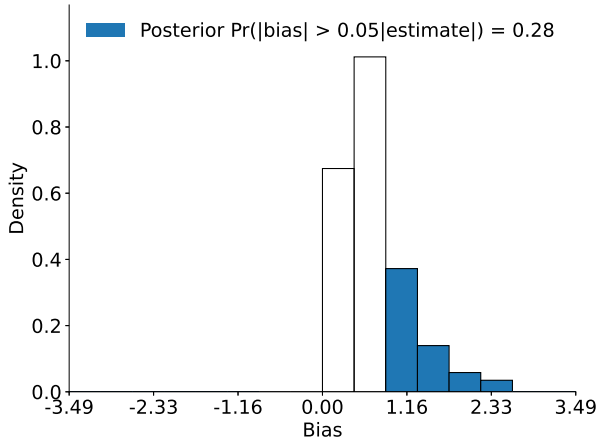
Figure 1: Critic's Assessment of Bias and Coverage in a Get-out-the-vote Example



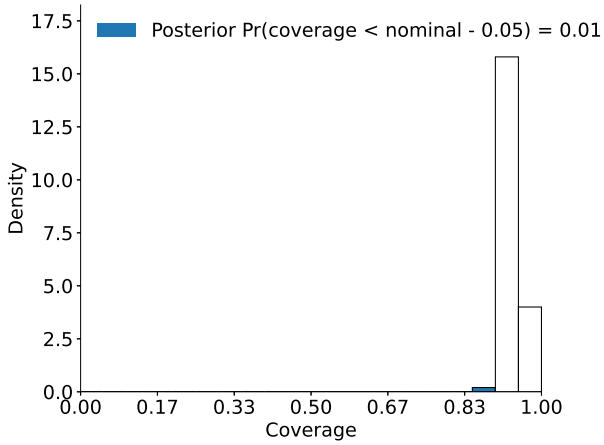
(a) Bias of the plug-in estimator of the treatment-control difference



(b) Coverage of the confidence interval for the treatment-control difference



(c) Bias of the plug-in estimator of the cost per vote



(d) Coverage of the confidence interval for the cost per vote

Note. Each plot shows a histogram of the critic's posterior distribution. For panels (a) and (b), the parameter of interest is the treatment-control difference. For panels (c) and (d), the parameter of interest is the cost per vote. Panels (a) and (c) show the posterior distributions of bias for the associated plug-in estimators, with shading denoting when bias is greater in magnitude than 0.05 times the point estimate. Panels (b) and (d) show the posterior distributions of coverage for the usual 95% confidence interval and the delta-method 95% confidence interval respectively, with shading denoting when coverage is more than 0.05 below nominal. All plots are based on applying Algorithm 2 using $D = 200$ draws from the posterior distribution and $S = 500$ samples from each draw.

distribution of bias is tightly concentrated around zero. For example, the critic assesses that the absolute value of the bias is less than 5 percent of the point estimate with probability 1.00, implying a doubt of 0.00 in this guarantee.

Asking the critic to assess the coverage of the associated confidence interval is just as easy. Panel (b) of Figure 1 reports the critic's assessment. The critic assesses that the usual confidence interval has approximately nominal coverage. The critic's posterior probability that coverage is more than five percentage points below nominal is 0.00. Again, this finding seems unsurprising to us, but might be reassuring to an analyst (or their audience).

Another parameter of interest is the cost per vote generated by a phone call, $\theta^K(P) = \frac{K}{\theta^{TC}(P)}$ for $K > 0$ the cost of a phone call. Given a cost per phone call of $K = \$0.83$, the plug-in estimated cost per vote is \$17.46. Plug-in procedures for this parameter, which is a nonlinear function of conditional probabilities, may perform poorly. While this fact is known theoretically (e.g., Andrews and Mikusheva, 2016) and discussed in some applications (e.g., Roberts and Schlenker, 2013, p. 2279), it may not be known to the analyst. Moreover, even an analyst well-versed in modern econometric theory will know that the performance of a given procedure depends on details of the setting, such as the true data distribution and the sample size, that can be hard to account for using analytic theory alone.

Fortunately, the critic can readily assess the properties of procedures for the cost per vote. Panel (c) of Figure 1 reports the critic's assessment of the bias of the plug-in estimator.[4] In contrast to the plug-in estimator of the treatment-control difference, the critic assesses an economically meaningful bias for the plug-in estimator of the cost per vote. The posterior expected bias is \$0.76, and the posterior probability that the absolute bias is more than 5 percent of the point estimate is 0.28.

Panel (d) of Figure 1 reports the critic's assessment of the coverage of the delta-method confidence interval for the cost per vote. In contrast to the confidence interval for the treatment-control difference, the critic has nonzero (though still small) doubt of 0.01 in the guarantee that coverage is no more than 5 percentage points below nominal.

Taking the critic's findings together, the analyst concludes that, if the critic is (close to)

[4]To avoid incoherent values for the cost per vote, we suppose that, under the critic's prior, $\theta^{TC}(P) > \underline{\theta}$ with probability one, where $\underline{\theta} \in (0, 1)$ is small (0.005). To impose this constraint, we modify Algorithm 2 to discard any draw d (0.00% of the total) or sample s (0.02% of the total) that violates it.

a plausible member of the analyst's audience, then the audience cannot trust frequentist guarantees for the procedures concerning the cost per vote. By contrast, the analyst does not find a reason for such an audience to doubt the analogous guarantees for the procedures concerning the treatment-control difference in voting probabilities.

Remark 11. (Agnosticism of the critic.) *Nickerson et al. (2006) report details of their experimental design including the method of randomization. The critic's low-precision Dirichlet process prior does not incorporate such details, so the critic represents an audience member who is a priori unsure of the experimental design. With an appropriate choice of prior, it is possible to model a critic who is certain of the experimental design, though with the consequence of making the approach less automated than the one we adopt here. We return to this possibility in Section 4.2.*

Remark 12. (Uses of the critic.) *The warm-up example illustrates several possible uses of the critic:*

1. Pilot analysis. *The analyst has access to data from a pilot of their get-out-the-vote experiment and plans to collect data from a larger, final sample. The analyst believes the DGP will be similar between the pilot and final sample. The analyst wishes to determine which procedures to include in their pre-analysis plan, and how large of a final sample to collect. Because the critic can evaluate a given procedure at any sample size, the critic can provide feedback on both questions.*[5]

2. Diagnostic statistics. *The analyst wishes to evaluate the guarantees for the procedure they use in the final data from the get-out-the-vote experiment. If the critic casts doubt on a procedure's guarantees, the analyst's audience may reasonably doubt the procedure as well. Anticipating this possibility, the analyst may wish to apply the critic to a hold-out sample to avoid the well-known consequences of post-hoc procedure selection.*

3. Methodological recommendations. *The analyst wishes to recommend procedures for use in future get-out-the-vote experiments. If the critic casts doubt on a procedure's guarantees, that procedure may not be trustworthy for use in analyzing future experiments.*

[5] *This use case relates to a literature on interim monitoring of clinical trials, reviewed for example by Saville et al. (2015), which uses Bayesian methods to probabilistically evaluate prospective trial outcomes, rather than to assess the frequentist properties of procedures.*

4 Applications with the Default Prior

4.1 Returns to Schooling

Angrist and Krueger (1991, Table V) estimate the returns to schooling in a homogeneous, linear model. The outcome is the log of earnings. The regressor of interest is the number of years of schooling. One estimator is an ordinary least squares (OLS) estimator (Angrist and Krueger, 1991, Table V, column 5). Another estimator is a two-stage least squares (2SLS) estimator, using as excluded instruments interactions between quarter-of-birth and year-of-birth (Angrist and Krueger, 1991, Table V, column 6). Inference is conducted using Eicker-Huber-White (EHW) standard errors. An extensive literature beginning with Bound et al. (1995) and Staiger and Stock (1997) argues that, because there are many instruments and the instruments are only weakly related to the endogenous regressor, the 2SLS estimator may be biased and its EHW confidence interval may undercover.

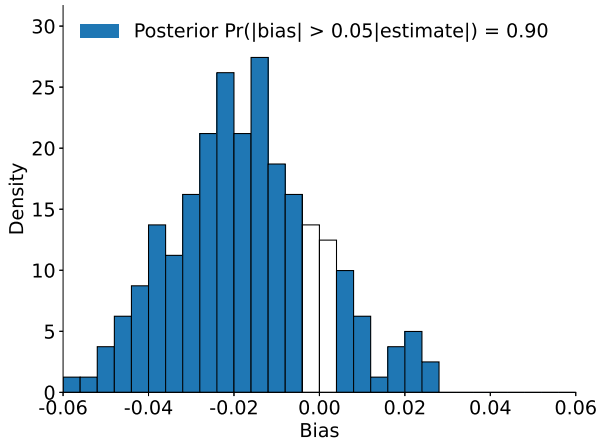
We ask the default critic to evaluate the bias of each of the two estimators and the coverage of their associated EHW confidence intervals. We take the statistical parameter of interest to be the value of the 2SLS estimator under the true DGP.

Panels (a) and (b) of Figure 2 report, respectively, the critic's posterior belief about the bias of each estimator relative to this parameter. For the OLS estimator, the posterior expected bias is -0.017, as compared to the 2SLS point estimate of 0.081. The posterior probability that the absolute bias of the OLS estimator is more than 5 percent of the 2SLS estimate is 0.90. For the 2SLS estimator, the posterior expected bias is -0.002, and the posterior probability that the absolute bias is more than 5 percent of the 2SLS estimate is 0.26.[6]

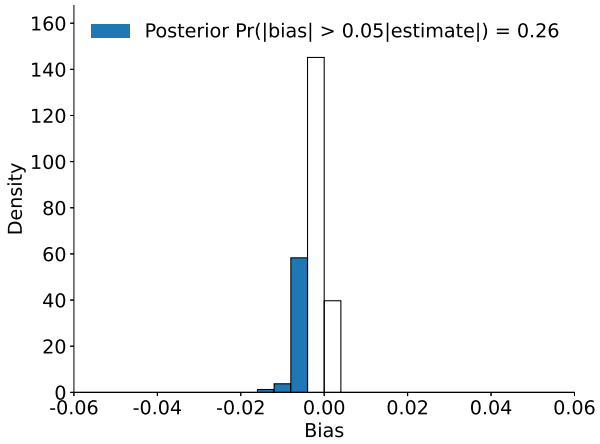
Panels (c) and (d) of Figure 2 report, respectively, the critic's posterior belief about the coverage of each EHW confidence interval. For the OLS estimator, the critic assesses that the EHW confidence interval severely undercovers, concluding with near certainty that coverage is more than five percentage points below the nominal 95% level. For the 2SLS estimator, the critic likewise doubts the coverage of the interval, assessing that the posterior probability

[6]The posterior expected bias of the 2SLS estimator is 14% of the posterior expected bias of the OLS estimator. For comparison, under their assumptions Staiger and Stock (1997, p. 581) estimate that the worst-case asymptotic bias of the 2SLS estimator is 21% of the bias of the OLS estimator.

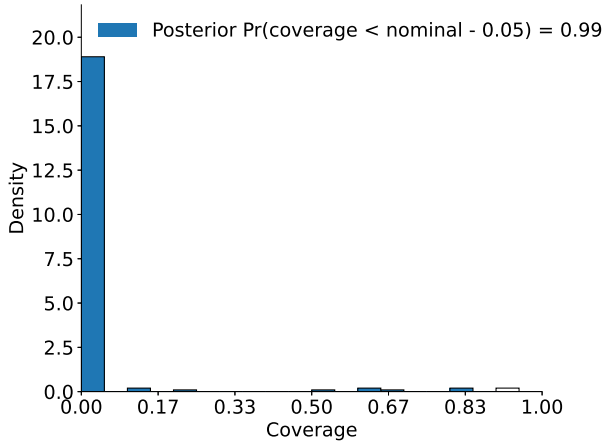
Figure 2: Critic's Assessment of Bias and Coverage in a Returns to Schooling Application (Angrist and Krueger, 1991)



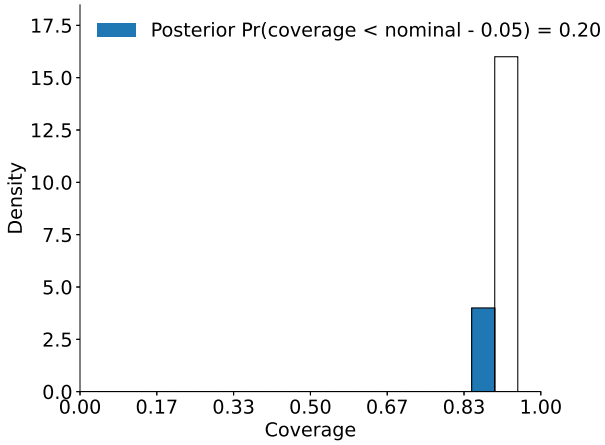
(a) Bias of the OLS estimator



(b) Bias of the 2SLS estimator



(c) Coverage of the OLS confidence interval



(d) Coverage of the 2SLS confidence interval

Note. Each plot shows a histogram of the critic's posterior distribution, taking the parameter of interest to be the value of the 2SLS estimator under the true DGP. Panels (a) and (b) show the posterior distribution of bias for the OLS and 2SLS estimator, respectively, with shading denoting when bias is greater in magnitude than 0.05 times the 2SLS point estimate. Panels (c) and (d) show the posterior distribution of coverage for a 95% EHW confidence interval corresponding to the OLS and 2SLS estimator, respectively, with shading denoting when coverage is more than 0.05 below nominal. All plots are based on applying Algorithm 2 using $D = 200$ draws from the posterior distribution and $S = 500$ samples from each draw.

that coverage is more than five percentage points below the nominal 95% level is 0.20. The critic therefore confirms the concerns in the literature regarding the coverage of conventional confidence intervals in this setting.

In addition to assessing the properties of the original procedures used by Angrist and Krueger (1991), the critic can assess the properties of alternative procedures. We focus on the three alternative inferential procedures studied in Kleibergen and Zhan (2025) for which critical values are available in closed form. Figure 3 reports the critic's assessment of each of these three alternative inferential procedures.

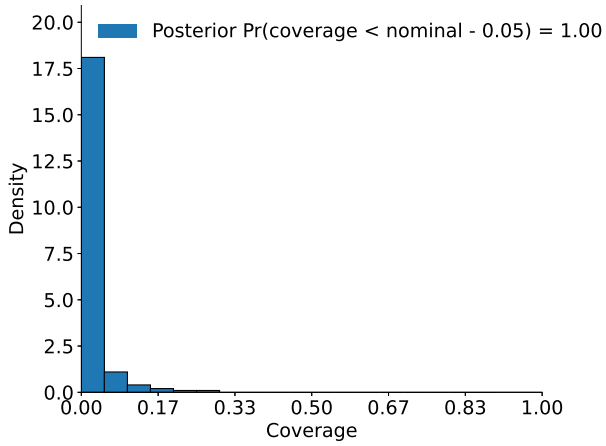
Panel (a) of Figure 3 reports the critic's posterior belief about the coverage of the S-statistic procedure of Stock and Wright (2000). The critic assesses that this procedure severely undercovers, concluding with near certainty that coverage is more than 5 percentage points below the nominal 95% level, and with probability 0.91 that it is more than 90 percentage points below.

The reason for the critic's doubt about the S-statistic procedure is instructive. Frequentist guarantees for the S-statistic procedure rely on assuming that the instrumental variables model's overidentifying restrictions hold, in the sense that the 2SLS estimand based on any subset of the instruments is equal to the 2SLS estimand based on the full set of instruments. Outside of knife-edge cases, the Dirichlet process prior and posterior place zero probability on DGPs satisfying such restrictions. Data generated from DGPs supported under the posterior tend to exhibit in-sample violations of the restrictions, such that the S-statistic procedure often returns an empty confidence set.

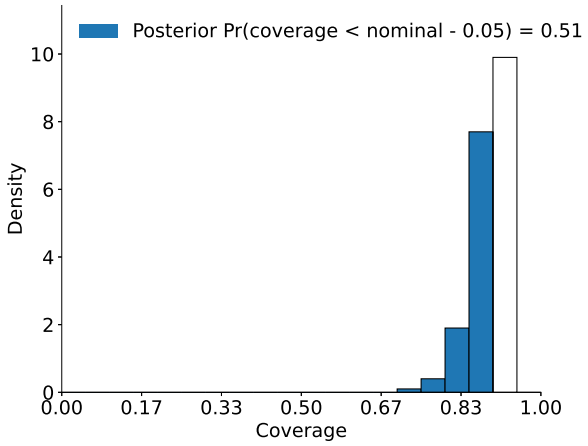
We regard this aspect of the default prior as a feature, because it seems to us unlikely that the overidentifying restrictions hold (i.e., are exactly satisfied) under the true DGP. An analyst whose audience is confident that the overidentifying restrictions do hold could alternatively query a critic whose prior imposes these restrictions, for example via the prior class considered in Schennach (2005).

Panel (b) of Figure 3 reports the critic's posterior belief about the coverage of the K-statistic procedure of Kleibergen (2005). The critic assesses that the probability that coverage is more than 5 percentage points below the nominal 95% level is 0.51. As with the S-statistic procedure, the K-statistic procedure's frequentist guarantees are based on the IV model's

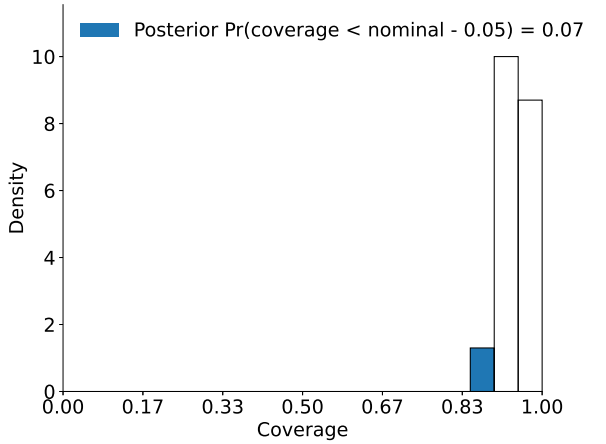
Figure 3: Critic's Assessment of Coverage of Alternative Procedures in a Returns to Schooling Application (Angrist and Krueger, 1991)



(a) Coverage of the S-statistic procedure of Stock and Wright (2000)



(b) Coverage of the K-statistic procedure of Kleibergen (2005)



(c) Coverage of the DRLM-statistic procedure of Kleibergen and Zhan (2025)

Note. Each plot shows a histogram of the critic's posterior distribution of coverage of a given 95% confidence procedure, taking the parameter of interest to be the value of the 2SLS estimator under the true DGP, with shading denoting when coverage is more than 0.05 below nominal. All plots are based on applying Algorithm 2 using $D = 200$ draws from the posterior distribution and $S = 500$ samples from each draw.

overidentifying restrictions.

Panel (c) of Figure 3 reports the critic's posterior belief about the coverage of the double robust Lagrange multiplier (DRLM)-statistic procedure of Kleibergen and Zhan (2025). The critic assesses that the probability that coverage is more than five percentage points below the nominal 95% level is 0.07. This improves substantially upon the coverage of the S-statistic and the K-statistic procedures. Unlike the S-statistic and the K-statistic procedures, frequentist guarantees for the DRLM-statistic procedure do not rely on the IV model's overidentifying restrictions.

Frequentist guarantees for the DRLM-statistic procedure treat as the parameter of interest the value of the continuous updating GMM (CUGMM) estimator under the true DGP. Online Appendix Table 1 shows that the critic assesses a probability of zero that the DRLM-statistic procedure has coverage more than 5 percentage points below nominal for this alternative parameter. Thus, the critic assesses that frequentist guarantees for this procedure are very likely to hold under the true DGP.

Taken together, these findings point to the value of the critic in assessing not only the concerns with the authors' original procedures, which are well-known, but also the likely performance of proposed alternatives in this particular setting, which is unknown.

4.2 Randomized Controlled Trials

Young (2019) studies the inferential procedures used in previous studies' analyses of data from randomized controlled trials (RCTs). Young (2019) shows simulation evidence that, especially in settings with high-leverage observations, common inferential procedures need not control size in the test of the null of no effect. Young (2019) then shows that, in a sample of published studies, randomization tests for the sharp null of no effect often yield different conclusions about statistical significance than the studies' original tests for the null of no effect.

We ask the critic to assess the coverage of the studies' original inference procedures. In each case, we take the parameter of interest to be the value of the given study's own estimator when applied to the true DGP. We use the replication files in Young (2018) and Andrews et al. (2025). We focus on those studies for which the necessary data are publicly available, and on

those procedures that, in the study's original implementation, accept weights.[7] We are left with 36 studies and 2,910 parameters, out of an original 53 studies and 4,044 parameters in Young (2019). Online Appendix Table 2 details the impact of each of our criteria on the sample of studies and parameters. Online Appendix Table 3 reproduces findings in Young (2019, Table V) for the set of studies and parameters in our sample.

Panel (a) of Table 1 summarizes the critic's doubt about coverage, which we define as the critic's posterior probability that the coverage of the study's original confidence procedure is more than 5 percentage points below nominal. We follow Young (2019, Table V) in averaging this posterior probability, and other statistics, weighting by the inverse of the number of parameters in the given study, so that each study in effect has equal weight. We also follow Young (2019, top of Table V) in grouping procedures into three terciles according to the average leverage of the highest-leverage observation, and in studying procedures with both one and five percent testing thresholds.

The critic assesses that, on average, the doubt is 0.13 for the study's original 99% confidence intervals, and 0.14 for the study's original 95% confidence intervals. Consistent with Young's (2019) emphasis on leverage, these average doubts increase, to 0.29 and 0.29 respectively, when we focus on specifications that have a high-leverage observation.

The first row of Figure 4 illustrates these findings. Each plot shows the Bayesian critic's posterior belief about coverage for one parameter's original 95% confidence interval. Panel (a) shows a case with low leverage. Panel (b) shows a case with high leverage. In the former case, the critic does not doubt coverage; in the latter case, the critic has substantial doubt.

While high leverage is associated with greater doubt about coverage, this association is not perfect. The second row of Figure 4 illustrates with examples that reverse the pattern in the first row. An analyst studying one of these settings could rely on the critic, rather than on leverage statistics or other rules-of-thumb, for an assessment of coverage.

In addition to assessing the properties of the authors' original procedures, the critic can assess the properties of alternative procedures. Simonsohn (2021) argues based on prior literature and simulation evidence that heteroskedasticity-robust (HC3; MacKinnon and White 1985) inference can address the shortcomings of the types of procedures studied by

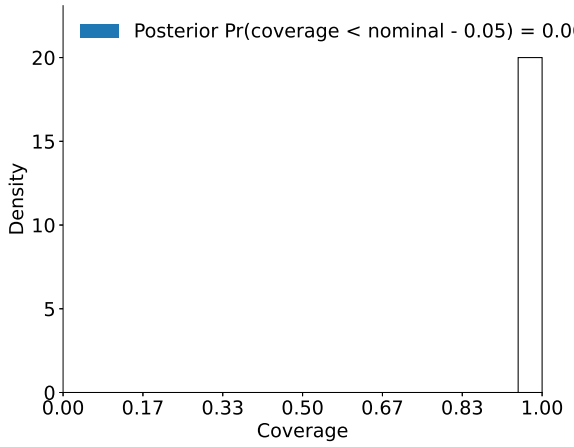
[7] We also exclude a handful of studies that involve a bootstrap or are memory-intensive.

Table 1: Critic's Assessment of Coverage in Randomized Controlled Trials (Young, 2019)

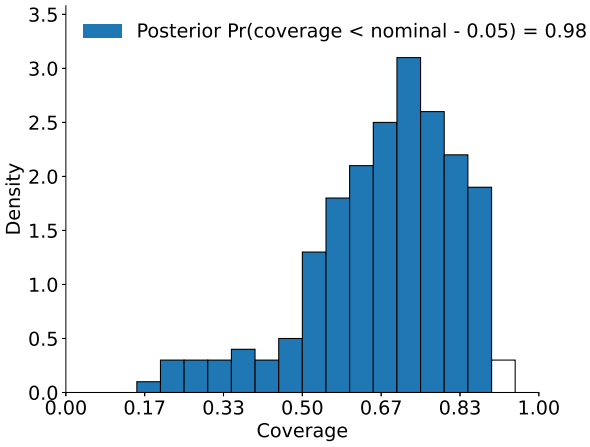
	All		Low leverage		Medium leverage		High leverage	
	99%	95%	99%	95%	99%	95%	99%	95%
Panel (a): Baseline								
Average posterior Pr(coverage < nominal - 0.05)								
Study's original procedure	0.13	0.14	0.04	0.04	0.07	0.10	0.29	0.29
(full analysis sample)	(36 studies)		(12 studies)		(12 studies)		(12 studies)	
Panel (b): Experiment with HC3								
Average posterior Pr(coverage < nominal - 0.05)								
Study's original procedure	0.05	0.07	0.04	0.04	0.12	0.17	0.00	0.00
HC3-robust inference	0.05	0.06	0.04	0.04	0.11	0.12	0.00	0.00
(restricted analysis sample)	(10 studies)		(5 studies)		(3 studies)		(2 studies)	

Note. The table summarizes the critic's doubt about coverage, defined as the critic's posterior probability that coverage is more than five percentage points below nominal, in a subset of the studies considered in Young (2019). We follow Young (2019) in averaging statistics weighting by the inverse of the number of parameters in a given study, and in classifying leverage (defined as the average, across parameters, of the leverage of the highest-leverage observation) by terciles in our analysis sample. The full analysis sample is restricted to those studies for which the necessary data are publicly available, and those procedures that, in the study's original implementation, accept weights (see Section 4.2 and Online Appendix Table 2). The restricted analysis sample further requires that procedures accept the option to compute HC3-robust confidence intervals. All values are based on applying Algorithm 2 using $D = 200$ draws from the posterior distribution and $S = 200$ samples from each draw.

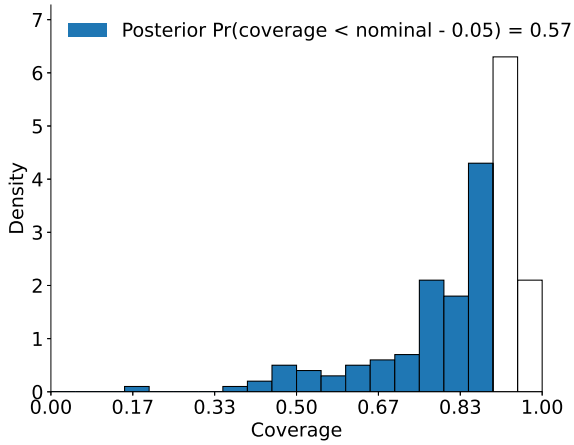
Figure 4: Critic's Assessment of Coverage in Selected Randomized Controlled Trials (Young, 2019)



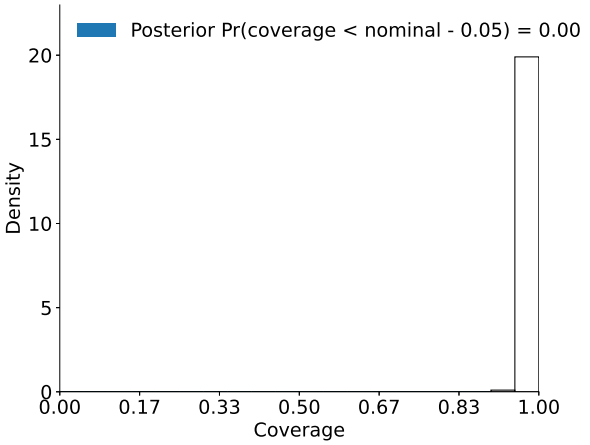
(a) Low leverage, low doubt about coverage



(b) High leverage, high doubt about coverage



(c) Low leverage, high doubt about coverage



(d) High leverage, low doubt about coverage

Note. Each plot shows a histogram of the critic's posterior distribution on the coverage of the given parameter's original 95% confidence interval, with shading denoting when coverage is more than 0.05 below nominal. Plots show examples of high and low leverage parameters with high and low doubt about coverage, where doubt is the posterior probability that coverage is more than 5 percentage points below nominal. We follow Young (2019) in classifying leverage by terciles in our analysis sample. All plots are based on applying Algorithm 2 using $D = 200$ draws from the posterior distribution and $S = 200$ samples from each draw.

Young (2019). We ask the critic to assess the likely performance of HC3-robust confidence intervals on the DGPs in the original settings.

Panel (b) of Table 1 summarizes the critic's assessment of both the original, and HC3-robust confidence intervals, for the subset of parameters for which the study's original procedure accepts an option to compute HC3-robust confidence intervals. For this subset of parameters, leverage, and doubt about coverage of the original confidence intervals, tend to be smaller than in the full set of parameters. For this subset of parameters, the critic's assessment of HC3-robust confidence intervals is similar to the critic's assessment of the studies' original confidence intervals.

Taken together, these findings show that the critic can deliver assessments that are customized to particular settings, and can also assess procedures not previously studied in these settings.

The default prior does not incorporate details of the experimental design (see Remark 11). For a subset of studies, Online Appendix Table 4 shows results for a prior which imposes the exact stratification and treatment assignment from the original sample, thus fully incorporating the design for procedures that do not use additional covariates (panel a), and partially incorporating the design for procedures that do use additional covariates (panel b).

5 A Critic with a Smoothing Prior

The Dirichlet process prior does not smooth across similar observations (see Remark 10). In some settings, such as when the goal of the analysis is to interpolate between similar observations, the Dirichlet process is therefore not an appropriate choice of prior. In this section we propose the Dirichlet process mixture as a flexible default prior class for the critic in such settings. We discuss the properties of this prior class, explain the resulting sampling algorithm, and illustrate the use of the prior class with an application to regression discontinuity.

5.1 A Default Smoothing Prior

Suppose that the critic's prior π is in the Dirichlet process mixture class, so that $\pi = DPM(\Psi, \alpha, G)$, where the kernel $\Psi : \Gamma \to \Delta(\mathcal{X})$ is a distribution parameterized by $\gamma \in \Gamma$, the parameter $\alpha > 0$ again controls the precision of the prior, and the centering measure $G \in \Delta(\Gamma)$ controls the location of the prior. Under such a prior, to draw a single observation, one may first draw a distribution $F \sim DP(\alpha, G)$, then draw a parameter $\gamma \sim F \in \Delta(\Gamma)$ according to that distribution, and then finally draw an observation $X_i \sim \Psi(\gamma) \in \Delta(\mathcal{X})$ according to the kernel with the given parameter.[8] Intuitively, the Dirichlet process mixture class modifies the Dirichlet process class so that the Dirichlet process describes, not a prior over the distribution of an observation itself, but a prior over the distribution of the parameter $\gamma \in \Gamma$.

Because the kernel may in principle be any distribution, the Dirichlet process mixture is a flexible class. Because the kernel may be a continuous distribution, this prior class allows a notion of smoothing across related observations. In the limiting case in which the kernel is a Dirac measure, the Dirichlet process mixture collapses to the Dirichlet process. The choice of kernel and centering measure can be guided by standard principles for prior selection and validated using posterior predictive checks (Gelman et al., 2013, Chapters 2 and 6).

Given suitable choice of kernel and centering measure, it is reasonably convenient to sample from the posterior corresponding to a Dirichlet process mixture prior.[9] Because the corresponding posterior distribution over γ is supported on a potentially infinite set, some (though not all) implementations call for restricting attention to a finite number K of such parameters. Algorithm 4 describes one such implementation.

Though reasonably convenient, in many settings Algorithm 4 involves more computation than Algorithm 2. In addition, Algorithm 4 requires specifying a centering measure G and kernel $\psi(\cdot)$, whereas Algorithm 2 requires no similar parameterization. We therefore recommend Algorithm 2 as a default, and Algorithm 4 for situations in which smoothing

[8] In practice, one may add additional richness, such as fixed unit-specific covariates (Ghosal and van der Vaart, 2017, Chapter 5.1).

[9] As evidence for this, one may observe that a standard textbook reference describes methods for sampling from the posterior (Ghosal and van der Vaart, 2017, Chapter 5.2), and that an existing Python library for Bayesian inference includes a corresponding vignette (Rochford and Das, 2021).

Algorithm 4 Numerical evaluation by a critic with a smoothing prior

- Select a number of kernel parameters $K \in \mathbb{N}$
- To draw $P_d \sim \pi(P|X)$, do:
  - Use the stick-breaking process to record, under a draw of the distribution $F_d \sim \pi(F|X)$:
    - the most probable values $\gamma_{d,1}, ..., \gamma_{d,K}$
    - the associated weights $\omega_{d,1}, ..., \omega_{d,K}$
  - Define $P_d = \sum_k \omega_{d,k} \Psi(\gamma_{d,k})$ as a mixture of kernels
- To sample $X_{d,s} \sim P_d$, independently for each $i \in \{1, ..., N\}$ do:
  - Draw $k_{d,s,i} \sim Multinomial(1; \omega_{d,1}, ..., \omega_{d,K})$
  - Draw $X_{d,s,i} \sim \Psi\left(\gamma_{d,k_{d,s,i}}\right)$

across similar observations is required.

Remark 13. (Posterior consistency of the Dirichlet process mixture prior.) *Theorems 7.2 and 7.15 in* Ghosal and van der Vaart (2017) *provide sufficient conditions for the Dirichlet process mixture posterior to be consistent with respect to the topology τ of total variation.*[10] *Under these conditions, and any frequentist guarantee that is uniformly asymptotically valid on τ-open sets of distributions $\mathcal{P}^U \subseteq \Delta(\mathcal{X})$, Proposition 1 thus again implies that a critic with a Dirichlet process mixture prior does not asymptotically raise false alarms.*

Remark 14. (Alternatives to the Dirichlet process mixture class.) *Alternative prior classes exist that may be used in place of the Dirichlet process mixture. Norets and Pelenis (2022), for example, propose mixtures of finite mixtures and characterize the contraction rates of the resulting posteriors.*

5.2 Applications to Regression Discontinuity

Gelman and Imbens (2019) study the properties of regression discontinuity designs in ap-

[10] *For instance, it suffices (but is not necessary) that the kernel $\Psi(\gamma)$ have density $\psi(\cdot;\gamma)$ with respect to a σ-finite measure for all γ, where $\psi(x;\gamma)$ is bounded and Lipschitz in γ for all x with a common Lipschitz constant, that Γ is bounded, and that the true density p_0 is bounded away from zero and belongs to the closure of $\left\{\int \psi(\cdot;\gamma)\, dF : F \in \Delta(\Gamma)\right\}$ with respect to the supremum norm.*

plications to incumbency advantage (Lee, 2008) and summer programs (Jacob and Lefgren, 2004; Matsudaira, 2008). All procedures they consider use a polynomial to describe the conditional expectation of the outcome given the running variable on either side of the cut-off. Gelman and Imbens (2019) argue that researchers should favor procedures based on local, lower-order polynomial fits over procedures based on global, higher-order polynomial fits. Gelman and Imbens (2019) fit local polynomials using a triangular kernel and the mean-squared-error-optimal bandwidth of Imbens and Kalyanaraman (2012). We consider their applications in turn and ask the critic to evaluate these and other procedures. We use replication files provided to us by Gelman and Imbens.

5.2.1 Incumbency Advantage

Lee (2008) estimates the incumbency advantage in US House elections using a regression discontinuity design. The running variable is the Democratic margin of victory in the previous election. The outcome variable is the Democratic vote share in the current election. We take as the parameter of interest the difference between the right and left limits of the conditional expectation of the outcome given the running variable, evaluated at the cutoff for Democratic victory in the previous election (Lee, 2008, Proposition 2b).

We ask the critic to assess the bias and coverage of the estimation and inference procedures, respectively, in Lee's (2008) setting. We specify the critic's prior as a Dirichlet process mixture. The kernel describes the running variable and the outcome variable as beta distributions, with mass points at the boundaries of their supports representing uncontested elections, and with the parameters describing the distribution of the outcome variable depending on the value of the running variable. Online Appendix C provides additional details and diagnostics on the prior specification.

Panel (a) of Table 2 presents the critic's assessment of the properties of procedures based on global polynomials of orders one through six. The critic assesses that the estimates may be severely biased. Across the different polynomial orders, the posterior expected bias ranges from -3.56 to 3.08 percentage points, as compared to a preferred point estimate in Lee (2008) of 7.70 percentage points. Across the different polynomial orders, the probability that bias is more than five percent of Lee's (2008) preferred point estimate is never less than 0.18. The

Table 2: Critic's Assessment of Bias and Coverage in an Application to Incumbency Advantage (Lee, 2008)

	Posterior E[bias]	Posterior Pr(|bias| > 0.05|estimate|)	Posterior Pr(coverage < nominal - 0.05)
Panel (a): Procedures based on global polynomials			
Order 1	3.08	1.00	1.00
Order 2	-3.56	1.00	1.00
Order 3	1.64	1.00	1.00
Order 4	0.12	0.18	0.02
Order 5	-0.69	0.89	0.19
Order 6	0.54	0.76	0.02
Panel (b): Procedures based on local polynomials			
Order 1	-0.03	0.00	0.00
Order 2	0.06	0.00	0.00
Panel (c): Alternative local procedures not studied by Gelman and Imbens (2019)			
Calonico et al. (2014)	0.02	0.00	0.00
Armstrong and Kolesár (2018, 2020)	-0.01	0.00	0.00

Note. The table shows the critic's assessment of the expected bias, the probability that bias is greater in magnitude than 0.05 times the Lee (2008) point estimate, and the probability that coverage is at least 5 percentage points below the nominal level of 95% for each estimator and associated confidence procedure. The estimators and associated confidence procedures are defined in Section 5.2. We use the Calonico et al. (2024) implementation of the mean-squared-error-optimal bandwidth of Imbens and Kalyanaraman (2012). To apply the procedure in Armstrong and Kolesár (2018, 2020), we trim the support of the running variable from $[-1, 1]$ to $[-0.99, 0.99]$. All values are based on applying Algorithm 4 using $D = 200$ draws from the posterior distribution, $S = 500$ samples from each draw, and the parameterization described in Online Appendix C.

critic likewise assesses that the confidence intervals may undercover. Across all polynomial orders the posterior probability that coverage is more than five percentage points below the nominal level ranges from 0.01 to near certainty.

Panel (b) of Table 2 presents the critic's assessment of the properties of procedures based on local polynomials of orders one and two. The critic assesses that the bias is small, with the posterior expected bias being always below 0.06 percentage points in absolute value, and the probability that the bias is more than five percent of Lee's (2008) preferred point estimate never above 0.00. The critic assesses probability 0.00 that coverage is more than five percentage points below the nominal level.

Taken together, Panels (a) and (b) of Table 2 show that the critic raises doubts about canonical frequentist guarantees for procedures based on global polynomials, but not for procedures based on local polynomials. In this sense the critic agrees, in this setting, with the recommendation in Gelman and Imbens (2019) to prefer procedures based on local polynomials.

Panel (c) of Table 2 presents the critic's assessments of the properties of two alternative procedures not studied by Gelman and Imbens (2019). These procedures are based on local polynomials, and differ from those studied in Gelman and Imbens (2019) in how they adjust for the effect of the bandwidth choice on the relevant asymptotic approximations. The first procedure is the one proposed by Calonico et al. (2014). The second is the one proposed by Armstrong and Kolesár (2018), using the data-driven tuning parameter proposed by Armstrong and Kolesár (2020).[11] The critic has little doubt about the guarantees for these procedures.

Panel (c) of Table 2 illustrates the value of the critic in assessing procedures not studied in Gelman and Imbens (2019). An analyst with access to the critic could learn whether the critic raises doubts about these alternative procedures, as we have done here.

Remark 15. (The Dirichlet process prior does not smooth, continued.) *Suppose that, instead of the Dirichlet process mixture prior considered here, we endow the critic with the Dirichlet*

[11] We trim the boundary regions of the support of the running variable because the assumptions proposed by Armstrong and Kolesár (2018) do not allow mass points. Online Appendix Table 5 presents results when we use the data-driven tuning parameter but include the boundary regions of the support of the running variable, and when we use an oracle tuning parameter that requires knowledge of the true conditional expectation function.

process prior considered in Section 3. Then, in the setting of Lee (2008), under the critic's posterior belief the parameter of interest is undefined with probability 1. The setting of Lee (2008) therefore illustrates that a Dirichlet process prior, though useful in many settings, is not always appropriate.

5.2.2 Summer Programs

Jacob and Lefgren (2004) and Matsudaira (2008) estimate the effect of summer programs on student achievement using regression discontinuity designs. In both settings, the running variable is discrete, which means the usual regression discontinuity estimand is formally undefined. For completeness, Online Appendix D presents results using a Dirichlet process mixture with a Gaussian kernel, where we again find that the critic raises more doubts about global than local procedures. We think a better approach, which we do not pursue here for brevity, could be to apply the default critic of Section 3 and take as the parameter of interest the estimand in Kolesár and Rothe (2018), which is defined without requiring continuity of the running variable.

6 Conclusions

We propose a Bayesian critic as both a thought experiment and a practical tool for assessing the properties of frequentist procedures. A default implementation using the Bayesian bootstrap is fully automated up to a choice of numerical precision. The default implementation proves surprisingly versatile in handling many interesting economic settings, in which the critic automatically recovers known concerns with standard procedures and also offers new insights on alternative procedures. An alternative implementation using a default smoothing prior likewise illustrates the practical value of the critic. Because the performance of frequentist procedures is inherently context-dependent, we see the critic as a useful complement to theoretical analysis for evaluating procedures in the settings where they are applied.

Appendix: Proofs of Statements Given in Main Text

Proposition 1

Proof. Since P_0 is in the τ-interior of $\mathcal{P}_N^U \cap \mathcal{P}^\pi$, there exists a τ-open set $\mathcal{U}$ such that $P_0 \in \mathcal{U} \subseteq \mathcal{P}_N^U \cap \mathcal{P}^\pi$. Because the sequence of sets $\mathcal{P}_N^U$ is non-decreasing in N, $\mathcal{U} \subseteq \mathcal{P}_{N+K}^U$ for all $K \geq 0$. Assumption 1 implies there exists K_0 such that, for all $K \geq K_0$,

$$\sup_{P \in \mathcal{U}} |M_{N,P}(T_N, \theta)| \leq \sup_{P \in \mathcal{P}_{N+K}^U} |M_{N+K,P}(T_{N+K}, \theta)| < \eta$$

if (1) holds and

$$\inf_{P \in \mathcal{U}} M_{N,P}(T_N, \theta) \geq \inf_{P \in \mathcal{P}_{N+K}^U} M_{N+K,P}(T_{N+K}, \theta) > -\eta$$

if (2) holds. Because $\mathcal{U}$ is τ-open and contains P_0, Assumption 2 implies that $\pi\left(\mathcal{U}|X^{N+K}\right) \to_p 1$ as $K \to \infty$, from which the result is immediate. □

Proposition 2

Proof. By Theorem 4.6 in Ghosal and van der Vaart (2017), for any $Q \in \Delta(\mathcal{X})$ and $\pi_{\alpha,Q} = DP(\alpha, Q)$, $\pi_{\alpha,Q}(P|X) = DP\left(\alpha + N, \frac{\alpha}{\alpha+N}Q + \frac{N}{\alpha+N}\hat{P}_N\right)$. By Theorem 4.16 of Ghosal and van der Vaart (2017), $\pi_{\alpha,Q}(P|X) \to_d \pi^B(P|X) = DP\left(N, \hat{P}_N\right)$ as $\alpha \to 0$, where $\to_d$ denotes convergence in distribution. By assumption, M_P is continuous in P (with respect to the topology of weak convergence) almost everywhere on the support of $\pi^B(P|X)$. Hence, by the continuous mapping theorem (e.g., Theorem 18.11 of van der Vaart, 1998), $\pi_{\alpha,Q}(M_P|X) \to_d \pi^B(M_P|X)$. □

References

Advani, A., Kitagawa, T., and Słoczyński, T. (2019). Mostly harmless simulations? Using Monte Carlo studies for estimator selection. *Journal of Applied Econometrics*, 34(6):893–910.

Andrews, I., Chen, J., and Tecchio, O. (Forthcoming). The purpose of an estimator is what it does: Misspecification, estimands, and over-identification. In *Advances in Economics and Econometrics: Thirteenth World Congress*.

Andrews, I., Essig Aberg, S., and Shapiro, J. M. (2025). Partial replication of Young (2019). `https://github.com/JMSLab/Young2019`. GitHub repository.

Andrews, I. and Mikusheva, A. (2016). A geometric approach to nonlinear econometric models. *Econometrica*, 84(3):1249–1264.

Andrews, I. and Shapiro, J. M. (Forthcoming). Communicating scientific uncertainty via approximate posteriors. *Econometrica*.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.

Armstrong, T., Kitagawa, T., and Tetenov, A. (Forthcoming). Statistical decision theory and empirical practice. *Journal of Political Economy: Microeconomics*.

Armstrong, T. B. and Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.

Armstrong, T. B. and Kolesár, M. (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics*, 11(1):1–39.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2024). Using Wasserstein Generative Adversarial Networks for the design of Monte Carlo simulations. *Journal of Econometrics*, 240(2):105076.

Blair, G., Cooper, J., Coppock, A., and Humphreys, M. (2019). Declaring and diagnosing research designs. *American Political Science Review*, 113(3):838–859.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

Busso, M., DiNardo, J., and McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, 96(5):885–897.

Calonico, S., Cattaneo, M. D., Farrell, M. H., Masini, R., and Titiunik, R. (2024). rdrobust: Statistical inference and graphical procedures for regression discontinuity designs. Python package.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Chamberlain, G. and Imbens, G. W. (2003). Nonparametric applications of Bayesian inference. *Journal of Business and Economic Statistics*, 21(1):12–18.

Efron, B. (1986). Why isn't everyone a Bayesian? *American Statistician*, 40(1):1–5.

Ferman, B. (2025). Assessing inference methods. *arXiv preprint arXiv:1912.08772*.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). Bayesian nonparametric inference of switching dynamic linear models. *IEEE Transactions on Signal Processing*, 59(4):1569–1585.

Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press, Cambridge.

Gasparini, M. (1995). Exact multivariate Bayesian bootstrap distributions of moments. *Annals of Statistics*, 23(3):762–768.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3 edition.

Gelman, A. and Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics*, 37(3):447–456.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Huber, M., Lechner, M., and Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, 175(1):1–21.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3):933–959.

Jacob, B. A. and Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, 86(1):226–244.

Kleibergen, F. (2005). Testing parameters in GMM without assuming that they are identified. *Econometrica*, 73(4):1103–1123.

Kleibergen, F. and Zhan, Z. (2025). Double robust inference for continuous updating GMM. *Quantitative Economics*, 16(1):295–327.

Knaus, M. C., Lechner, M., and Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *Econometrics Journal*, 24(1):134–161.

Kolesár, M. and Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, 108(8):2277–2304.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

Lechner, M. and Wunsch, C. (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics*, 21:111–121.

Lee, D. S. (2008). Randomized experiments from non-random selection in US House elections. *Journal of Econometrics*, 142(2):675–697.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics*, 142(2):829–850.

Müller, U. K. and Norets, A. (2016). Credibility of confidence sets in nonstandard econometric problems. *Econometrica*, 84(6):2183–2213.

Nickerson, D. W., Friedrichs, R. D., and King, D. C. (2006). Partisan mobilization campaigns in the field: Results from a statewide turnout experiment in Michigan. *Political Research Quarterly*, 59(1):85–97.

Norets, A. and Pelenis, J. (2022). Adaptive Bayesian estimation of discrete-continuous distributions under smoothness and sparsity. *Econometrica*, 90(3):1355–1377.

Parikh, H., Varjao, C., Xu, L., and Tchetgen Tchetgen, E. (2022). Validating causal inference methods. In *International Conference on Machine Learning*, pages 17346–17358. PMLR.

Pfanzagl, J. (1994). *Parametric Statistical Theory*. De Gruyter, New York.

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(2):169–203.

Rainforth, T., Cornish, R., Yang, H., Warrington, A., and Wood, F. (2018). On nesting Monte Carlo estimators. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, pages 4267–4276.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, New York, 2 edition.

Roberts, M. J. and Schlenker, W. (2013). Identifying supply and demand elasticities of agricultural commodities: Implications for the US ethanol mandate. *American Economic Review*, 103(6):2265–2295.

Rochford, A. and Das, A. (2021). Dirichlet process mixtures for density estimation. In PyMC Team, editor, *PyMC Examples*.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9(1):130–134.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons.

Saville, B. R., Connor, J. T., Ayers, G. D., and Alvarez, J. (2015). The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clinical Trials*, 11(4):485–493.

Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92(1):31–46.

Schorfheide, F. (2000). Loss function-based evaluation of DSGE models. *Journal of Applied Econometrics*, 15(6):645–670.

Schorfheide, F. and You, Z. (Forthcoming). Uncertainty in empirical economics. *Journal of Political Economy: Microeconomics*.

Schuler, A., Jung, K., Tibshirani, R., Hastie, T., and Shah, N. (2017). Synth-validation: Selecting the best causal inference method for a given dataset. *arXiv preprint arXiv:1711.00083*.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *American Statistician*, 55(1):62–71.

Simonsohn, U. (2021). Hyping Fisher: The most cited 2019 QJE paper relied on an outdated Stata default to conclude regression p-values are inadequate. Data Colada (blog). Post #99. Published October 13, 2021; updated October 27, 2021.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stock, J. H. and Wright, J. H. (2000). GMM with weak identification. *Econometrica*, 68(5):1055–1096.

Taddy, M., Gardner, M., Chen, L., and Draper, D. (2016). A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business and Economic Statistics*, 34(4):661–672.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Young, A. (2018). Replication Data for: 'Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results'. Harvard Dataverse, V1. Available at `https://doi.org/10.7910/DVN/JX6HCJ`.

Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, 134(2):557–598.

Online Appendix for "A Bayesian Critic for Frequentist Procedures"

Isaiah Andrews, *MIT and NBER*[1]
Simon Essig Aberg, *Harvard University*
Jesse M. Shapiro, *Harvard University and NBER*

Abstract

This online appendix contains additional results and details to accompany the material in Andrews, Essig Aberg, and Shapiro (2026), "A Bayesian Critic for Frequentist Procedures."

A Additional Theoretical Results

A.1 Bayesian Foundations of Frequentist Guarantees

Suppose that a frequentist guarantee holds on the set $\mathcal{P} \subseteq \Delta(\mathcal{X})$ in the sense that

$$M_P(T, \theta) \in \mathcal{M}_\eta$$

for all $P \in \mathcal{P}$ and for $\mathcal{M}_\eta = [-\eta, \eta]$ (for a two-sided guarantee) or $\mathcal{M}_\eta = [-\eta, \infty)$ (for a one-sided guarantee). Since $\mathcal{M}_\eta$ is convex by definition, it follows that any Bayesian audience member with beliefs $\pi^* \in \Delta(\mathcal{P})$ trusts the guarantee in the sense that

$$M_{\pi^*}(T, \theta) \in \mathcal{M}_\eta$$

for $M_{\pi^*}(T, \theta) = \mathrm{E}_{P \sim \pi^*}[M_P(T, \theta)]$. This elementary fact connects to several classic foundations of frequentist guarantees.

Example. (Bias and coverage; Pratt 1965) If $M_P(T, \theta)$ is the bias of an estimator $T(\cdot)$, and the guarantee is two-sided, $\mathcal{M}_\eta = [-\eta, \eta]$, then a frequentist guarantee on $\mathcal{P}$ implies that any

[1]E-mail: iandrews@mit.edu, sessigaberg@g.harvard.edu, jesse_shapiro@fas.harvard.edu.

audience member with beliefs $\pi^* \in \Delta(\mathcal{P})$ trusts that the bias is, on average, of magnitude no greater than η.

Likewise, if $M_P(T, \theta)$ is the coverage (relative to nominal) of a confidence set $T(\cdot)$, and the guarantee is one-sided, $\mathcal{M}_\eta = [-\eta, \infty)$, then a frequentist guarantee on $\mathcal{P}$ implies that any audience member with beliefs $\pi^* \in \Delta(\mathcal{P})$ trusts that coverage is, on average, no more than η below nominal.

These arguments directly extend those of Pratt (1965), who addresses the case where $\eta = 0$. △

Example. (Risk; Savage 1954) If $T(\cdot)$ is a decision rule and $m : \mathcal{T} \times \Theta \to \mathbb{R}_{\leq 0}$ is the (negative) associated loss, then $M_P(T, \theta) = \mathrm{E}_P[m(T(X), \theta(P))]$ is the (negative) frequentist risk for DGP P. A (one-sided) frequentist guarantee on $\mathcal{P}$ implies that any audience member with beliefs $\pi^* \in \Delta(\mathcal{P})$ has Bayes risk bounded above by η. The procedure associated with the smallest such η is the minimax procedure, which Savage (1954, Chapter 10) justifies along similar lines. △

A.2 Relationship of the Critic to the Audience

Suppose that a critic doubts the guarantee. We can connect such doubt to the foundations of frequentist guarantees in Online Appendix A.1.

Example. (Bias and coverage; Pratt 1965, continued) If $\pi(\mathcal{P}^\eta | X) = \pi(M_P(T, \theta) \in \mathcal{M}_\eta | X) < 1$, then it is immediate that there exists some belief $\pi^* \in \Delta(\Delta(\mathcal{X}))$ such that an audience member with belief π^* does not trust the guarantee, $M_{\pi^*}(T, \theta) \notin \mathcal{M}_\eta$. Moreover, if $\pi(\mathcal{P}^\eta | X) \leq 1 - \delta$ for some $\delta > 0$, then $M_{\pi^*}(T, \theta) \notin \mathcal{M}_\eta$ for some belief π^* in, respectively, a $\left(1 - \frac{\delta}{2}\right)$-total variation neighborhood of $\pi(\cdot | X)$ (for two-sided guarantees) or a $(1 - \delta)$-total variation neighborhood of $\pi(\cdot | X)$ (for one-sided guarantees). △

Example. (Risk; Savage 1954, continued) If $M_{\pi(\cdot|X)}(T, \theta) \notin \mathcal{M}_\eta$ and the belief $\pi(\cdot | X)$ is in the audience, then at least one audience member has Bayes risk in excess of η. Moreover, if the (negative) loss $m : \mathcal{T} \times \Theta \to [-\overline{m}, 0]$ is bounded and $M_{\pi(\cdot|X)}(T, \theta) < -\eta - \delta\overline{m}$ for some $\delta \geq 0$, then $M_{\pi^*}(T, \theta) \notin \mathcal{M}_\eta$ for all beliefs π^* in a $\delta-$total variation neighborhood of $\pi(\cdot | X)$. △

A.3 No Asymptotic False Assurances

Online Appendix Proposition 1. *Suppose Assumption 2 holds, and that for some τ-open set $\mathcal{U} \subseteq \mathcal{P}^{\pi}$,*

$$\limsup_{N\to\infty} \sup_{P\in\mathcal{U}} M_{N,P}\left(T_N, \theta\right) < 0.$$

Then there exists $\eta > 0$ such that for any $P_0 \in \mathcal{U}$, and $X_1, ..., X_N \overset{iid}{\sim} P_0$ for all N,

$$1 - \pi\left(\mathcal{P}_N^{\eta} | X^N\right) \geq \pi\left(\left\{P : M_{N,P}\left(T_N, \theta\right) < -\eta\right\} | X^N\right) \to_p 1. \tag{3}$$

Proposition 1.

Proof. Recall that $\mathcal{P}_N^{\eta} = \left\{P \in \Delta\left(\mathcal{X}\right) : M_{N,P}\left(T_N, \theta\right) \in \mathcal{M}_{\eta}\right\}$, and observe that the inequality in (3) binds for $\mathcal{M}_{\eta} = [-\eta, \infty)$, but may otherwise be strict. Set $\eta > 0$ equal to $-\frac{1}{2}\limsup_{N\to\infty}\sup_{P\in\mathcal{U}} M_{N,P}\left(T_N, \theta\right)$. The definition of lim sup implies that there exists N_{η} such that for all $N \geq N_{\eta}$,

$$\sup_{P\in\mathcal{U}} M_{N,P}\left(T_N, \theta\right) < -\eta.$$

Thus, for all $N > N_{\eta}$

$$\pi\left(\left\{P : M_{N,P}\left(T_N, \theta\right) < -\eta\right\} | X^N\right) \geq \pi\left(\mathcal{U} | X^N\right).$$

Since Assumption 2 implies that $\pi\left(\mathcal{U} | X^N\right) \to_p 1$ as $N \to \infty$, the result follows. □

B Additional Findings in Applications

Online Appendix Table 1: Critic's Assessment of Coverage of Alternative Procedures in a Returns to Schooling Application (Angrist and Krueger, 1991)

	2SLS parameter	CUGMM parameter
Posterior Pr(coverage < nominal - 0.05)		
EHW confidence interval, OLS	0.99	1.00
EHW confidence interval, 2SLS	0.20	0.75
S-statistic procedure	1.00	1.00
K-statistic procedure	0.51	0.05
DRLM-statistic procedure	0.07	0.00

Note. The table shows the critic's assessment of the probability that coverage is at least 5 percentage points below the nominal level of 95% for each type of confidence procedure (row) and each parameter (column). The confidence procedures are defined in Section 4.1. The parameters are the value under the true DGP of, respectively, the two-stage least squares estimator (2SLS) and the continuous updating GMM estimator. All values are based on applying Algorithm 2 using $D = 200$ draws from the posterior distribution and $S = 500$ samples from each draw.

Online Appendix Table 2: Sample Selection for Analysis of Randomized Controlled Trials (Young, 2019)

	Number of studies	Number of parameters
Young (2019)	53	4044
Reason for exclusion		
Data unavailable	10	353
Weights not accepted	3	237
Bootstrap	2	432
Memory intensive	2	112
Analysis sample	36	2910

Note. The table shows statistics for the sample of studies in Young (2019) and for those in our analysis sample. Procedures where weights are not accepted are those for which the Stata command used in the study's original implementation does not accept weights. Memory-intensive procedures are those for which the study's original implementation has peak RAM use exceeding 300MB.

Online Appendix Table 3: Partial Replication of Young (2019, Table V)

Significance level	.01	.05	.01	.05	.01	.05	.01	.05
	All papers (53 papers)		Low leverage (18 papers)		Medium leverage (17 papers)		High leverage (18 papers)	
Authors' p-value	0.216	0.354	0.199	0.310	0.164	0.313	0.283	0.437
Randomization-t	0.78	0.87	0.96	0.98	0.79	0.96	0.65	0.74

(a) Young (2019) Sample

Significance level	.01	.05	.01	.05	.01	.05	.01	.05
	All papers (36 papers)		Low leverage (12 papers)		Medium leverage (12 papers)		High leverage (12 papers)	
Authors' p-value	0.251	0.386	0.224	0.332	0.165	0.312	0.365	0.515
Randomization-t	0.81	0.88	1.00	1.01	0.75	0.92	0.71	0.79

(b) Analysis Sample

Note. Panel (a) reproduces results from Young (2019, Table V). Panel (b) reports analogous results for our analysis sample. Within each panel, the first row contains the average fraction of the parameters that are significant at the specified level under the authors' original procedure. The second row contains the fraction of the parameters that are significant at the specified level based on randomization inference, divided by the corresponding value from the first row.

Online Appendix Table 4: Critic's Assessment of Coverage in Randomized Controlled Trials (Young, 2019), Incorporating Randomization Design

	All		Low leverage		Medium leverage		High leverage	
	99%	95%	99%	95%	99%	95%	99%	95%
Panel (a): No controls beyond stratum								
Average posterior Pr(coverage < nominal - 0.05)								
Baseline	0.05	0.05	0.09	0.08	0.00	0.00	0.00	0.00
Fully incorporate design	0.34	0.43	0.23	0.29	0.50	0.65	0.54	0.71
	(8 studies)		(5 studies)		(2 studies)		(1 study)	
Panel (b): Other controls								
Average posterior Pr(coverage < nominal - 0.05)								
Baseline	0.07	0.10	0.05	0.06	0.10	0.14	0.13	0.19
Partially incorporate design	0.41	0.54	0.37	0.48	0.51	0.66	0.31	0.42
	(12 studies)		(7 studies)		(4 studies)		(1 study)	

Note. The table summarizes the critic's doubt about coverage, defined as the critic's posterior probability that coverage is more than five percentage points below nominal, in a subset of the studies considered in Young (2019). We follow Young (2019) in averaging statistics weighting by the inverse of the number of parameters in a given study, and in classifying leverage (defined as the average, across parameters, of the leverage of the highest-leverage observation) by terciles in our analysis sample. The "baseline" approach follows Table 1 in applying Algorithm 2 using $D = 200$ draws from the posterior distribution and $S = 200$ samples from each draw. The "incorporate design" approach instead takes draws from the posterior separately by cells defined by the interaction of treatment and stratum. We restrict attention to those studies for which the necessary data are publicly available, and those procedures that, in the study's original implementation, accept weights (see Section 4.2 and Online Appendix Table 2) and do not cluster inference by stratum.

Online Appendix Table 5: Variations on Alternative Procedures in an Application to Incumbency Advantage (Lee, 2008)

	Posterior E[bias]	Posterior Pr(|bias| > 0.05|estimate|)	Posterior Pr(coverage < nominal - 0.05)
Armstrong and Kolesár (2018, 2020)			
Include boundary of support	-0.01	0.00	0.00
Oracle tuning parameter	-0.01	0.01	0.00

Note. The table shows the critic's assessment of the expected bias, the probability that bias is greater in magnitude than 0.05 times the Lee (2008) point estimate, and the probability that coverage is at least 5 percentage points below the nominal level of 95% for each estimator and associated confidence procedure. The procedures are variants of the procedure of Armstrong and Kolesár (2018, 2020) defined in Section 5.2. We use the Calonico et al. (2024) implementation of the mean-squared-error-optimal bandwidth of Imbens and Kalyanaraman (2012). In the first row ("Include boundary of support"), the procedure using the data-driven tuning parameter proposed by Armstrong and Kolesár (2020) is calculated on the entire sample without trimming the boundary regions of the support of the running variable. In the second row ("Oracle tuning parameter"), the data-driven tuning parameter is replaced with its oracle value under the given draw from the posterior distribution. All values are based on applying Algorithm 4 using $D = 200$ draws from the posterior distribution, $S = 500$ samples from each draw, and the parameterization described in Online Appendix C.

C Details on Priors for Regression Discontinuity Applications

The data $X_i = (R_i, Y_i)$ consist of a running variable R_i and an outcome variable Y_i. For the application to incumbency advantage, $R_i \in [-1, 1]$ and $Y_i \in [0, 1]$. For the application to summer programs, $R_i \in \mathbb{R}$ and $Y_i \in \mathbb{R}$. We define $W_i = \mathbf{1}[R_i \geq \overline{r}]$, where $\overline{r}$ denotes the cutoff in the running variable. We define the vector $Z_i = (1, R_i - \overline{r}, W_i, (R_i - \overline{r}) W_i)$.

A parameterization of the Dirichlet process mixture prior requires a kernel $\Psi(\cdot)$, a centering measure G, and a precision α.

For the application to incumbency advantage, we specify the parameter γ of the kernel to consist of four subvectors: (i) $\gamma_0^R = \left(\gamma_{0,L}^R, \gamma_{0,M}^R, \gamma_{0,H}^R\right) \in \Delta(\{1,2,3\})$, (ii) $\gamma_1^R = \left(\gamma_{1,a}^R, \gamma_{1,b}^R\right) \in \mathbb{R}_{>0}^2$, (iii) $\gamma_0^{Y|R} = \left(\gamma_{0,L}^{Y|R}, \gamma_{0,H}^{Y|R}\right) \in \mathbb{R}^4 \times \mathbb{R}^4$, and (iv) $\gamma_1^{Y|R} = \left(\gamma_{1,a}^{Y|R}, \gamma_{1,b}^{Y|R}\right) \in \mathbb{R}^4 \times \mathbb{R}^4$. We specify the kernel so that, under the distribution $\Psi(\gamma)$, we have

$$R_i|\gamma \sim \begin{cases} \delta(-1) & \text{with probability } \gamma_{0,L}^R, \\ 2Beta\left(\exp\left(\gamma_{1,a}^R\right), \exp\left(\gamma_{1,b}^R\right)\right) - 1 & \text{with probability } \gamma_{0,M}^R, \\ \delta(1) & \text{with probability } \gamma_{0,H}^R, \end{cases}$$

and

$$Y_i|R_i, \gamma \sim \begin{cases} \delta(0) & \text{with probability } \frac{\exp\left(Z_i'\gamma_{0,L}^{Y|R}\right)}{1+\exp\left(Z_i'\gamma_{0,L}^{Y|R}\right)+\exp\left(Z_i'\gamma_{0,H}^{Y|R}\right)}, \\ Beta\left(\exp\left(Z_i'\gamma_{1,a}^{Y|R}\right), \exp\left(Z_i'\gamma_{1,b}^{Y|R}\right)\right) & \text{with probability } \frac{1}{1+\exp\left(Z_i'\gamma_{0,L}^{Y|R}\right)+\exp\left(Z_i'\gamma_{0,H}^{Y|R}\right)}, \\ \delta(1) & \text{with probability } \frac{\exp\left(Z_i'\gamma_{0,H}^{Y|R}\right)}{1+\exp\left(Z_i'\gamma_{0,L}^{Y|R}\right)+\exp\left(Z_i'\gamma_{0,H}^{Y|R}\right)}, \end{cases}$$

where $\delta(\cdot)$ denotes the Dirac measure. We specify the centering measure G so that, under G, we have $\gamma_0^R \sim Dir(1,1,1)$, $\gamma_{1,a}^R, \gamma_{1,b}^R \sim Gamma(2,1)$, and $\left(\gamma_0^{Y|R}, \gamma_1^{Y|R}\right) \sim N(0, I_{16})$, independently across components. Finally, we specify the precision as $\alpha = 2$.

For the application to summer programs, we specify the parameter γ of the kernel to consist of four subvectors: (i) $\gamma_0^R \in \mathbb{R}$, (ii) $\gamma_1^R \in \mathbb{R}$, (iii) $\gamma_0^{Y|R} \in \mathbb{R}^4$, and (iv) $\gamma_1^{Y|R} \in \mathbb{R}^2$. We

specify the kernel so that, under the distribution $\Psi(\gamma)$, we have

$$R_i|\gamma \sim N\left(\gamma_0^R,\, \exp\left(\gamma_1^R\right)^2\right)$$

and

$$Y_i|R_i,\gamma \sim N\left(Z_i'\gamma_0^{Y|R},\, \exp\left(0.5\,(1,W_i)'\,\gamma_1^{Y|R}\right)^2\right).$$

We specify the centering measure G so that, under G, each component of γ is independently normally distributed across components. For the Jacob and Lefgren (2004) summer program application, we specify $\gamma_0^R \sim N(3,\, 2^2)$,$\gamma_1^R \sim N(0,\, 1^2)$, $\gamma_0^{Y|R} \sim N((-0.5, 0, 0, 0),\, I_4)$, and $\gamma_1^{Y|R} \sim N((0,0),\, I_2)$. For the Matsudaira (2008) summer program application, we specify $\gamma_0^R \sim N(10,\, 50^2)$, $\gamma_1^R \sim N(3.7,\, 1^2)$, $\gamma_0^{Y|R} \sim N(0,\, (1^2, 0.03^2, 1^2, 0.03^2) \odot I_4)$, and $\gamma_1^{Y|R} \sim N((0,0),\, I_2)$. Finally, we specify the precision as $\alpha = 2$.

To implement Algorithm 4, we sample from the posterior using the PyMC implementation of the No U-Turn variation of the Hamiltonian Monte Carlo algorithm. We specify the number of kernel parameters as $K = 8$. Online Appendix Table 6 reports diagnostics.

Online Appendix Table 6: Diagnostics for Applications to Incumbency Advantage (Lee, 2008) and Summer Programs (Jacob and Lefgren, 2004; Matsudaira, 2008)

	Lee	Jacob-Lefgren	Matsudaira
Posterior E[minimum weight]	0.001	0.023	0.001
Posterior predictive checks			
Global linear RD estimate			
Empirical value	0.118	-0.024	0.167
Posterior 95% predictive interval	[0.109, 0.143]	[-0.038, -0.007]	[0.091, 0.129]
Global linear RD 95% CI width			
Empirical value	0.021	0.034	0.028
Posterior 95% predictive interval	[0.021, 0.024]	[0.034, 0.035]	[0.036, 0.040]
Local linear RD estimate			
Empirical value	0.063	-0.195	-0.075
Posterior 95% predictive interval	[0.073, 0.117]	[-0.190, -0.124]	[-0.102, -0.041]
Local linear RD 95% CI width			
Empirical value	0.043	0.069	0.049
Posterior 95% predictive interval	[0.036, 0.051]	[0.064, 0.081]	[0.049, 0.058]

Note. The minimum weight is the smallest mixture weight across the $K = 8$ components. The posterior predictive checks report statistics computed on the original data and the corresponding 95% posterior predictive intervals. Each interval is constructed by drawing a data-generating process from the posterior, sampling a dataset of the same size as the original, computing the statistic on the sampled dataset, and repeating for D draws of the data-generating process. The interval is then formed as the 2.5th and 97.5th quantiles of the distribution of the statistic across the D draws. The local linear RD specification uses the Imbens and Kalyanaraman (2012) bandwidth. Posterior calculations are based on $D = 200$ draws from the posterior distribution and the parameterization described in Online Appendix C.

D Application to Summer Programs

Jacob and Lefgren (2004) and Matsudaira (2008) estimate the effect of summer programs on student achievement using regression discontinuity designs. In both settings, the running variable is a function of test scores and the outcome is a subsequent measure of academic performance. We take as the parameter of interest the difference between the left and right limits of the conditional expectation of the outcome given the running variable, evaluated at the cutoff for passing test scores. We specify the critic's prior as a Dirichlet process mixture. We specify a Gaussian kernel for both the running variable and the outcome variable, with the conditional mean of the outcome depending linearly on the running variable on each side of the cutoff. Online Appendix C provides further details and diagnostics.

Online Appendix Table 7 presents the critic's assessment of the bias and coverage of regression discontinuity procedures based on global and local polynomials. The critic casts substantial doubt on the properties of procedures based on global polynomials. Across polynomial orders and settings, the critic assesses at least a 0.47 posterior probability that the bias of global procedures exceeds 0.05 times the authors' point estimates of 0.11 (Jacob and Lefgren, 2004) and 0.09 (Matsudaira, 2008).[2]

The critic's doubt regarding the properties of procedures based on local polynomials is less extreme. For Jacob and Lefgren (2004) and Matsudaira (2008), respectively, the critic assesses a posterior probability of 0.00 and 0.00 that the local quadratic procedure exhibits substantial bias. The critic's remaining doubts about local procedures may arise in part due to the discreteness of the running variable, which means that even adjacent values of the running variable can be associated with fairly different mean outcomes.

[2]The Jacob and Lefgren (2004) point estimate is the difference in one-year math gain immediately left versus right of the cutoff (Table 2, columns 2–3), averaged across the third and sixth grade samples. The Matsudaira (2008) point estimate is the fifth grade math reduced-form second-stage coefficient (Table 2).

Online Appendix Table 7: Critic's Assessment of Bias and Coverage in Application to Summer Programs (Jacob and Lefgren, 2004; Matsudaira, 2008)

	Posterior E[bias]		Posterior Pr(|bias| > 0.05|estimate|)		Posterior Pr(coverage < nominal - 0.05)	
	Jacob-Lefgren	Matsudaira	Jacob-Lefgren	Matsudaira	Jacob-Lefgren	Matsudaira
Panel (a): Procedures based on global polynomials						
Order 1	-0.13	-0.19	1.00	1.00	1.00	1.00
Order 2	0.05	-0.03	1.00	1.00	1.00	1.00
Order 3	0.05	0.03	1.00	1.00	1.00	1.00
Order 4	0.01	0.01	1.00	0.93	0.67	0.01
Order 5	-0.01	-0.02	0.48	1.00	0.00	1.00
Order 6	-0.01	-0.01	0.67	0.99	0.00	0.00
Panel (b): Procedures based on local polynomials						
Order 1	0.00	-0.00	0.00	0.62	0.00	0.07
Order 2	0.00	0.00	0.00	0.00	0.00	0.00

Note. The table shows the critic's assessment of the expected bias, the probability that bias is greater in magnitude than 0.05 times the authors' point estimate, and the probability that coverage is at least 5 percentage points below the nominal level of 95% for each estimator and associated confidence procedure. The estimators and associated confidence procedures are defined in Section 5.2. We use the Calonico et al. (2024) implementation of the mean-squared-error-optimal bandwidth of Imbens and Kalyanaraman (2012). All values are based on applying Algorithm 4 using $D = 200$ draws from the posterior distribution, $S = 500$ samples from each draw, and the parameterization described in Online Appendix C.

Appendix References

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.

Armstrong, T. B. and Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.

Armstrong, T. B. and Kolesár, M. (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics*, 11(1):1–39.

Calonico, S., Cattaneo, M. D., Farrell, M. H., Masini, R., and Titiunik, R. (2024). rdrobust: Statistical inference and graphical procedures for regression discontinuity designs. Python package.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3):933–959.

Jacob, B. A. and Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, 86(1):226–244.

Lee, D. S. (2008). Randomized experiments from non-random selection in US House elections. *Journal of Econometrics*, 142(2):675–697.

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics*, 142(2):829–850.

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(2):169–203.

Savage, L. J. (1954). *The Foundations of Statistics.* John Wiley & Sons.

Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, 134(2):557–598.