Supplementary Materials for
Macartan Humphreys and Alan M. Jacobs, "Mixing Methods: A Bayesian Approach," *American Political Science Review*

A Probabilistic logics in existing work using process tracing

The Bayesian formalization of process tracing developed in the paper builds on an empirical logic that is already common in current accounts and, to a lesser extent, applications of process tracing.

Van Evera's hoop, smoking gun, and doubly decisive tests have played a central role in accounts of the logic of process tracing. Consistent with the logic of process tracing embedded in the BIQQ framework, these tests involve reasoning about the likelihood of observing particular pieces of evidence under alternative hypotheses. A notable feature of early accounts is that they described process tracing tests as grounded in an empirical logic of necessity and sufficiency. In a hoop test for a hypothesis, H, for instance, the observation of a given piece of evidence, K, was characterized as being *necessary* for the survival of H (though not sufficient for its affirmation). This corresponds in our framework to $\Pr(K = 1|H) = 1$. Likewise, in a smoking gun test, the observation of a piece of evidence, K, was characterized as being *sufficient* for the confirmation of H (though not necessary for its survival), formally $\Pr(K = 1|\neg H) = 0$ (Collier, 2011, 825-827; Mahoney, 2012, 5 and 7; Bennett, 2010, 210). Thus, some test outcomes—the passage of a smoking gun test, the failure of a hoop test, and either outcome in a double decisive test—were seen as having categorical (pass/fail) consequences for the hypothesis.

These early accounts thus suggested a deterministic feature of process tracing tests different from that taken in this paper.

Even in these accounts, however, the relationship between hypotheses and evidence was at least implicitly understood as partially uncertain; and some test outcomes were understood as incrementally shifting beliefs in a manner more consistent with Bayesian reasoning. The passage of a hoop test or the failure of a smoking gun test were seen as somewhat strengthening and weakening a hypothesis, respectively. They thus implicitly reflected probabilities for $\Pr(K = 1|\neg H)$, for the hoop test, and $\Pr(K = 1|H)$, for the smoking gun test, that lay between 0 and 1 (exclusive). Mahoney (2012) also points out that

hoop tests and smoking gun tests may be more or less difficult, depending on the commonness of the evidence (pp. 7 and 9). This consideration corresponds directly to the denominator in Bayes' rule.

Further, "straw in the wind" tests—where, in effect, $\Pr(K = 1|H) \neq \Pr(K = 1|\neg H)$ both lie strictly between 0 and 1—were seen as shifting beliefs somewhat toward or against the hypothesis, though treated as modest sources of inferential leverage. Indeed, Zaks (2013) points to the central importance of "pieces of evidence that are not quite definitive enough to qualify as either sufficient or necessary", but that nonetheless "may lend support to (or undermine) an explanation" (p. 11). Zaks relabels these as "leveraging tests" to highlight the key probative role that evidence with uncertain implications typically plays in process tracing.

Collier (2011, p. 827), moreover, addresses the possibility of ambiguity about the test type to which a causal process observation (CPO) corresponds, arising from different background theories of or reasoning about the data-generating process giving rise to the CPOs. Importantly, the BIQQ framework can capture this kind of uncertainty not just by allowing the ϕ parameters to take on intermediate values but also by allowing probability distributions over many possible values.

We note that the focus of these accounts is, in effect, on the likelihood of observing within-case data given different underlying processes. This focus does not imply the use of a fully Bayesian approach. And, as we demonstrate elsewhere in the Supplementary Materials (§J), a similar type of analysis to that we propose can be undertaken using a maximum likelihood framework. The Bayesian approach, however, better captures the accounts in this work of how inferences are drawn from clue information (in terms of lending support for one or another explanation) and, as noted above, the Bayesian approach provides an intuitive way of handling uncertainty over probative value.

While largely compatible with these accounts, the approach taken in the BIQQ framework is closest to that presented in recent accounts that embrace a fully probabilistic view of process tracing. Bennett (2015), Beach and Pedersen (2013, pp. 83ff), and Rohlfing (2012, p. 187–198) shift from the language of

sufficiency and necessity to a wholly probabilistic understanding of the relationship between hypotheses and pieces of evidence. Beach and Pedersen (2013, pp. 102), for instance, construe Van Evera's concepts of "certainty" and "uniqueness" as continua, representing differing probabilities of making a given observation under a hypothesis. These accounts, moreover, formalize the inferential reasoning involved in process tracing in Bayesian terms — as a function of the probability of the evidence, given the hypothesis; the unconditional probability of the evidence; and the prior probability of the hypothesis.[18] To our knowledge, current approaches do not allow for *uncertainty* over probative value (as opposed to intermediate probative value) or for joint uncertainty over causal types, probative value, and assignment processes.

To date, surprisingly few *substantive* works using process tracing feature explicit reasoning about the test types to which the search for particular pieces of evidence correspond or about the likelihood of observing pieces of evidence under alternative hypotheses. Two rare exceptions are Lengfelder (2012) and Fairfield (2013), which explicitly analyze CPOs in relation to Van Evera's test types. Fairfield (2013)'s treatment, moreover, treats inference from CPOs in an explicitly probabilistic fashion. Reasoning about "how surprising the evidence would be if a hypothesis were correct," Fairfield distinguishes among hoop and smoking gun tests according to their strength and the degree to which they "increase or decrease the likelihood that a hypothesis is correct to varying degrees" (p. 55). So far, at least, clear probabilistic reasoning about qualitative evidence has yet make the leap from principles to common practice.

[18]Critically, a probabilistic understanding of process tracing does not imply a probabilistic understanding of *causal relations*. A Bayesian approach to inference is fully compatible, for instance, with a set-theoretic, deterministic approach to causation. Indeed, the potential outcomes framework underlying the BIQQ approach itself assumes deterministic outcomes conditional on treatment.

B Bayesian process tracing: intuition and illustration

B.1 The role of priors in Bayesian updating

We note in the text of the paper that the amount of learning that results from a given piece of new data depends on prior beliefs.

Figure 5 illustrates these points. In each subgraph, we show how much learning occurs under different scenarios. The horizontal axis indicates the level of prior confidence in the hypothesis and the curve indicates the posterior belief that arises if we do (or do not) observe the clue. As can be seen, the amount of learning that occurs—the shift in beliefs from prior to posterior—depends a good deal on what prior we start out with. For a smoking gun test, the amount of learning is highest for values roughly in the 0.2 to 0.4 range—and then declines as we have more and more prior confidence in our hypothesis. For a hoop test, the amount of learning when the clue is *not* observed is greatest for hypotheses in which we have middling-high confidence (around 0.6 to 0.8), and minimal for hypotheses in which we have a very high or a very low level of confidence.

The implication here is that our inferences with respect to a hypothesis must be based not just on the search for a clue predicted by the hypothesis but also on the *plausibility* of the hypothesis, based on other things we know. Suppose, for instance, that we fail to observe evidence that we are 90 percent sure we *should* observe if a hypothesized causal effect has occurred: a strong hoop test is failed. But suppose that the existing literature has given us a very high level of confidence that the hypothesis *is* right. This high prior confidence, sometimes referred to as a "base rate," is equivalent to believing that the causal effect exists in a very high proportion of cases. Thus, while any given case with a causal effect has only a 0.1 chance of not generating the clue, the high base rate means that the vast majority of cases that we observe without the clue will nonetheless be cases with causal effects. Thus, the failure of even a strong hoop test, involving a highly certain prediction, should only marginally reduce

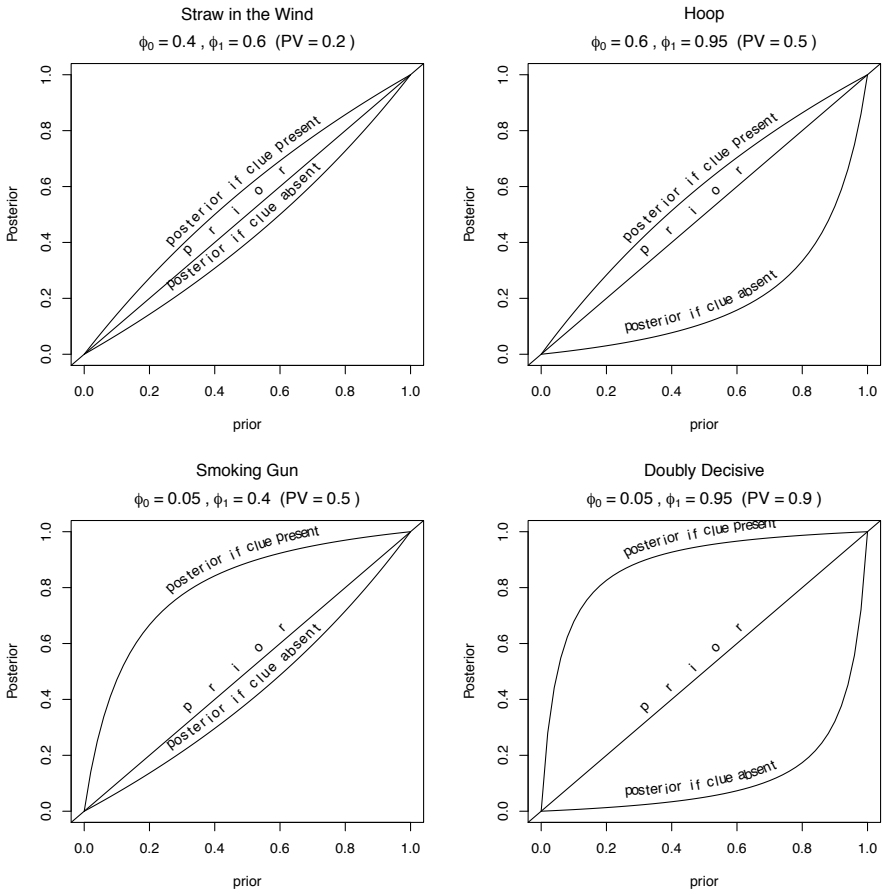

Figure 5: Figure shows how the learning from different types of tests depends on priors regarding the proposition. A "smoking gun" test has the greatest impact on beliefs when priors are middling low and the clue is observed; a "hoop test" has the greatest effect when priors are middling high and the clue is not observed. Here "PV" denotes the probative value of the test.

our confidence in a hypothesis that we strongly expect to be true.

A similar line of reasoning applies to smoking-gun tests involving hypotheses that prior evidence suggests are very unlikely to be true. Innocent people may be very unlikely to be seen holding smoking guns after a murder. But if a very high proportion of people observed are known to be innocent, then a very high proportion of those holding smoking guns will in fact be innocent—and a smoking-gun clue will be far from decisive.

We emphasize two respects in which these implications depart from common intuitions. First, we cannot make *general* statements about how decisive dif-

ferent categories of test, in Van Evera's framework, will be. It is commonly stated that hoop tests are devastating to a theory when they are failed, while smoking gun tests provide powerful evidence in favor of a hypothesis. But, in fact the degree of belief change depends not just on features of the clues but also on prior beliefs.

Second, although scholars frequently treat evidence that goes against the grain of the existing literature as especially enlightening, in the Bayesian framework the contribution of such evidence may sometimes be modest, precisely because received wisdom carries weight. Thus, although the discovery of *disconfirming* evidence—an observation thought to be strongly inconsistent with the hypothesis—for a hypothesis commonly believed to be true is more informative (has a larger impact on beliefs) than *confirming* evidence, this does not mean that we learn more than we would have if the prior were weaker. When it comes to very strong hypotheses, the "discovery" of disconfirming evidence is very likely to be a false negative; likewise, the discovery of supporting evidence for a very implausible hypothesis is very likely to be a false positive. The Bayesian approach takes account of these features naturally. We note, however, that one common intuition—that little is learned from disconfirming evidence on a low-plausibility hypothesis or from confirming evidence on a high-plausibility one—*is* correct.

B.2 Joint updating over ϕ and type

Here we elaborate on the intuition of multi-parameter Bayesian process tracing, in which updating occurs over both causal type (j) and beliefs about the probabilities with which clues are observed for each type (ϕ values). The illustration in the text makes clear how updating over type occurs, given beliefs about ϕ values. But how does updating over ϕ occur?

Suppose that we observe a case with values $X = 1, Y = 1$. We begin by defining a prior probability distribution over each parameter. Suppose that we establish a prior categorical distribution reflecting uncertainty over whether the case is a b type (e.g., setting a probability of 0.5 that it is a b and 0.5 that

is a d type). We also start with priors on ϕ_b and ϕ_d. For concreteness, suppose that we are certain that the clue is unlikely for a d type ($\phi_d = .1$), but we are very uncertain about ϕ_b; in particular, we have a uniform prior distribution over $[0, 1]$ for ϕ_b. Note that, even though we are very uncertain about ϕ_b, the clue still has probative value, arising from the fact that the expected value of ϕ_b is higher than that of ϕ_d.

Suppose that we then look for the clue in the case and observe it. This observation shifts posterior weight away from a belief that the case is a b. See Figure 6 for an illustration. Yet it *simultaneously* shifts weight toward a higher value for ϕ_b and a lower value for ϕ_d. The reason is that the observed clue has a relatively high likelihood *both* for combinations of parameter values in which the case is a d and ϕ_b is low *and* for combinations in which the case is a b and ϕ_b is *high* (or, equivalently, in this example, where ϕ_d is low). Since we now are more confident that the case is a b, however, the marginal posterior distribution of ϕ_b will be shifted upward relative to its prior marginal distribution. The joint posterior distribution will also reflect a dependency between the probability that the case is a b vs. a d, on the one hand, and ϕ_b and ϕ_d on the other.

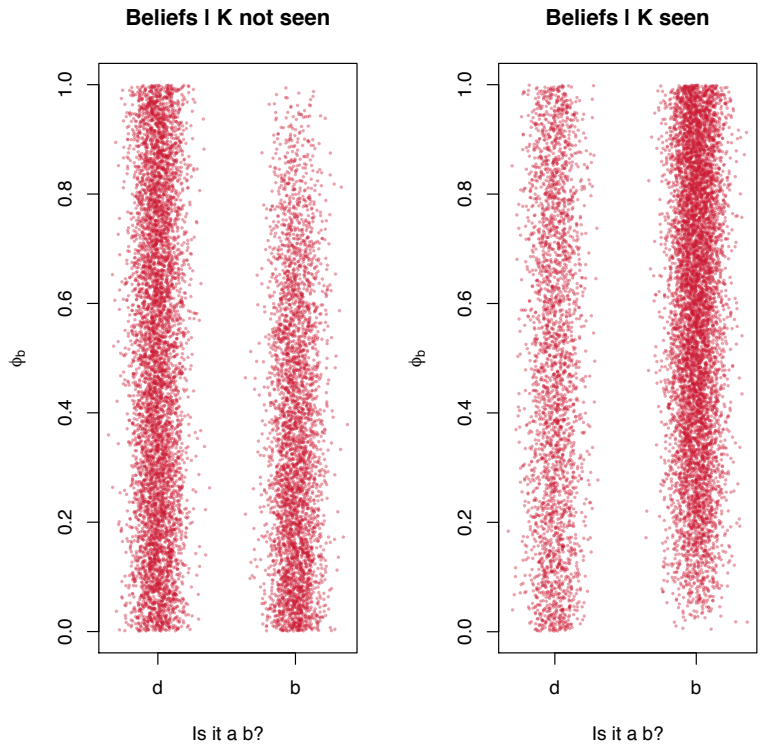

Figure 6: Joint posteriors distribution on whether a case is a b or d and on the probability of seeing a clue for a b type (ϕ_b).

C Bayesian correlational inference: intuition and illustration

In the text, we note that Bayesian updating is commonly used to draw causal inferences from correlational X, Y data. Here we elaborate on the intuition underlying correlational Bayesian inference for the binary problem and provide a simple illustration.

Formally, we update our beliefs over θ using Bayes' rule:

$$p(\theta|\mathcal{D}) = \frac{\Pr(\mathcal{D}|\theta)p(\theta)}{\int \Pr(\mathcal{D}|\theta')p(\theta')d\theta'} \tag{6}$$

Intuitively, we treat each set of possible values of our parameters of interest—each θ vector, that is—as a hypothesis and apply Bayes' rule to assess its probability given the data, that is, the posterior.[19] We use three quantities to calculate the posterior.

First, we ask, if this set of parameter values is true, how likely were the observed X, Y values to have emerged? [20] Consider a hypothesis (a specific value of θ) in which most authoritarian countries are assumed to be either susceptible to a regime-collapsing effect of economic crisis or destined to collapse anyway—i.e., a θ in which λ_b and λ_d are very high and λ_a and λ_c very low. Suppose we then observe data in which a large proportion of countries display values $X = 1$ and $Y = 0$—they experienced crisis and did not collapse—which pegs them as either a or c types. The probability of these data under the hypothesized θ — $\Pr(\mathcal{D}|\theta)$ — will then be low, reducing our confidence in this hypothesis. On the other hand, such data are far more likely under any θ vector in which λ_a or λ_c is high, boosting our confidence in such hypotheses.

[19] More generally we might think of a hypothesis as being a subset of values of θ — e.g. "there is a positive treatment effect" corresponds to the set of values for which $\lambda_b > \lambda_a$.

[20] This calculation in our binary framework is simple. For example, the probability of observing the event $X = 1, Y = 1$ for a single randomly selected case is given by event probability $w_{11} = b\pi_b + d\pi_d$. Note that we assume in this example that each *type* is drawn independently as would be the case if cases under study were randomly sampled from a large population.

Second, we ask, how likely were we to observe these data, $\mathcal{D}$, regardless of whether this particular θ is true? This value appears in the denominator, where we take into account the likelihood of observing these data for *all* of the possible values of θ, weighted by their prior probabilities. More formally, under the assumption of independence, the probability of observing $\mathcal{D}$, that is, a particular collection of X, Y data, is given by the corresponding value of the multinomial distribution given the event probabilities implied by θ.

The more likely the data are in general—whether the hypothesis is true or not—the smaller the effect of these data on our beliefs. On the other hand, if the observation of lots of crisis-suffering, collapsing regimes was generally *unlikely* across all θ's, then observing these data will generate a larger shift in our confidence toward any particular θ vector with which the data are relatively consistent.

Third, we multiply the ratio of these first two quantities by our confidence in the values in this θ prior to seeing the data ($p(\theta)$). The more prior confidence we have in a hypothesis, the greater the probability that evidence consistent with and unique to the hypothesis in fact indicates that the hypothesis is true. Thus, for instance, suppose that prior evidence and logic suggest that a high proportion of authoritarian regimes in the world are susceptible to a regime-collapsing effect of crisis (are b types). This strong prior belief in a high λ_b increases the likelihood that any data pattern consistent with a high λ_b—say, many $X = 1, Y = 1$ cases—has *in fact* been generated by a large set of b cases.

We can illustrate Bayesian correlational inference with a simple case. Suppose we observe for all postwar authoritarian regimes, whether they did or did not suffer economic crisis and did or did not collapse. Say for simplicity we know that all authoritarian regimes were "assigned" to economic crisis with a 0.5 probability during the period under analysis (thus assignment is known to be *as if* random). And assume that, prior to observing X, Y data we believe that each of two propositions is true with 0.5 probability. Under proposition (θ_1), all regimes are of type b (and so the average treatment effect is 1); under proposition (θ_2) 50% of regimes are of type c and 50% are of type d (and so

the average treatment effect is 0).[21]

Suppose we draw a random sample of $n = 2$ cases and observe one case in which $X = Y = 0$ and one case in which $X = Y = 1$. That is, we observe a perfect correlation between X and Y but only two cases. What then should we infer?

Applying Bayes' rule, our posterior probability on proposition θ_1, having observed the data, is:

$$\Pr(\theta_1|\mathcal{D}) = \frac{\Pr(\mathcal{D}|\theta_1)\Pr(\theta_1)}{\Pr(\mathcal{D}|\theta_1)\Pr(\theta_1) + \Pr(\mathcal{D}|\theta_2)\Pr(\theta_2)}$$

or equivalently:

$$\Pr(b = 1|\mathcal{D}) = \frac{\Pr(\mathcal{D}|\lambda_b = 1)\Pr(\lambda_b = 1)}{\Pr(\mathcal{D}|\lambda_b = 1)\Pr(\lambda_b = 1) + \Pr(\mathcal{D}|\lambda_b = 0)\Pr(\lambda_b = 0)}$$

The event probabilities of each of the observed events is 0.5 under θ_1 but just 0.25 under θ_2. Using the binomial distribution (a special case of the multinomial for this simple case) we know that the chances of such data arising are 1 in 2 under θ_1 but only 1 in 8 under θ_2. Our posterior would then be:

$$\Pr(\lambda_b = 1|\mathcal{D}) = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{1}{2} \times \frac{1}{2} + \frac{1}{8} \times \frac{1}{2}} = \frac{4}{5}$$

The key difference between this example and more general applications is simply that in the general case we allow for uncertainty — and updating — not simply over whether λ_b is 0 or 1, but over a range of possible values for multiple parameters of interest. Though this adds complexity, it does not change the fundamental logic of updating.

[21] In this simple case, we can think of θ as being constrained to take on only one of two possible values: $\theta \in \{\theta_1 = \{a = 0, b = 1, c = 0, d = 0, \pi_a = 0.5, \pi_b = 0.5, \pi_c = 0.5, \pi_d = 0.5\}, \{\theta_2 = \{a = 0, b = 0, c = .5, d = .5, \pi_a = 0.5, \pi_b = 0.5, \pi_c = 0.5, \pi_d = 0.5\}\}$.

D BIQQ Inference Step by Step

Here we use a simplified setup to illustrate the basic mapping from priors on causal types, assignment probabilities, the probative values of clues, and the validity of theoretical propositions to posteriors on these quantities.

Table 7, describes a prior distribution that includes only three possible values for θ—three possible states of the world. Here the parameters include λ, π and ϕ as in our baseline model.

In addition we include a new pair of parameters η, where η^L and η^M are indicators that one of two rival causal accounts, L or M, is correct.

In this example, the researcher considers two theories, L and M. In one state of the world, θ_1, L is true. In θ_1, the treatment also has strong causal effects (ATE = 0.8), all cases receive treatment with equal probability (0.5 each), and the clue is equally likely to be observed for all types and treatment states (clues have no probative value). We can also read the table as indicating that theory L implies no uncertainty about the other parameters, reflected in the fact that L is associated with only one θ. A prior probability of one-third is placed on theory L and all associated parameter values.

Theory M is believed to be true in θ_2 and θ_3, each of which is also given a prior probability of 0.33. Yet theory M is associated with some uncertainty about the other parameters: across θ_2 and θ_3, expected treatment effects vary between weak and negative, and beliefs about the probative value of clues vary.[22]

Note that each admissible state for this example is consistent with a belief that the population incidence of clue K is 0.5, the population incidence of the treatment condition $X = 1$ is 0.5, and the population incidence of outcome $Y = 1$ is 0.5. Uncertainty over three constrained combinations nonetheless allows for relatively complex priors over causal-type proportions, assignment processes, theories, and clue predictions as well as correlations among all of these.

[22]Note that we do not post an index to explicitly associate theories with different values of ϕ, though the association can nevertheless be derived from the probability distribution.

	Parameter	State of the World: θ_1	θ_2	θ_3	Prior expectation
Prior on θ_j:		0.33	0.33	0.33	
Types	λ_a	0	0.1	0.2	0.1
	λ_b	0.8	0.1	0	0.3
	λ_c	0.1	0.4	0.4	0.3
	λ_d	0.1	0.4	0.4	0.3
Assignment Propensities	π_a	0.5	0.5	0.5	0.5
	π_b	0.5	0.5	0.5	0.5
	π_c	0.5	0.1	0.4	0.33
	π_d	0.5	0.9	0.6	0.67
Clues	ϕ_{a0}	0.5	0.9	0.6	0.67
	ϕ_{a1}	0.5	0.1	0.4	0.33
	ϕ_{b0}	0.5	0.9	0.6	0.67
	ϕ_{b1}	0.5	0.1	0.4	0.33
	ϕ_{c0}	0.5	0.1	0.4	0.33
	ϕ_{c1}	0.5	0.9	0.6	0.67
	ϕ_{d0}	0.5	0.1	0.4	0.33
	ϕ_{d1}	0.5	0.9	0.6	0.67
Theory	η^L	1	0	0	0.33
	η^M	0	1	1	0.67
Values implied by priors:					
Population incidence of clue		0.5	0.5	0.5	0.5
Population incidence of treatment		0.5	0.5	0.5	0.5
Population incidence of outcome		0.5	0.5	0.5	0.5
ATE		0.8	0	-0.2	0.2

Table 7: Illustration (Part 1 of 3): Illustration of prior beliefs over potential outcomes, assignment, and theoretical validity. In this illustration it is assumed that the researcher starts out uncertain over only three combinations of parameters.

Given the priors described in Table 7, researchers can use Bayes' rule to form posteriors on all the parameters listed in Table 7, for any realization of the data.

Table 8 shows the possible posteriors for a design in which a researcher collects data on a *single* case randomly drawn from the population, and observes the value taken by X, Y, and K.

The table shows the posterior distribution of weights that we would then place on each state of the world θ_j. Each row represents one possible set of observed

X, Y, and K values for the selected case. Within each row, the first sub-row indicates the prior probability, under each θ_j, that this set of values would be observed. The second sub-row provides the *posterior* weight we then place on each θ_j if that set of observations is in fact made.

	θ_1	θ_2	θ_3	Total
Prior on θ_j	0.33	0.33	0.33	1
Prob $X=1, Y=1, K=1$, for each θ_j (type b or d)	0.23	0.33	0.14	0.23
Posterior on θ_j	0.32	0.47	0.21	1
Prob $X=0, Y=1, K=1$, for each θ_j (type a or d):	**0.03**	0.05	0.12	0.07
Posterior on θ_j	0.13	0.25	0.63	1
Prob $X=1, Y=0, K=1$, for each θ_j (type a or c)	0.03	0.04	0.14	0.07
Posterior on θ_j	0.12	0.2	0.67	1
Prob $X=0, Y=0, K=1$, for each θ_j (type b or c):	0.23	0.08	0.1	0.13
Posterior on θ_j	0.56	0.2	0.24	1
Prob $X=1, Y=1, K=0$, for each θ_j (type b or d):	0.23	0.08	0.1	0.13
Posterior on θ_j	0.56	0.2	0.24	1
Prob $X=0, Y=1, K=0$, for each θ_j (type a or d):	0.03	0.04	0.14	0.07
Posterior on θ_j	0.12	0.2	0.67	1
Prob $X=1, Y=0, K=0$, for each θ_j (type a or c):	0.03	0.05	0.12	0.07
Posterior on θ_j	0.13	0.25	0.63	1
Prob $X=0, Y=0, K=0$, for each θ_j (type b or c):	0.23	0.33	0.14	0.23
Posterior on θ_j	0.32	0.47	0.21	1

Table 8: Illustration (Part 2 of 3). Given the prior information provided in Table 7, the probability of each combination of X, Y, and K observations can be calculated for each possible state of the world (θ_j). This in turn allows for the calculation of the posterior probability of each state of the world for each possible pattern of data using Bayes' rule.

Table 9 provides our posteriors on all parameters, with each column representing one possible realization of the data for a single case.

A few implications of this simple exercise are apparent from Table 9. *First, the example demonstrates that observing data—even for a single case—allows for updating on all parameters of interest:* causal effects, the merits of different theoretical accounts, assignment propensities, and the probative value of clues. Note, for instance, how our belief about the ATE falls dramatically when we observe a single $X=1$ case in which $Y=0$ and the clue is present. This shift results from the fact that this pattern of evidence was much more likely under

θ_3 than under θ_1 or θ_2. With the shift in confidence toward θ_3, we in turn place more weight on that vector's constituent beliefs in a relatively high λ_a and a low λ_b, and this updating brings down our estimate of the ATE $(\lambda_b - \lambda_a)$.

The exercise demonstrates the sometimes counter-intuitive nature of Bayesian updating. In particular, we see that *a belief can gain support from data that are unlikely under that belief—as long as those data are even more unlikely under the alternatives.* Consider, for instance, what happens when we observe $X = Y = 1$ and $K = 0$. This observation pushes beliefs in theory L above 50%, even though such an observation was expected under θ_1, in which L was believed to be true, with only a 0.23 probability. Yet this observation was even *more* unexpected under theory M (θ_2 and θ_3).

An important implication for case selection also follows. Theory L might seem to have had only a modest stake in a case with $X = Y = 1$ and $K = 0$, which it predicted to be somewhat but not extremely unlikely. However, even when a belief makes no strong prediction about a particular configuration of X, Y, or K values, *such a case will have large implications for the belief as long as the alternative implies a sharp and divergent prediction about the case's likelihood of occurring.*

Further, the example shows that, *in some situations, the most significant updating occurs over analytical assumptions rather than substantive causal effects.* For instance, where $X = Y = K = 0$ or $X = Y = K = 1$, there is a small loss in confidence in theory L relative to theory M and a small increase in the expected treatment effect. But there is a larger gain in confidence in the probative value of clues. This latter updating occurs because of a substantial shift from θ_3 to θ_2, which both contain theory M, and which yields a shift in support between the alternative clue probabilities that we believed might be associated with M.

Event		X=1, Y=1, K=1	X=0, Y=1, K=1	X=1, Y=0, K=1	X=0, Y=0, K=1	X=1, Y=1, K=0	X=0, Y=1, K=0	X=1, Y=0, K=0	X=0, Y=0, K=0
Probability of event:		0.233	0.066	0.067	0.134	0.134	0.067	0.066	0.233
	λ_a	0.09	0.15	0.15	0.07	0.07	0.15	0.15	0.09
	λ_b	0.31	0.13	0.12	0.47	0.47	0.12	0.13	0.31
	λ_c	0.30	0.36	0.36	0.23	0.23	0.36	0.36	0.30
	λ_d	0.30	0.36	0.36	0.23	0.23	0.36	0.36	0.30
	π_a	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	π_b	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	π_c	0.29	0.34	0.35	0.40	0.40	0.35	0.34	0.29
Posterior on:	π_d	0.71	0.66	0.65	0.60	0.60	0.65	0.66	0.71
	ϕ_{a0}	0.71	0.66	0.65	0.60	0.60	0.65	0.66	0.71
	ϕ_{a1}	0.29	0.34	0.35	0.40	0.40	0.35	0.34	0.29
	ϕ_{b0}	0.71	0.66	0.65	0.60	0.60	0.65	0.66	0.71
	ϕ_{b1}	0.29	0.34	0.35	0.40	0.40	0.35	0.34	0.29
	ϕ_{c0}	0.29	0.34	0.35	0.40	0.40	0.35	0.34	0.29
	ϕ_{c1}	0.71	0.66	0.65	0.60	0.60	0.65	0.66	0.71
	ϕ_{d0}	0.29	0.34	0.35	0.40	0.40	0.35	0.34	0.29
	ϕ_{d1}	0.71	0.66	0.65	0.60	0.60	0.65	0.66	0.71
	η^L	0.32	0.13	0.12	0.56	0.56	0.12	0.13	0.32
	η^M	0.68	0.87	0.88	0.44	0.44	0.88	0.87	0.68
Implied posterior on:	ATE	0.22	-0.02	-0.04	0.40	0.40	-0.04	-0.02	0.22

Table 9: Illustration (Part 3 of 3) Posteriors on states of the world, (θ_j), as calculated in Table 8 given data (X, Y, and K), imply posteriors over all quantities of interest.

E Code to Implement baseline BIQQ model using Stan

Here we provide R code that calls Stan, and builds a BIQQ function that generates draws from a posterior distribution given user supplied priors and data. This code is for the baseline model in "Mixing Methods," in which a fixed number of cases are assumed to be randomly sampled from a population.

```
# Sample code to implement Bayesian Integration of Quantitative and Qualitative data.
# Provided as Supplementary Material for:
# Macartan Humphreys and Alan M Jacobs, 2015. "Mixing Methods: A Bayesian Approach."
#
# This code has been developed for R (Version 3.1.1).
# Requirements: rstan package (Version 2.4.0).

library(rstan)

# Code used to build Stan model
biqq.stan.code <- '

# User supplies the following data:
data {

# Counts of the different XY and XYK outcomes observed:

int<lower=0> XYK[8];    # summary of N of XYK outcomes
int<lower=0> XY[4];     # summary of N of XY outcomes

# And the hyperparameters for the prior:

vector[4] alpha_prior;  # Dirichlet shape parameters for share of types (abcd)

vector[4] pi_alpha;     # Beta shape parameters for assignment probabilities
vector[4] pi_beta ;

vector[4] phi0_alpha ;  # Beta shape parameters for probative value of clues
vector[4] phi0_beta ;
vector[4] phi1_alpha ;
```

```
vector[4] phi1_beta ;
}

# The parameters to be modelled are defined:

parameters {

simplex[4]  abcd;             # The share of abcd types in the population
                              # (constrained to sum to 1)

real<lower=0,upper=1> pi_a;   # The probabilities of assignment
real<lower=0,upper=1> pi_b;   # (constrained between 0 and 1 inclusive)
real<lower=0,upper=1> pi_c;
real<lower=0,upper=1> pi_d;

real<lower=0,upper=1> phi_a0; # The probative values of clues for each type
real<lower=0,upper=1> phi_b0; # (constrained between 0 and 1 inclusive)
real<lower=0,upper=1> phi_c0;
real<lower=0,upper=1> phi_d0;

real<lower=0,upper=1> phi_a1;
real<lower=0,upper=1> phi_b1;
real<lower=0,upper=1> phi_c1;
real<lower=0,upper=1> phi_d1;
}

# The multinomial event probabilities are calculated:

transformed parameters {

simplex[4] w_XY;
simplex[8] w_XYK;

w_XY[1] <- (1-pi_b)*abcd[2]            + (1-pi_c)*abcd[3]           ; # Pr(00*)
w_XY[2] <- (1-pi_a)*abcd[1]            + (1-pi_d)*abcd[4]           ; # Pr(01*)
w_XY[3] <- pi_a*abcd[1]                + pi_c*abcd[3]               ; # Pr(10*)
w_XY[4] <- pi_b*abcd[2]                + pi_d*abcd[4]               ; # Pr(11*)

w_XYK[1]<- (1-pi_b)*(1-phi_b0)*abcd[2] + (1-pi_c)*(1-phi_c0)*abcd[3]; # Pr(000)
```

```
w_XYK[2]<- (1-pi_b)*phi_b0*abcd[2]     + (1-pi_c)*phi_c0*abcd[3]    ; # Pr(001)

w_XYK[3]<- (1-pi_a)*(1-phi_a0)*abcd[1] + (1-pi_d)*(1-phi_d0)*abcd[4]; # Pr(010)
w_XYK[4]<- (1-pi_a)*phi_a0*abcd[1]     + (1-pi_d)*phi_d0*abcd[4]    ; # Pr(011)

w_XYK[5]<- pi_a*(1-phi_a1)*abcd[1]     + pi_c*(1-phi_c1)*abcd[3]    ; # Pr(100)
w_XYK[6]<- pi_a*phi_a1*abcd[1]         + pi_c*phi_c1*abcd[3]        ; # Pr(101)

w_XYK[7]<- pi_b*(1-phi_b1)*abcd[2]     + pi_d*(1-phi_d1)*abcd[4]    ; # Pr(110)
w_XYK[8]<- pi_b*phi_b1*abcd[2]         + pi_d*phi_d1*abcd[4]        ; # Pr(111)
}

# The parameters are then modeled as follows:

model {

abcd   ~ dirichlet(alpha_prior);             # Priors for abcd types

pi_a   ~ beta(pi_alpha[1], pi_beta[1]);      # Priors for assignment probs
pi_b   ~ beta(pi_alpha[2], pi_beta[2]);
pi_c   ~ beta(pi_alpha[3], pi_beta[3]);
pi_d   ~ beta(pi_alpha[4], pi_beta[4]);

phi_a0 ~ beta(phi0_alpha[1], phi0_beta[1]);  # Priors for clues
phi_a1 ~ beta(phi1_alpha[1], phi1_beta[1]);

phi_b0 ~ beta(phi0_alpha[2], phi0_beta[2]);
phi_b1 ~ beta(phi1_alpha[2], phi1_beta[2]);

phi_c0 ~ beta(phi0_alpha[3], phi0_beta[3]);
phi_c1 ~ beta(phi1_alpha[3], phi1_beta[3]);

phi_d0 ~ beta(phi0_alpha[4], phi0_beta[4]);
phi_d1 ~ beta(phi1_alpha[4], phi1_beta[4]);

XY     ~ multinomial(w_XY);                  # Likelihood Part 1
XYK    ~ multinomial(w_XYK);                 # Likelihood Part 2
}
'
```

```
# Compile the model, giving it simple starting values
ones <- rep(1,4)

biqq.stan.model <- stan(

# The code for the model
     model_code = biqq.stan.code,

# The initial data
     data       = list(XY  = c(  1,     # Number of 00*
                                 1,     # Number of 01*
                                 1,     # Number of 10*
                                 1      # Number of 11*
                                   ),
                       XYK = c(  1,  1, # Number of 000 & 001
                                 1,  1, # Number of 010 & 011
                                 1,  1, # Number of 100 & 101
                                 1,  1  # Number of 110 & 111
                                       ),

# Shape parameters for flat priors:
                       alpha_prior = ones,
                       pi_alpha    = ones, pi_beta     = ones,
                       phi0_alpha  = ones, phi0_beta   = ones,
                       phi1_alpha  = ones, phi1_beta   = ones),

# Run a low number of iterations to warm the model up:
                 iter   = 100,
                 chains = 4
)

# BIQQ function for drawing from posterior distribution using STAN:

biqq <- function(fit         = biqq.stan.model,

# Define the default arguments of the function:
                 XY          = c(  1,  1,  1, 1),
                 XYK         = c(  1,  1,
```

```
                                    1,  1,
                                    1,  1,
                                    1,  1),
                    alpha_prior = ones,
                    pi_alpha    = ones, pi_beta     = ones,
                    phi0_alpha  = ones, phi0_beta   = ones,
                    phi1_alpha  = ones, phi1_beta   = ones,
                    iter        = 1000, chains      = 4,
                    warmup      = 100,  seed        = 100
)  {
# Function takes inputs on data and priors and implements the baseline BIQQ model using rstan

# Define the model inputs, based on user-supplied or default arguments
     data <- list(
          XY          = XY,
          XYK         = XYK,
          alpha_prior = alpha_prior,
          pi_alpha    = pi_alpha,    pi_beta     = pi_beta,
          phi0_alpha  = phi0_alpha,  phi0_beta   = phi0_beta,
          phi1_alpha  = phi1_alpha,  phi1_beta   = phi1_beta
     )

# Use Stan to sample from the posterior distribution:
     posterior <-
          stan(fit     = fit,
               data    = data,
               iter    = iter,
               chains  = chains,
               warmup  = warmup,
               seed    = seed
               )

# Display the results
     return(posterior)
}

# Demonstration 1:
# Simple demonstration using default values
biqq()
```

```
# Demonstration 2:
# In this example we assume that we know that b types leave a clue in the
# treatment condition. Compare inferences on lambda_b and on q_d for cases
# where clues are or are not seen in the X=Y=1 cases
biqq(XYK = c(5,0, 0,0, 0,0, 0,5), XY = rep(0,4), q1_alpha    = c(1,20,1,1))
biqq(XYK = c(5,0, 0,0, 0,0, 5,0), XY = rep(0,4), q1_alpha    = c(1,20,1,1))
```

F Extensions

F.1 Alternative data generating processes

The likelihood function contains information about how cases are selected for the overall study and also how cases are selected for qualitative analysis. In this appendix we demonstrate how the likelihood function can change given different research strategies.

F.1.1 Independent case selection strategy

In the text we considered a situation in which the researcher examines a fixed number of cases for clue information. An alternative strategy that produces a simpler likelihood is one in which each case is selected for within-case data gathering with some independent probability. The likelihood below introduces a case selection probability κ_{xy} that covers this situation and allows for the possibility that selection probabilities are different for different X, Y combinations.

We assume again that X, Y data is observed for all n cases under study, but that K data may be sought for only a random subset of these (we use the wildcard symbol "$*$" to denote that the value of the clue is unknown). Unlike in our baseline model, however, the number of cases for which clue data is sought is not fixed. We let n_{xyk} denote the number of cases with each possible data realization. Then, assuming the data are independently and identically distributed, the likelihood is:

$$\begin{aligned}\Pr(\mathcal{D}|\theta) \quad &= \text{Multinomial}((n_{000}, n_{001}, n_{00*}, n_{010}, n_{010}, n_{01*}, n_{100}, n_{101}, n_{10*}, n_{110}, n_{111}, n_{11*}) \\ &\qquad |n, (w_{000}, w_{001}, w_{00*}, w_{010}, w_{010}, w_{01*}, w_{100}, w_{101}, w_{10*}, w_{110}, w_{111}, w_{11*}))\end{aligned}$$

where the event probabilities are now given by:

$$\begin{pmatrix} w_{000} \\ w_{001} \\ \vdots \\ w_{11*} \end{pmatrix} = \begin{pmatrix} \lambda_b(1-\pi_b)\kappa_{00}(1-\phi_{b0}) + \lambda_c(1-\pi_c)\kappa_{00}(1-\phi_{c0}) \\ \lambda_b(1-\pi_b)\kappa_{00}\phi_{b0} + \lambda_c(1-\pi_c)\kappa_{00}\phi_{c0} \\ \vdots \\ \lambda_b\pi_b(1-\kappa_{11}) + \lambda_d\pi_d(1-\kappa_{11}) \end{pmatrix}$$

We use a Greek symbol to denote the case selection probabilities to highlight that these may also be unknown and an object of inquiry, entering into the vector of parameters, θ.

F.1.2 Non-random XY Sample Selection

While we have assumed in our baseline model that cases are selected at random for quantitative analysis, this need not be the case. Suppose instead that each case of type j is selected into the study with probability ρ_j. In that situation, assuming independent selection of cases for qualitative analysis, the likelihood function is now:

$$\begin{aligned} \Pr(\mathcal{D}|\theta) \quad &= \text{Multinomial}((n_{000}, n_{001}, n_{00*}, n_{010}, n_{010}, n_{01*}, n_{100}, n_{101}, n_{10*}, n_{110}, n_{111}, n_{11*}) \\ &\qquad |n, (w_{000}, w_{001}, w_{00*}, w_{010}, w_{010}, w_{01*}, w_{100}, w_{101}, w_{10*}, w_{110}, w_{111}, w_{11*})) \end{aligned}$$

where the event probabilities are now, given by:

$$\begin{pmatrix} w_{000} \\ w_{001} \\ \vdots \\ w_{11*} \end{pmatrix} = \begin{pmatrix} \frac{\rho_b\lambda_b}{\rho_a\lambda_a+\rho_b\lambda_b+\rho_c\lambda_c+\rho_d\lambda_d}(1-\pi_b)\kappa_{00}(1-\phi_{b0}) + \frac{\rho_c\lambda_c}{\rho_a\lambda_a+\rho_b\lambda_b+\rho_c\lambda_c+\rho_d\lambda_d}(1-\pi_c)\kappa_{00}(1-\phi_{c0}) \\ \frac{\rho_b\lambda_b}{\rho_a\lambda_a+\rho_b\lambda_b+\rho_c\lambda_c+\rho_d\lambda_d}(1-\pi_b)\kappa_{00}\phi_{b0} + \frac{\rho_c\lambda_c}{\rho_a\lambda_a+\rho_b\lambda_b+\rho_c\lambda_c+\rho_d\lambda_d}(1-\pi_c)\kappa_{00}\phi_{c0} \\ \vdots \\ \frac{\rho_b\lambda_b}{\rho_a\lambda_a+\rho_b\lambda_b+\rho_c\lambda_c+\rho_d\lambda_d}\pi_b(1-\kappa_{11}) + \frac{\rho_d\lambda_d}{\rho_a\lambda_a+\rho_b\lambda_b+\rho_c\lambda_c+\rho_d\lambda_{11}}\pi_d(1-\kappa_{11}) \end{pmatrix}$$

We use a Greek symbol for the selection probabilities to highlight that these probabilities may be unknown and could enter into the set of parameters of

interest, θ.

F.1.3 Conditional random case selection

Finally, consider the likelihood for a design in which a researcher selects cases in which to search for clues as a function of the X, Y values. This is a somewhat harder situation because the size of each X, Y group will be stochastic. Let $n_{xy} = n_{xy0} + n_{xy1} + n_{xy*}$ denote the number of cases with particular values on X and Y, and let $n_{XY} = (n_{00}, n_{01}, n_{10}, n_{11})$ denote the collection of n_{xy} values.

Say now that, conditional on the X, Y observations, a researcher sets a target of $k_{xy}(n_{XY})$ cases for clue examination (note here that the number of clues sought for a particular X, Y combination can be allowed to depend on what is observed across all X, Y combinations). Then the likelihood is:

$$\text{Multinomial}(n_{XY}|n, w_{XY}) \prod_{x\in\{0,1\},y\in\{0,1\}} \text{Binom}(n_{xy1}|k_{xy}(n_{xy}), \psi_{xy})$$

The multinomial part of this expression gives the probability of observing the particular X, Y combinations; the event probabilities for these depend on λ and π only — for example $w_{11} = \lambda_b\pi_b + \lambda_d\pi_d$. The subsequent binomials give the probability of observing the clue patterns conditional on searching for a given number of clues ($k_{xy}(n_{xy})$) and given an event probability ψ_{xy} for seeing a clue given that the clue is sought for an x, y combination; thus for example:

$$\psi_{11} = \frac{\lambda_b\pi_b}{\lambda_b\pi_b + \lambda_d\pi_d}\phi_{b1} + \frac{\lambda_d\pi_d}{\lambda_b\pi_b + \lambda_d\pi_d}\phi_{d1}$$

F.2 Multiple Causes

Here we provide additional intuition for how the BIQQ framework can handle multiple causal variables characterized by either equifinality (multiple potential causes of the same outcome) or interaction effects. The core approach is to expand the number of types to take into account the more complex combinations of causal conditions for which potential outcomes must now be defined.

Table 10 displays the set of potential outcomes for two binary causal variables. With two causes, X_1 and X_2, we now have 16 types as defined by the potential outcomes under alternative combinations of causal conditions.

Type	Label	$(Y\|X_1 = 0, X_2 = 0)$	$(Y\|X_1 = 1, X_2 = 0)$	$(Y\|X_1 = 0, X_2 = 1)$	$(Y\|X_1 = 1, X_2 = 1)$
1	chronic	0	0	0	0
2	jointly-beneficial	0	0	0	1
3	2-alone-beneficial	0	0	1	0
4	2-beneficial	0	0	1	1
5	1-alone-beneficial	0	1	0	0
6	1-beneficial	0	1	0	1
7	any-alone-beneficial	0	1	1	0
8	any-beneficial	0	1	1	1
9	any-adverse	1	0	0	0
10	any-alone-adverse	1	0	0	1
11	1-adverse	1	0	1	0
12	1-alone-adverse	1	0	1	1
13	2-adverse	1	1	0	0
14	2-alone-adverse	1	1	0	1
15	jointly-adverse	1	1	1	0
16	destined	1	1	1	1

Table 10: Types given two treatments (or one treatment and one covariate)

Taking interaction effects first, Type 3 (2-alone-beneficial), for instance, is a type in which $X_2 = 1$ causes $Y = 1$ only when $X_1 = 0$, and not when $X_1 = 1$. The hypothesis of no-interaction-effects is the hypothesis that all cases are of type 1, 4, 6, 11, 13, or 16 (that is chronic, destined, 1-beneficial, 2-beneficial, 1-adverse, or 2-adverse). Note that the binary outcome framework excludes possibilities, such as two countervailing or two additive effects of X_1 and X_2.

Turning now to equifinality, in the simple typological setup in the main paper, the difference between a *b* and a *d* type already implies equifinality: for a *b* type, the positive outcome was caused by treatment; for a *d* type, the same outcome is caused by some other (unspecified) cause. Table 10, however, explicitly builds multiple causes into the framework. Suppose, for instance, that we have two cases, one of Type 4 and one of Type 6, and that $X_1 = X_2 = 1$ in both cases. For both cases, we will observe $Y = 1$. However, for the Type 4 case,

the outcome was caused by X_2 (in the sense that it would not have occurred if X_2 had been 0, but would have even if X_1 was 0) whereas the outcome in the Type 6 case was caused by X_1, but not by X_2.

The parameters in the model would now be defined in terms of these 16 types and two causal variables.

For estimating population-level causal effects, we would state priors about the population proportions of these 16 types.

Assignment probabilities would be expressed, separately for each independent variable, as the probability that each type is assigned to the value 1 on that variable, yielding 32 π values in total.

Clue probabilities, finally, would be supplied for each type. In principle, these ϕ values could be made conditional on the combination of X_1 and X_2 values, potentially yielding 64 ϕ values. In practice, greater structure might facilitate analysis. For example, if a given clue's likelihood depends only on the value of one of the independent variables, rather than that of both, this greatly reduces the required number of ϕ priors.

G Notes on Applications

G.1 Notes on Application 1

For the Kreuzer application we treat the ϕ values as known (or at least as known with a very high level of certainty). Unfortunately we do not have, nor do Kreuzer's analyses provide, an empirical basis for claims about ϕ values. Instead we present here an approach to "filling in the values" in the absence of empirical information, based on reasoning about the theoretical logic and background information about the world. The conclusions in the analysis can be understood to be conditional on the *phi* values given below, though not on the specific the reasoning that led to these values.

We begin with a few general comments:

- We refer throughout to the reasons for supporting PR that Boix (1999) attributes to non-socialist parties under conditions of high left threat as "electoral engineering" (EE) motives.
- We assume that there exist non-EE reasons why a governing coalition of parties might prefer PR over single-member districts. These might include, for instance, normative beliefs that PR is more democratic or a response to mass demands for PR.
- We assume the socialist party never constitutes by itself a winning coalition for enacting electoral reform, but that the non-socialist parties may in some cases require socialist support to enact reform (see a-type cases).
- In our illustration, the "clue" is found if all three of the process-tracing tests in Kreuzer (2010) are passed.

$\phi_{a0} = 0.1$: Probability of the clue for an untreated a type

An a (adverse) case is one in which strong left threat will prevent PR from being adopted. We assume that adverse effects happen via the strategic calculations and veto power of the left.

When the socialists are electorally strong and the right is divided, the left seeks to *stop* any switch to PR since plurality rules now advantage them. Adverse cases are those in which the non-socialist parties cannot enact electoral reform without the support of the socialists — i.e., there the left has the institutional leverage to veto a move to PR initiated by the right.

In an untreated "adverse" case ($X = 0, Y = 1$), the right does not have an EE motive to favor PR. The left does favor PR for EE reasons, but by assumption does not by itself form a winning coalition for enactment. Thus, PR must emerge via support from non-socialist parties that is motivated by non-EE reasons. However, PR can emerge without the support of all non-socialist parties; all that is needed is some winning coalition.

The question for generating ϕ_{a0}, then, is in what proportion of untreated a cases will all non-socialist parties favor a move to PR and do so shortly after the expansion of the franchise. In other words, in what proportion of such cases will there be non-EE motivations for PR that apply to all non-socialist parties at this historical point in time, thus generating all three of Kreuzer's clues?

Assuming that this is very unlikely, we assign a ϕ_{a0} value of 0.1.

$\phi_{a1} = 0.95$: Probability of the clue for a treated a type

Under the adverse-effect logic above, we should see the clues in a treated a case with a greater probability than in a treated b case (see below). This is because there are two logics that might generate the clue in a treated a case.

1. In an a case with high left threat, right parties have the same EE incentives that they do in a treated b case, where the clue probability is set at 0.9.
2. In addition, as explained for ϕ_{a0}, a cases are those in which a winning coalition – which must include non-socialists – has non-EE reasons to prefer PR. Further, we have assumed above that with 0.1 probability those non-EE reasons apply to all non-socialist parties and emerge with the expansion of suffrage, thus generating the clue.

We assume that reason 2 for the clue coexists with reason 1 in half of cases with reason 2. Thus, the total probability of seeing the clue in a treated a type is $0.9 + 0.1 \times 0.5 = 0.95$.

$\phi_{b0} = 0.1$: Probability of the clue for an untreated b type

The lack of a left threat removes EE motives for the ruling parties under Boix's logic, thus reducing the likelihood of observing the clue in any untreated case. Further, however, b (beneficial) cases are those in which no move to PR will happen without high left threat. Thus, b cases are those in which there is also no non-EE reason for adopting PR that obtains for any winning coalition of parties.

Under Boix's logic, therefore, it is very unlikely that we would see unanimous non-socialist party support for a move to PR for a b case when left threat is low. We thus set ϕ_{b0} to 0.1, allowing only for a margin of measurement error.

$\phi_{b1} = 0.9$: Probability of the clue for a treated b type

A treated b case is one in which PR emerges under high left threat for precisely the EE reasons that Boix outlines. As Kreuzer's clues are tightly linked to Boix's logic, we place a high probability (0.9) on observing the clue for such cases, allowing only for a margin of measurement error (e.g., the possibility that we cannot find evidence of a right party's support for PR even when that support was in fact present).

$\phi_{c0} = 0.05$: Probability of the clue for an untreated c type

Cases of c (chronic) type are those in which PR will never emerge, even if left threat is great. We posit that two processes may cause a case to be chronic:

1. Institutional obstacles: the right parties may prefer PR for EE or other reasons, but some other actor has veto power over the decision. Here we might observe the clue; or,

2. Preferences: a significant share of right parties have reasons, outside of the logic of electoral engineering, for opposing PR. Here the clue should never be observed.

We assume that each process is responsible for chronicness in 0.5 of all chronic cases.

In an untreated case, there can be no EE reasons for the clue to emerge.

There could, however, be non-EE reasons for unanimous right-party support in (and only in) the "institutionally" chronic cases. We have assumed that 0.5 of chronic cases are "institutional." Then, consistent with the assumptions made for a cases about the prevalence of non-EE reasons, we posit that in 0.1 of all "institutional" chronic cases, there are non-EE reasons that generate unanimous right-party support for PR.

The total probability of observing the clue in an untreated c case, then, is $0.1 \times 0.5 = 0.05$.

$\phi_{c1} = 0.475$: Probability of the clue for a treated c type

In treated chronic cases of the "preference" variety (0.5 of c cases), the clue will never emerge.

However, there are two ways in which the clue could emerge in institutionally chronic cases:

1. As discussed above, 0.1 of such cases will have non-EE reasons for generating the clue.
2. In addition, however, in a treated institutionally Chronic case, the clue could emerge via the same logic that generates EE motives in 0.9 of treated b cases.

As we do for treated a cases, we assume that reason 2 for the clue coexists with reason 1 in half of cases with reason 2. We thus get 0.95 of institutionally chronic cases having the clue, giving $\phi_{c1} = 0.5 \times 0.95 = 0.475$.

$\phi_{d0} = 0.3$: Probability of the clue for an untreated d type

In d (destined) cases, PR will always emerge, regardless of whether left threat is high.

This means that in d cases some winning coalition always has a non-EE reason sufficient to support PR. This coalition might:

1. include a unified right and take place at the time of suffrage expansion, in which case we will observe the clue, or
2. be a coalition of part of the right and the left *or* happen at some other time, in either of which case we will not observe the clue.

So the question is: how common will situation 1 be in d cases? Since d cases are those that always generate PR, we assume that this is because the normative and other non-EE pressures to adopt PR are inherently stronger in these cases than in the other types. We thus assume a higher probability that such pressures will apply to all non-socialist parties, and we posit that such pressures are likely to be strongest of all at the "democratic moment" that produces the universal male franchise.

We thus put the probability of observing the clue in such a case at the low-moderate value of 0.3.

$\phi_{d1} = 0.5$: Probability of the clue for a treated d type

We see two processes that might produce the clue in a treated d case.

1. In a d case, any EE motives for supporting PR must be causally redundant: there must be a winning coalition for PR for reasons outside Boix's framework. As discussed for untreated d cases, any winning coalition generated by non-EE reasons might not include all right parties. We have assumed above that non-EE reasons affect all non-socialist parties following the suffrage expansion, and thus produce the clue, in 0.3 of cases.

2. At the same time, in a treated d case some EE strategic considerations may also operate on right parties. Thus, high left threat in a d case may bring the remaining right parties (those unmoved by non-EE motives) into support of PR, generating the clue. We assume that this configuration is relatively rare, obtaining in 0.2 of cases.

We add the probabilities of the two processes together, yielding $\phi_{d1} = 0.3 + 0.2 = 0.5$.

G.2 Notes on Application 2

G.2.1 Case selection

To integrate the inferences from Ross's (2004) analysis with the wider Collier and Hoeffler (2004) population of cases we faced two challenges that affected the set of cases that we could use for this analysis.

First, we had to create a binary measure of natural resource wealth. This is done implicitly in Ross's analysis but not explicitly. Consulting Collier and Hoeffler's data, we could find no threshold that separated cases with high and low levels of natural resources consistent with Ross's analysis. Most cases in the Ross sample had relatively high levels of natural resources, but one, Afghanistan, did not — even though experts linked natural resources to the conflict in this case. This coding raises a subtle issue: it is possible that X is causally linked to Y in a given case even though X takes on an exceptionally low value in the case. Should such a case count as evidence that natural resource wealth causes conflict or not? We take it that the general proposition X causes Y should be interpreted as a claim that some minimal *level* of X causes Y. Under this reading cases with very low levels of X in which $Y = 1$ are not supportive of the proposition that X causes Y, even if paths can be found between X and Y. Under this interpretation, we selected a threshold level of resource dependence that separated cases in a way as consistent with Ross's coding as possible; in doing so we were forced to omit Afghanistan from the qualitative analysis.

Second, we had to identify the population of cases from which Ross's subset were drawn. We discuss one selection criterion used by Ross — expert judgments — in the main text. A second selection criterion for Ross was the timing of the conflicts. Ross selected cases of conflicts that started *or were ongoing* in a given time period. Including cases that began before the period in question, on the grounds that they had a conflict that continued into the period, makes identification of the population of cases difficult because we do not know whether a given case that did not have a conflict prior to the period would have had a conflict long enough to take it into the sample, conditional on having had a conflict. To deal with this problem, we limited the population to all cases in which there was not a conflict ongoing in the 1990s and interpreted the outcome variable as an indicator of a conflict start in this period.

G.2.2 Probative values of clues and graphical representation of posteriors

For our analyses we considered two sets of values (informative and uninformative) for each of two types of clue, K_1 and K_2. For K_1 uninformative, we supposed that $\phi^1_{b1} = \phi^1_{d1} = 0.5$; for K_1 informative, we set $\phi^1_{b1} = .9$, $\phi^1_{d1} = 0.3$, making clue 1 a strong (i.e., difficult) hoop test. For K_2 uninformative, we supposed that $\phi^2_{b1} = \phi^2_{d1} = 0.5$; for K_2 informative, we set $\phi^2_{b1} = .99$, $\phi^2_{d1} = 0.01$, making clue 2 strongly doubly decisive.

In the text we provide summary statistics of the posterior distributions under these different assumptions of the informativeness of the clues. Figure 7 presents information on these posteriors graphically for each set of assumptions.

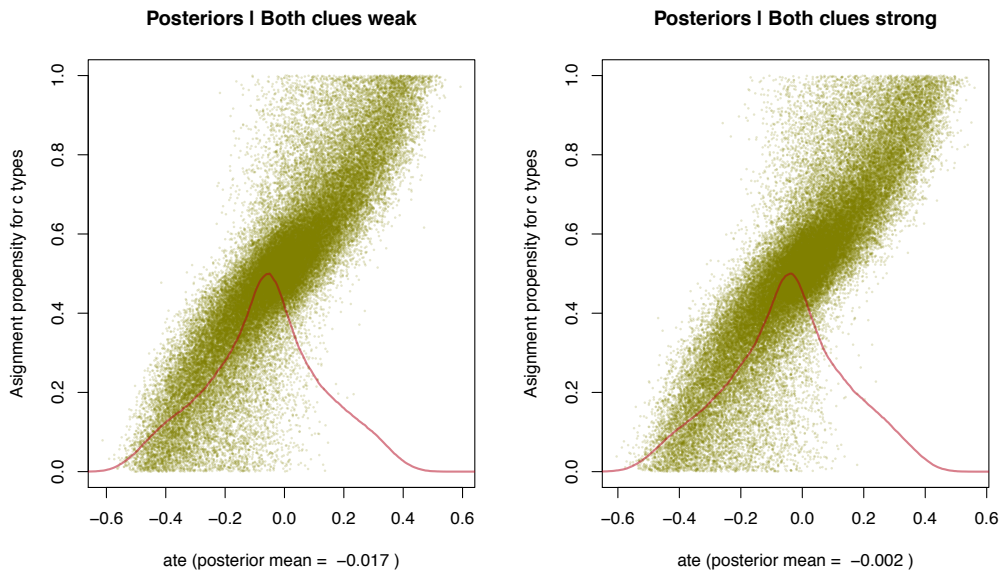

Figure 7: Joint distribution of posterior beliefs on the average effect of natural resources on conflict (i.e., $\lambda_b - \lambda_a$) and the beliefs regarding differential assignment probabilities, given clues of varying probative value. Plots also show the marginal posterior on the average causal effect.

H Notes on Simulations

Here we provide statistical details and some further interpretation for the paper's simulations assessing the benefits of different designs conditional on different priors on: the probative value of clues, the heterogeneity of causal effects, uncertainty regarding assignment probabilities, and uncertainty regarding the probative value of clues. Table 11 provides details on all parameters used in simulations; Table 12 provides detail on the number of runs, iterations, and related information used in the estimation.

H.1 Probative values

For these simulations we simultaneously vary the probative value for tests for all X, Y combinations. Specifically, we vary the differences between ϕ_{b0} and ϕ_{c0} (for $X = Y = 0$ cases), between ϕ_{a0} and ϕ_{d0} (for $X = 0, Y = 1$ cases); between ϕ_{a1} and ϕ_{c1} (for $X = 1, Y = 0$ cases); and between ϕ_{b1} and ϕ_{d1} (for $X = Y = 1$

cases). For each X, Y combination, we compare the relevant ϕ pairs across values of $(.5, .5)$ (no probative value), $(.25, .75)$ (middling probative value) and $(0.01, 0.99)$ (strong probative value). Using the definition of probative value (PV) (see footnote 5), these correspond to cases with probative value of 0, .5, and close to 1 respectively.

H.2 Effect heterogeneity

We note that heterogeneity makes going "wide" relatively more beneficial for two reasons. First, when all cases are affected either positively or negatively, all of the information needed to identify types is provided by information on X and Y. If $X = Y$, then a case was (or could have been) positively affected; if $X \neq Y$ then a case was (or could have been) negatively affected. In this extreme case of maximal heterogeneity, causal process information provides no additional inferential gains. Where there is high homogeneity, on the other hand, the core difficulty is distinguishing a and b types, from c and d types. Then, the information contained in clues may provide greater benefits (see Table 2). Second, the more heterogeneous effects are across cases, the less we learn about *population-level* causal effects by getting an individual case right. Thus, again, we would expect greater relative gains to more extensive analysis as heterogeneity increases.

H.3 Uncertainty about assignment processes

Note that in our binary setup, infinite bias cannot arise, and the harm done by uncertainty over selection processes can be more moderate. In this set of simulations, the expected value of π_j is fixed at 0.5 and we vary the variance in π_j between 0 and a maximum of 0.289.

H.4 Uncertainty regarding the probative value of clues

In this experiment, the expected probability that a clue will be observed is set to 0.85 if one hypothesis is right, and 0.15 if the alternative hypothesis is

correct. The simulations vary from a situation in which those probabilities are known with certainty (uncertainty low) to a situation in which the researcher admits the possibility of many possible values of ϕ. Uncertainty is simultaneously varied for all pairs of ϕ values (see §H.1). The displayed results suggest that uncertainty about the probative value of clues plays a muted role in the assessment of optimal strategies.

H.5 Details on simulation experiments

			1: m or n	2: Probative Value			3: Effect Heterogeneity			4: Assignment Uncertainty			5: Clue Uncertainty		
θ	Dist	arg		Low	$\rightarrow$	High	Low	$\rightarrow$	High	Low	$\rightarrow$	High	Low	$\rightarrow$	High
λ_a	Dirichlet	α_a	1	0.20	$\rightarrow$	0.20	0.10	$\rightarrow$	**2.00**	1	$\rightarrow$	1	1	$\rightarrow$	1
λ_b		α_b	1	0.20	$\rightarrow$	0.20	02.10	$\rightarrow$	**4.00**	1	$\rightarrow$	1	1	$\rightarrow$	1
λ_c		α_c	1	0.20	$\rightarrow$	0.20	02.00	$\rightarrow$	**0.10**	1	$\rightarrow$	1	1	$\rightarrow$	1
λ_d		α_d	1	0.20	$\rightarrow$	0.20	02.00	$\rightarrow$	**0.10**	1	$\rightarrow$	1	1	$\rightarrow$	1
π_a	Beta	μ	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50
		σ	0.10	0.10	$\rightarrow$	0.10	0.10	$\rightarrow$	0.10	0.01	$\rightarrow$	**0.289**	0.10	$\rightarrow$	0.10
π_b	Beta	μ	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50
		σ	0.10	0.10	$\rightarrow$	0.10	0.10	$\rightarrow$	0.10	0.01	$\rightarrow$	**0.289**	0.10	$\rightarrow$	0.10
π_c	Beta	μ	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50
		σ	0.10	0.10	$\rightarrow$	0.10	0.10	$\rightarrow$	0.10	0.01	$\rightarrow$	**0.289**	0.10	$\rightarrow$	0.10
π_d	Beta	μ	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50	0.50	$\rightarrow$	0.50
		σ	0.10	0.10	$\rightarrow$	0.10	0.10	$\rightarrow$	0.10	0.01	$\rightarrow$	**0.289**	0.10	$\rightarrow$	0.10
ϕ_{a0}	Beta	μ	0.01	0.50	$\rightarrow$	**0.01**	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.15	$\rightarrow$	0.15
		σ	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.0005	$\rightarrow$	**0.0638**
ϕ_{a1}	Beta	μ	0.99	0.50	$\rightarrow$	**0.99**	0.99	$\rightarrow$	0.99	0.99	$\rightarrow$	0.99	0.85	$\rightarrow$	0.85
		σ	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.0005	$\rightarrow$	**0.0638**
ϕ_{b0}	Beta	μ	0.01	0.50	$\rightarrow$	**0.01**	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.15	$\rightarrow$	0.15
		σ	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.0005	$\rightarrow$	**0.0638**
ϕ_{b1}	Beta	μ	0.99	0.50	$\rightarrow$	**0.99**	0.99	$\rightarrow$	0.99	0.99	$\rightarrow$	0.99	0.85	$\rightarrow$	0.85
		σ	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.0005	$\rightarrow$	**0.0638**
ϕ_{c0}	Beta	μ	0.99	0.50	$\rightarrow$	**0.99**	0.99	$\rightarrow$	0.99	0.99	$\rightarrow$	0.99	0.85	$\rightarrow$	0.85
		σ	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.0005	$\rightarrow$	**0.0638**
ϕ_{c1}	Beta	μ	0.01	0.50	$\rightarrow$	**0.01**	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.15	$\rightarrow$	0.15
		σ	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.0005	$\rightarrow$	**0.0638**
ϕ_{d0}	Beta	μ	0.99	0.50	$\rightarrow$	**0.99**	0.99	$\rightarrow$	0.99	0.99	$\rightarrow$	0.99	0.85	$\rightarrow$	0.85
		σ	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.0005	$\rightarrow$	**0.0638**
ϕ_{d1}	Beta	μ	0.01	0.50	$\rightarrow$	**0.01**	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.15	$\rightarrow$	0.15
		σ	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.01	$\rightarrow$	0.01	0.0005	$\rightarrow$	**0.0638**

Table 11: Simulation parameters. Each column details parameters used to generate prior distributions for one of the simulations below. The prior distribution for the full parameter vector is formed from independent draws from Beta distributions for all probabilities and the Dirichlet distribution for shares. Note that the mean and standard deviation parameterization we provide for Beta distributions can be mapped directly to the more standard α, β parameterization. For the Clue Uncertainty experiments the parameters correspond to a shift from $Beta(.15 \times 2^i, .85 \times 2^i)$ for $i \in \{0, 4, 8\}$.

Experiment	j steps per exp.	k sims per step	Comments
1: Varying N or m	29	5,200	The 5,200 k simulations for each θ_j were split into 26 runs of 200 k sims, and then compiled through averaging. Datapoints at N=3 and N=4 were added using 14,200 k simulations.
2: Probative Value	30	5,200	The 5,200 k simulations for each θ_j were split into 26 batches of 200 k sims, and then compiled through averaging.
3: Effect Heterogeneity	30	5,200	The 5,200 k simulations for each θ_j were split into 26 batches of 200 k sims, and then compiled through averaging.
4: Assignment Uncertainty	30	5,200	The 5,200 k simulations for each θ_j were split into 26 batches of 200 k sims, and then compiled through averaging.
5: Clue Uncertainty	30	10,200	The 10,200 k simulations for each θ_j were split into 26 batches of 200 k sims and 10 batches of 500, then compiled through averaging.

Table 12: *Note:* Each experiment takes j steps through different values of θ. At each θ_j, the data is simulated k times. For each simulation, a call is made to the Stan model and HMC (Hamiltonian Monte Carlo) sampling is used to approximate the posterior distribution. In each such call to Stan, we run 4 chains with 6000 iterations, and 1000 warmup draws.

I Learning from non-discriminating clues

We noted in the text that when priors over clue probabilities do not discriminate between causal types, then learning clue values does not affect learning over other parameters when $n = 1$. However, learning is possible for $n > 1$ even when priors over clue probabilities do not discriminate between causal types.

To see why, divide θ into two parts, $\theta_{-\phi}$ and θ_{ϕ}, where $\theta_{-\phi}$ denotes the vector of parameters excluding $\{\phi_{jx}\}_{j\in\{a,b,c,d\},x\in\{0,1\}}$, and θ_{ϕ} the complement.

We represent non-discriminating priors over the clues probabilities as follows. Assume that the prior distribution over θ is given by $p(\theta_{-\phi})\prod_{j\in\{a,b,c,d\},x\in\{0,1\}} f(\phi_{jx})$ — thus the marginal prior distribution over each ϕ_{jx}, $j \in \{a, b, c, d\}, x \in \{0, 1\}$ is given identically by some distribution f. Thus for this claim the key feature is not that f is flat, but simply that it is the same for different causal types.

We consider a situation in which we observe $X = Y = K = 1$ for a case and show that the posterior distribution over $\theta_{-\phi}$ is the same as it would be if we observed X, Y, data only. The same analysis can be conducted for any other combination of X, Y data.

With $X = Y = K = 1$ the posterior marginal distribution is:

$$\begin{aligned} p(\theta_{-\phi}|X=Y=K=1) &= \frac{\int\int\left(\left(\lambda_b\pi_b\phi_{b1}+\lambda_d\pi_d\phi_{d1}\right)p(\theta_{-\phi})f(\phi_{b1})f(\phi_{d1})\right)d\phi_{b1}d\phi_{d1}}{\int\int\int\left(\lambda_b\pi_b\phi_{b1}+\lambda_d\pi_d\phi_{d1}\right)p(\theta_{-\phi})\,f(\phi_{b1})f(\phi_{d1})d\theta_{-\phi}d\phi_{b1}d\phi_{d1}} \\ &= \frac{\left(\lambda_b\pi_b\int\phi_{b1}f(\phi_{b1})d\phi_{b1}+\lambda_d\pi_d\int\phi_{d1}f(\phi_{d1})d\phi_{d1}\right)p(\theta_{-\phi})}{\int\left(\lambda_b\pi_b\int\phi_{b1}f(\phi_{b1})d\phi_{b1}+\lambda_d\pi_d\int\phi_{d1}f(\phi_{d1})d\phi_{d1}\right)p(\theta_{-\phi})d\theta_{-\phi}} \end{aligned}$$

Since $\int \phi_{b1}f(\phi_{b1})d\phi_{b1} = \int \phi_{d1}f(\phi_{d1})d\phi_{d1}$, this simplifies to the posterior that obtains when information on K is disregarded entirely:

$$\frac{\left(\lambda_b\pi_b+\lambda_d\pi_d\right)p(\theta_{-\phi})}{\int\left(\lambda_b\pi_b+\lambda_d\pi_d\right)p(\theta_{-\phi})d\theta_{-\phi}}$$

A critical step in this simple proof is our ability to move the integrals given the fact that the ϕ terms enter the likelihood in an additive way.

This is not the case for $n > 1$. For example say there were two cases, each with $X = Y = K = 1$ we would then have:

$$p(\theta_{-\phi}|X = Y = K = (1,1)) = \frac{\int\int \left((\lambda_b\pi_b\phi_{b1} + \lambda_d\pi_d\phi_{d1})^2 p(\theta_{-\phi}) f(\phi_{b1}) f(\phi_{d1})\right) d\phi_{b1} d\phi_{d1}}{\int\int\int (\lambda_b\pi_b\phi_{b1} + \lambda_d\pi_d\phi_{d1})^2 p(\theta_{-\phi})\, f(\phi_{b1}) f(\phi_{d1}) d\theta_{-\phi} d\phi_{b1} d\phi_{d1}}$$

which does not admit the same simplification. For a counterexample, suppose that the only parameters over which there is uncertainty are λ_b, ϕ_{b1} and ϕ_{d1}. Assume that $f()$ is uniform, that priors over λ_b are also given by a uniform distribution over $[0, 1]$, that $\lambda_d = 1 - \lambda_b$, and that $\pi_b = \pi_d$. Note that in this simple world, observing two instances of $X = Y = 1$ does not provide information on whether b types are more or less common than d types since conditional on $X = 1$ both types produce $Y = 1$ (equivalently, both types produce data like this with probability π_b, π_d). The question is whether information on K can shift beliefs even though there is no prior information to lead one to expect K to be observed with greater probability for a b or a d type. The posterior marginal distribution over λ_b is now:

$$p(\lambda_b|X = Y = K = (1,1)) = \frac{\int\int (\lambda_b\phi_{b1} + (1-\lambda_b)\phi_{d1})^2 d\phi_{b1} d\phi_{d1}}{\int\int\int (\lambda_b\phi_{b1} + (1-\lambda_b)\phi_{d1})^2 d\lambda_b d\phi_{b1} d\phi_{d1}}$$

Solving out yields:

$$p(\lambda_b|X = Y = K = (1,1)) = \frac{6(2 + \lambda_b^2 - \lambda^b)}{11}$$

This symmetric U-shaped posterior distribution suggests that observing K does not shift the expected share of b and d types in this situation. It does, however, result in greater weight placed on *extreme* values of λ_b — that is, after seeing the data, we now believe that it is more likely that there are *either* very many

or very few b types. Intuitively, the discrimination arises because, given the independent priors on ϕ_b and ϕ_d, it is more likely that clue probabilities are high for one type than that they are high for two types.

J Maximum Likelihood Integration

Although we favor a Bayesian approach to integrating inferences from qualitative and quantitative data, we note that a similar approach can be implemented within a maximum likelihood framework. The key point is that information on the probative value on clues places a structure on the likelihood which affects inferences under both a Bayesian and a maximum likelihood analysis.

We illustrate here with an example that imposes considerable structure on the likelihood.

Suppose that it is known for certain that there are only b and d types in a population and that a clue K is observed with probability q if the unit is a b type in treatment, and with probability zero otherwise (we use q rather than ϕ to highlight the fact that in this example probative value is given and is not a parameter to be estimated). Say, moreover, that all types are assigned to treatment with probability .5.

In that case data can be summarized by a vector $(n_{00}, n_{01}, n_{110}, n_{111})$ and the likelihood of the data is given by:

$$L = \frac{n!}{n_{00}!n_{01}!n_{110}!n_{111}!}(.5\lambda_b)^{n_{00}}(.5(1-\lambda_b))^{n_{01}}(.5\lambda_b(1-q)+.5(1-\lambda_b))^{n_{110}}(.5\lambda_b q)^{n_{111}}$$

and so:

$$L \propto \lambda_b^{n_{00}+n_{111}}(1-\lambda_b)^{n_{01}}(1-q\lambda_b)^{n_{110}}$$

The maximum likelihood estimate (MLE) is the maximum of the log likelihood:

$$\max\left((n_{00+}n_{111})\ln(\lambda_b) + n_{01}\ln(1-\lambda_b) + n_{110}\ln(1-q\lambda_b)\right)$$

First-order conditions are:

$$\frac{n_{00}+n_{111}}{\lambda_b}-\frac{n_{01}}{1-\lambda_b}-\frac{qn_{110}}{1-q\lambda_b}=0$$

The second-order condition is satisfied since:

$$-\frac{n_{00}+n_{111}}{\lambda_b^2}-\frac{n_{01}}{(1-\lambda_b)^2}-q^2\frac{n_{110}}{(1-q\lambda_b)^2}<0$$

We note that if there is only one case or one sort of case (e.g., only $X=Y=K=1$ cases), then the first-order condition cannot be satisfied: the maximum is at the boundary. An implication of this is that, for single-case analysis, the MLE estimate can be insensitive to the probative value of clues.

For illustrative purposes, suppose $n_{01}=n_{00}=n_{110}=n_{110}>0$. Then the first order condition is satisfied uniquely[23] by:

$$\lambda_b^*=\frac{3(q+1)-\sqrt{9(q+1)^2-32q}}{8q}$$

This solution falls from $\lambda_b=\frac{2}{3}$ to $\lambda_b=\frac{1}{2}$ as q varies from 0 to 1. These differences reflect the fact that in this example the n_{111} $X=Y=K=1$ cases and the n_{00} $X=Y=0$ cases are known to be b types; and the n_{01} cases with $X=0,Y=1$ are known to be d types. The only uncertainty arises for the n_{110} cases with $X=Y=1$ and $K=0$. If b types have a low probability of exhibiting the clue when $X=1$, then many of these cases are likely to be b types.

[23]A second solution exists but exceeds 1 for admissible values of q.