# What do the numbers really say: A Structural and Simulation-Based Reassessment of False-Confession Wrongful-Conviction Rates

***Daniel Malmgren*[1], *Laurence Alison*[2], *Emily Alison*[2] *and Frances Surmon-Böhr*[2]**

[1]*Department of Psychology, Lund University*
[2]*Faculty of Healthy and Life Sciences, University of Liverpool*

**Corresponding author:**
francessurmonbohr@gmail.com

## ABSTRACT

**Abstract**

False-confession wrongful convictions (FCWCs) are among the most serious outcomes associated with police interrogation practices. Quantifying their prevalence is therefore important for legal policy, investigative training, and public debate. Recent work by Mourtgos and Adams (2026) argues that FCWC prevalence "clusters below one percent." We examine the probabilistic structure underlying this claim and argue that the published estimate is best understood as a conviction-conditioned quantity rather than a direct estimate of interrogation-conditioned probability.

Our analysis identifies two related issues. First, different probability factorizations can yield different substantive interpretations even when the calculations are internally coherent. Second, evaluating interrogation practices requires process-aligned quantities that reflect how techniques operate during interrogations, including their effects on information quality, error generation, and downstream outcomes. We further argue that technique-specific interrogation-conditioned probabilities cannot be recovered from conviction-conditioned quantities without explicitly modeling the interrogation-to-conviction sequence.

We do not aim to estimate the true prevalence of FCWCs. Instead, we use fixed-value examples, Monte Carlo simulations, and a generative interrogation-to-conviction model to clarify how alternative estimands behave under explicit assumptions. We conclude that the key question is not whether a pooled conviction-conditioned quantity appears numerically small, but whether such quantities obscure meaningful differences across interrogation contexts when used to inform interrogation policy.

***Key Words**: false confessions; wrongful convictions; investigative interviewing; Bayesian modeling; probability theory*

## Introduction

Recent work by Mourtgos and Adams (2026) makes the provocative claim that lawful police interrogation tactics produce false-confession wrongful convictions (FCWCs) at a rate that “clusters below one percent.” Their analysis raises an important methodological problem: how to reason about FCWC risk in the absence of complete population-level data. However, we argue that the published estimate is best understood as a conviction-conditioned quantity rather than a direct estimate of interrogation-conditioned risk. However, the controversy extends beyond academic debate, as the published estimate has been presented in professional policing forums and may be interpreted as evidence that existing interrogation practices pose negligible risk, thereby framing reform as unnecessary or disproportionate. Our central argument is that the calculation used by Mourtgos and Adams relies on a reference class defined primarily on convictions rather than interrogations, making it unsuitable for direct interpretation as the probability that an interrogation produces an FCWC.

An earlier preprint version of the model contained an additional algebraic redundancy that was later corrected prior to publication. We mention this only because both versions circulated in professional and academic contexts. The present analysis focuses primarily on the published formulation and its implications for estimating an interrogation-conditioned probability. Specifically, we argue that the published model estimates a conviction-linked prevalence quantity rather than the probability that an interrogation results in a false-confession wrongful conviction. Treating the former as a proxy for the latter can lead to misleading interpretations.

The case for reforming interrogation practices rests on multiple considerations, including ethical and legal concerns, as well as a growing body of evidence that science-based methods can improve information gathering and mitigate cognitive biases. The argument for changing practice is therefore not reducible to the prevalence of FCWCs alone. Debates about interrogation reform also concern whether particular techniques disproportionately increase the likelihood of unreliable information or false confessions relative to available alternatives.

However, the quantitative claim made by Mourtgos and Adams warrants direct scrutiny because it may be interpreted as implying that the risk of false confessions is negligible and that reform is therefore unnecessary. If a central argument against reform is that the probability of a FCWC is empirically small, then both the empirical uncertainty surrounding wrongful convictions and the alignment between the probability model and the interrogation process become critical.

Estimating the true rate of wrongful conviction is notoriously difficult because most errors remain a “dark figure” that never becomes observable. Nonetheless, methodologically rigorous attempts suggest that error rates are not negligible in high-stakes contexts. For example, Gross et al. (2014) leveraged the unusual level of post-conviction scrutiny in U.S. death-sentence cases and used survival-analysis logic to estimate that at least 4.1% of death-sentenced defendants are innocent. This kind of benchmark does not settle the empirical prevalence of false-confession wrongful convictions, nor does it generalize mechanically beyond capital cases. It does, however, illustrate why probabilistic claims about "negligible risk" require careful alignment between denominators, conditioning events, and the underlying interrogation-to-conviction process. When the underlying quantity is difficult to observe directly, the interpretation of the estimand itself becomes especially important.

Against this background, our aim is not to estimate the real-world interrogation-conditioned prevalence of FCWCs, but to evaluate whether the published factorization can support claims about interrogation-conditioned risk. Mourtgos and Adams (2026) proposed a Bayesian framework to estimate a "base-rate" of FCWCs. The present paper examines whether the quantity estimated by such a framework corresponds accurately to the prospective, interrogation-conditioned probability of FCWCs.

In the final published version of their article, Mourtgos and Adams define the FCWC quantity as:

$$\boldsymbol{P(FCWC) = P(Wrongful\ conviction) \times P(False\ confession \mid Wrongful\ Conviction)}$$

This formulation yields a conviction-conditioned quantity: the proportion of convictions involving false confessions. Importantly, this quantity is not a direct empirical observation but the output of a hierarchical model that pools heterogeneous studies using explicit assumptions about recency, sample size, and evidentiary type, including a linearly scaled recency weight that can materially affect the pooled distribution.

The present analysis focuses on whether the formulation used by Mourtgos and Adams corresponds to the interrogation-conditioned probability that interrogation practices produce false-confession wrongful convictions, rather than to the proportion of convictions involving false confessions.

Our argument is not that the published factorization is mathematically incoherent, but rather that it estimates a different quantity than the interrogation-conditioned probability implied in some interpretations of the "below 1%" claim. This distinction is important because conviction-conditioned quantities and interrogation-conditioned probabilities answer different substantive questions and therefore have different implications for policy discussions surrounding interrogation practices.

In the sections that follow, we examine how the factorization used by Mourtgos and Adams differs from an interrogation-aligned decomposition and how those differences affect what the resulting estimates mean. We first analyze the structure of their factorization and distinguish between two analytically separate issues: (a) reference-class mismatch and (b) conditioning on quantities that may not correspond to confession-based cases. We then demonstrate the quantitative implications with three complementary demonstrations: fixed-value numerical examples, a Monte Carlo sensitivity exercise that explores a wide but plausible range of component probabilities (without treating any single set of values as authoritative), and a generative interrogation-to-conviction pipeline in which the true interrogation-conditioned probability is known by construction.

We then extend the simulation framework by introducing technique heterogeneity to illustrate how pooling across interrogation contexts can attenuate apparent technique-specific differences and obscure meaningful distinctions between interrogation approaches. We further distinguish between varying the prevalence of tactic exposure across interrogations and varying the conditional probability that tactics contribute to false confessions or wrongful convictions, as these represent analytically distinct forms of sensitivity analysis. These demonstrations are diagnostic rather than inferential: they are intended to clarify how alternative estimands behave under explicit assumptions, rather than to estimate the empirical prevalence of FCWCs in real-world settings.

## Methodology

**Defining an Interrogation-Conditioned Estimand**

In the published version, Mourtgos and Adams explicitly motivate their approach as a correction to 'visible-tragedy' reasoning: they argue that risk cannot be estimated from FCWC cases alone and must incorporate the far larger set of interrogations that do not produce FCWCs. We agree with this general principle. Indeed, any meaningful estimate of an interrogation-conditioned quantity must be defined on the full space of interrogations, rather than solely on observed errors. The central issue, however, is whether the published formulation, instead estimates a conviction-conditioned quantity that differs from the interrogation-conditioned quantity discussed in many policy interpretations.

As noted in the introduction, the policy-relevant target in the present analysis is the probability that an interrogation produces a false-confession wrongful conviction. Using standard notation where:

- I = The case involves an interrogation
- C = A confession is obtained
- W = The conviction is wrongful
- F = The confession is false

The event of interest is the joint probability: $P(F \cap W \cap C \mid I)$. Throughout this paper, we distinguish between interrogation-conditioned probabilities, defined on the population of interrogations, and conviction-conditioned quantities, defined on the subset of cases resulting in conviction. By the chain rule of probability, this event can be decomposed as:

$$P(F \cap W \cap C \mid I) = P(C|I) \times P(F|C,I) \times P(W|F,C,I)$$

The interrogation-conditioned FCWC probability is therefore:

$$P(FCWC|I) = P(C|I) \times P(F|C,I) \times P(W|F,C,I).$$

This decomposition aligns each component with the underlying interrogation-to-conviction process:

**Interrogation -> Confession -> False Confession -> Wrongful Conviction**

Conceptually, the falsity of a confession precedes the wrongful conviction because a confession is false at the moment it is made. However, in empirical practice, false confessions are often identified retrospectively through exoneration or post-conviction review. The present analysis adopts the causal ordering above while recognizing that many FCWC datasets are constructed retrospectively from conviction outcomes.

**Comparison with the Authors' Formulation**

In the published version of their article, Mourtgos and Adams (2026) define the FCWC base rate as: $P(FCWC) = P(W) \times\ P(F \mid W)$. This expression is a valid probability factorization if $P(W)$, and $P(F|W)$ are defined on the same reference class. The key question, however, is what reference class the resulting quantity represents.

As used in the published article, the formulation is best understood as estimating the share of wrongful-conviction cases, or convictions more broadly, that involve false confessions. It therefore corresponds to a conviction-conditioned or conviction-linked quantity rather than the interrogation-conditioned probability defined above. In other words, it does not directly estimate: $P(FCWC|I)$

This distinction matters because interrogation-conditioned probabilities and conviction-conditioned quantities answer different questions. The former ask how often an interrogation process produces an FCWC; the latter ask what proportion of convictions, or wrongful convictions, involve false confessions. Both are meaningful quantities, but they are not interchangeable.

To translate a conviction-conditioned quantity into an interrogation-conditioned probability, one must explicitly model the pathway from interrogation to conviction. At minimum, this requires incorporating the probability that an interrogation results in conviction, $P(V = 1|I)$, where V = 1 denotes that a conviction occurs, and ensuring that all conditioning events remain aligned with the same population of cases.

We therefore frame the issue not as a mathematical error in Bayes' theorem or in the multiplication rule, but as an estimand-alignment problem. The published formulation may be informative about the composition or prevalence structure of conviction outcomes, but it cannot by itself answer the prospective policy question of how often interrogation practices produce false-confession wrongful convictions.

An earlier preprint version of the article included an additional confession-prevalence term, $P(C)$, in the base-rate expression. That term was removed prior to publication and is therefore not central to the present critique. We note the earlier formulation only because both versions of the work have circulated in professional and academic discussions and because examining the evolution of the model helps clarify the need for the distinction between conviction-conditioned quantities and an interrogation-conditioned probability. The present analysis, however, focuses on the published formulation and the estimand it represents.

**Mourtgos and Adams' Modeling Assumptions**

The resulting estimates in both the preprint and the published article depend substantially on modeling and weighting assumptions embedded in the hierarchical synthesis procedure. Although framed as a pooled estimate derived from the literature, the FCWC quantity modeled by Mourtgos and Adams is the output of a hierarchical Beta model with strong structural assumptions. In particular, individual studies are weighted multiplicatively by sample size, evidence type, and a linearly scaled recency factor, such that more recent studies contribute substantially greater influence on the pooled estimate than older studies. This procedure gives comparatively greater weight to recent data points, even when studies may differ in case-selection procedures, detection mechanisms, jurisdictional context, or interrogation practices.

As a result, the modeled FCWC quantity reflects not only empirical uncertainty in the underlying literature but also analytic assumptions regarding how different forms of evidence should be weighted and aggregated. These assumptions materially affect the interpretation of the resulting quantity and therefore warrant explicit scrutiny.

## Analysis

**Downstream Consequences: Estimand Misalignment in the 'Inverse-Probability'**

The consequence of the Mourtgos and Adams' conviction-conditioned formulation extends to the subsequent inverse-probability analysis used to estimate interrogation-technique-specific quantities. The Bayesian updating procedure itself is mathematically valid. However, the interpretation of the resulting posterior distribution depends critically on whether the prior quantity corresponds to the estimand of interest.

In the published framework, the FCWC base-rate functions as a prior input for estimating technique-conditional probabilities, such as: $P(FCWC|T)$, where T denotes the use of a particular interrogation tactic or technique. The concern raised here is not that Bayes' theorem is incorrectly applied, but that the prior quantity being propagated through the model is defined on a conviction-conditioned reference class rather than on the population of interrogations. As a result, the posterior distribution inherits the assumptions and limitations of that underlying estimand. Consequently, posterior inferences about technique-specific interrogation-conditioned probabilities remain dependent on the alignment between the prior quantity and the interrogation process being modeled.

Furthermore, this dependence on modeling assumptions is amplified by the nature of the additional inputs. As the authors explicitly acknowledge, direct empirical estimates of interrogation-technique sensitivity and specificity are "limited and contested." Accordingly, the supplementary analyses define broad plausibility intervals and sample sensitivity and specificity values across those ranges rather than estimating them directly from population-representative interrogation data.

The resulting posterior distributions should therefore be interpreted as conditional projections under a particular set of structural and probabilistic assumptions, rather than as direct empirical estimates of interrogation-conditioned FCWC quantities. More specifically, the model characterizes the implications of assumed relationships between interrogation tactics and FCWC outcomes within the chosen reference class structure, not the empirical measurement of interrogation-conditioned FCWC probability across the population of interrogations.

**Distinguishing Tactic Prevalence from Tactic Risk**

In their supplementary robustness analyses, Mourtgos and Adams further vary the assumed prevalence of potentially problematic interrogation tactics across interrogations. Importantly, varying the prevalence of tactic exposure is analytically distinct from varying the probability that a tactic contributes to false confessions or wrongful convictions.

Formally, the Mourtgos and Adams' robustness analyses primarily vary:

$P(T)$

The proportion of interrogations in which a tactic is used. However, this differs from modeling quantities such as:

$$P(F|T,C,I) \text{ or } P(W|F,T,C,I)$$

which represent the conditional probability that a tactic contributes to a false confession or wrongful conviction within the interrogation process itself.

This distinction is important because a tactic may be relatively uncommon while still carrying a disproportionately elevated conditional probability of contributing to FCWCs. Sensitivity analyses that only vary tactic prevalence therefore evaluate how pooled population-level quantities change as exposure frequency changes, but they do not directly identify how strongly a tactic contributes to FCWC probability conditional on its use.

**Distinguishing Reference-Class and Conditioning Issues in the Published Factorization**

We identify two analytically distinct issues in the Mourtgos and Adams formulation. The first concerns the reference class on which the FCWC quantity is defined. The second concerns the conditioning structure used to relate confessions, falsity, and wrongful conviction. These issues are related, but they have different implications and should not be treated as a single source of bias.

***Issue 1: Reference-Class Mismatch***

The primary concern with the published formulation is that it estimates a quantity defined on a conviction-conditioned reference class, while policy and training questions are typically posed at the level of interrogations. The authors define:

$$P(FCWC) = P(W) \times P(F|W).$$

This expression can be interpreted coherently if $P(W)$ and $P(F|W)$ are defined on a common reference class. For example, if $P(W)$ denotes the proportion of convictions that are wrongful, and $P(F|W)$ denotes the proportion of wrongful convictions involving a false confession, then their product estimates the proportion of convictions that are false-confession wrongful convictions. This is a meaningful quantity. However, it is not equivalent to the interrogation-conditioned probability: $P(FCWC|I)$.

The difference is not merely terminological. A conviction-conditioned or conviction-linked quantity answers a retrospective system-level question: among cases that culminate in conviction, what proportion involve false-confession wrongful convictions?

By contrast, an interrogation-conditioned quantity answers a prospective process question: how often does an interrogation process generate a false-confession wrongful conviction?

To translate a conviction-conditioned estimate into an interrogation-conditioned probability, the pathway from interrogation to conviction must be explicitly modeled. At minimum, this requires incorporating the probability that an interrogation results in conviction: $P(V = 1|I)$, as well as ensuring that the conditioning events are consistently defined over the same population of cases. Without this mapping, a conviction-conditioned estimate cannot be interpreted as the probability that an interrogation produces an FCWC.

This is the reference-class issue at the centre of our critique. It does not imply that the published quantity is meaningless or mathematically incoherent. Rather, it implies that the quantity estimated by the published formulation differs from the interrogation-conditioned probability often invoked in discussions of interrogation policy, prevention, and training.

An earlier preprint version included an additional confession-prevalence term, $P(C)$, that was removed prior to publication. The present analysis therefore focuses primarily on the published formulation. As a probability bounded between 0 and 1, the additional $P(C)$ term necessarily reduced the magnitude of the resulting estimate.

***Issue 2: Conditioning on Overall Rather than Confession-Linked Error Rates***

A second and analytically distinct issue concerns the conditioning structure used in the published formulation, Specifically, the model uses an overall wrongful conviction quantity, $P(W)$, rather than a quantity conditioned on the presence of a confession.

False-confession wrongful convictions can arise only in cases where a confession is obtained and contributes to the conviction outcome. From the perspective of the interrogation-to-conviction process, the most directly relevant error quantity is therefore not the overall wrongful-conviction prevalence, but the prevalence of wrongful conviction among confession-based cases: $P(W|C, I)$

This matters because confession-based cases may differ systematically from the broader population of convictions. In the false-confession literature, wrongful convictions involving false confessions are often concentrated in cases where confessions play a substantial evidentiary role (Kassin et al., 2012). Under such conditions, it is plausible that:

$$P(W|C) > P(W).$$

However, this relationship should not be treated as a mathematical necessity. The magnitude and even direction of the difference depends on how confession-based cases differ from the broader population of convictions, including the strength of corroborating evidence and the role confessions play within particular investigative contexts.

The key point is therefore not that the published formulation necessarily underestimates FCWC probability in every setting, but that the use of an overall wrongful-conviction quantity may not correspond to the subset of cases relevant to false-confession wrongful convictions. If confession-based cases differ systematically from the broader conviction population, then replacing $P(W|C, I)$ with $P(W)$ may materially alter the resulting estimate.

Accordingly, the issue is best understood as a conditioning-alignment problem. The relevant error structure for evaluating FCWC probability depends on the population of cases under consideration and on how confessions function within the interrogation-to-conviction sequence.

### Illustrative Consequences of Alternative Conditioning Structures

If the goal is to estimate the probability that an interrogation produces a false-confession wrongful conviction, the estimand must be defined on the population of interrogations. As noted above, one interrogation-aligned decomposition is:

$$P(F \cap W \cap C \mid I) = P(C \mid I)\, P(F \mid C, I)\, P(W \mid F, C, I).$$

This decomposition aligns the conditioning structure with the interrogation-to-conviction sequence:

**Interrogation → Confession → False Confession → Wrongful Conviction**

The purpose of the present section is not to estimate the true empirical prevalence of FCWCs, but to illustrate how alternative conditioning structures can produce materially different quantities even when based on similar conceptual inputs.

To demonstrate this point, we present a fixed-value numerical example using plausible parameter values drawn from the literature. These values are illustrative rather than authoritative and are intended solely to clarify how different formulations behave under explicit assumptions.

### Conviction-Conditioned Estimand

First, consider the published formulation used by Mourtgos and Adams:

$$P(FCWC) = P(W) \times P(F \mid W).$$

Using representative values in the range commonly discussed in the literature (Gross et al., 2014; Garrett, 2011), for example an overall wrongful-conviction rate of 3% and a false-confession share among wrongful convictions of 15%, the formulation yields:

$P(FCWC \mid Conviction) = 0.03 \times 0.15 = 0.0045$ or 0.45%, corresponding to roughly one FCWC for every 222 convictions. By construction, this quantity is conviction conditioned. It therefore describes the proportion of convictions involving false confessions rather than the probability that an interrogation produces an FCWC.

### Interrogation-Conditioned Estimand

Now, consider an interrogation-aligned decomposition using plausible parameter values consistent with the literature:

- $P(C \mid I) = 0.60$: 60% of interrogations produce a confession
- $P(F \mid C, I) = 0.05$: 5% of confessions obtained during interrogation are false. We use a higher value than the overall prevalence implied by the conviction-conditioned formulation because false confessions are concentrated within the subset of confession-based cases.
- $P(W \mid F, C, I) = 0.30$: 30% of false confessions contribute to wrongful conviction outcomes. This value is also higher than the overall average of 15%, as we are now conditioning on the specific subset of cases involving false confessions.

This yields:

$P(FCWC \mid I) = 0.60 \times 0.05 \times 0.30 = 0.009$ or 0.9%, corresponding to about one FCWC per 111 interrogations.

The important point is not that one estimate is “correct” and the other "incorrect." Rather, the two formulations answer different questions because they are defined on different conditioning structures and reference classes. Importantly, these examples are diagnostic rather than inferential. They are intended to clarify the implications of alternative estimands and conditioning assumptions, not to provide a definitive estimate of real-world FCWC prevalence.

These numerical examples are illustrative rather than inferential. The examples demonstrate how alternative factorizations applied to identical component values yield substantively different conclusions solely due to algebraic structure.

**Monte Carlo Illustration of Reference-Class Mappings**

The purpose of the simulations is not to estimate the true prevalence of FCWCs, but to examine how conviction-conditioned quantities behave when interpreted at the level of the interrogation process. In particular, we assess whether estimands defined on the space of convictions can be interpreted as interrogation-conditioned quantities without an explicit transformation.

To this end, we conducted a Monte Carlo simulation implemented in R (R Core Team, 2024). Figures were generated using the ggplot2 (Wickham, 2016). The simulation fixes the conviction-conditioned FCWC quantity while varying the probability of conviction given interrogation, $P(V = 1|I)$ across a plausible range of values. Specifically, we defined:

$$P(FCWC|V = 1) = P(W|V = 1) \times P(F|W)$$

Using illustrative values within the range discussed in the literature and examined how the implied interrogation-conditioned quantity changes as conviction probability varies.

We then compared three quantities:

**(1) The conviction-conditioned estimand**

$$P(FCWC \mid V = 1) = P(W \mid V = 1) \times P(F \mid W)$$

**(2) Direct reinterpretation as an interrogation-conditioned quantity**

Treating $P(FCWC \mid V = 1)$ as if it directly represented $P(FCWC \mid I)$

**(3) An explicitly mapped interrogation-conditioned quantity**

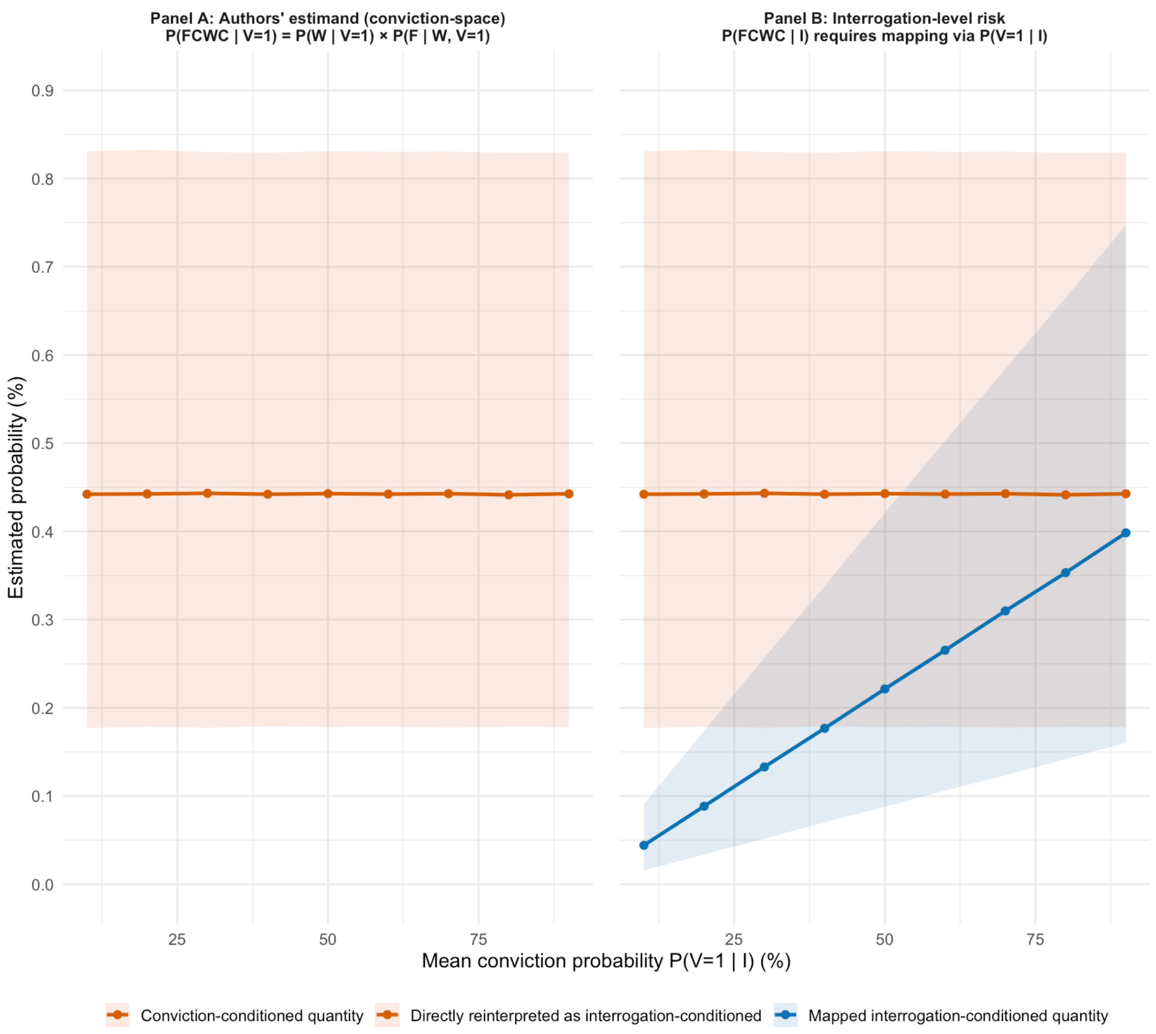
$$P(FCWC \mid I) = P(V = 1 \mid I) \times P(FCWC \mid V = 1)$$

Figure 1 illustrates that the authors' published estimand remains invariant to the conviction probability by construction, because it is defined conditional on conviction rather than interrogation. Panel B demonstrates the consequence of interpreting this quantity as an interrogation-conditioned probability without explicitly incorporating the probability of conviction. When the conviction-conditioned quantity is treated as if it directly represented an interrogation-conditioned probability, the resulting interpretation differs from the explicitly mapped interrogation-conditioned quantity. The magnitude and direction of this difference depend on the conviction probability $P(V = 1|I)$.

The important point is therefore not that the conviction-conditioned quantity is mathematically invalid, but that it answers a different question than the interrogation-conditioned quantity relevant to prevention, training, and policy evaluation.

Crucially, the differences shown in Figure 1 arise from conditioning structure and reference class, not from uncertainty in component probabilities. No choice of priors or sensitivity analysis can resolve this mismatch without redefining the estimand itself.

**Figure 1.** ***Conviction-conditioned and interrogation-conditioned quantities are not interchangeable***




*Note.* Panel A displays the conviction-space estimand $P(FCWC \mid V = 1) = P(W \mid V = 1) \times P(F \mid W)$, representing the proportion of convictions involving a false-confession wrongful conviction. Because this quantity is defined conditional on conviction, it is invariant to changes in the overall conviction probability $P(V = 1 \mid I)$. Panel B displays the interrogation-conditioned quantity: $P(FCWC \mid I) = P(V = 1 \mid I) \times P(FCWC \mid V = 1)$, obtained by explicitly incorporating the probability of conviction. Treating the conviction-conditioned quantity as if it were an interrogation-conditioned probability implicitly assumes $P(V = 1 \mid I) = 1$.

The figure illustrates how different reference classes and conditioning structures produce different implied interpretations even when the underlying component probabilities remain internally consistent.

All curves are based on Monte Carlo simulation across a plausible range of conviction probabilities. The differences across panels arise from the fact that conviction-conditioned and interrogation-conditioned quantities are defined on different conditioning structures and reference classes.

Figure 1 does not reflect disagreement about the proportion of convictions involving FCWCs. Rather, it demonstrates that conviction-conditioned and interrogation-conditioned quantities answer different questions and therefore are not directly interchangeable without explicit mapping through the probability of conviction.

From the perspective of evaluating interrogation practices prospectively, interrogation-conditioned quantities may be more directly aligned with questions concerning prevention, training, and investigative decision-making because they explicitly incorporate the pathway from interrogation to conviction. The simulations demonstrate that failure to distinguish between these quantities can lead to materially different interpretations even when all component probabilities are internally consistent.

**Generative model: Interrogation-Conditioned and Conviction-Conditioned Quantities**

$P(FCWC \mid I)$ denotes the interrogation-conditioned probability that a given interrogation will result in an FCWC. This quantity is defined on the population of interrogations and is closely aligned with prospective questions concerning interrogation practices, prevention, training, and investigative decision-making.

$P(FCWC \mid V = 1)$ denotes the conviction-conditioned proportion of convictions that involve FCWCs. This quantity is defined on the population of convictions and reflects a retrospective system-level perspective.

$P(FCWC \mid T)$ denotes a technique-conditional quantity, representing the probability of an FCWC conditional on the use of a particular interrogation tactic or approach.

Because interrogation tactics operate during the interrogation process, estimates of technique-conditional quantities require explicit specification of the reference class and conditioning structure used to define the quantity of interest.

To demonstrate the quantitative impact of this reference-class mismatch in a controlled environment, we constructed a generative model of an interrogation-to-conviction pipeline. Unlike robustness analyses that vary priors or study weights within a fixed estimand, the simulations presented here vary the conditioning structure and reference class themselves, allowing us to examine how different formulations map onto the interrogation-to-conviction process. The parameter values used in this simulation are illustrative rather than empirical. The goal is not to estimate the real-world prevalence of FCWCs, but to create a setting in which the underlying data-generating process is known by construction. This allows us to evaluate how alternative estimands behave relative to a known interrogation-conditioned quantity under explicitly defined assumptions.

**Model Specification and Simulation-Defined Quantity**

For each of N = 1,000,000 simulated interrogations, we constructed a stylized but plausible event sequence:

1. **Guilt Status (G)**: A suspect's underlying guilt was drawn with $P(G = 1) = 0.80$.

2. **Confession (C):** Confessions were generated conditional on guilt, with guilty suspects confessing at a higher rate ($P(C = 1 \mid G = 1) = 0.50$) than innocent suspects ($P(C = 1 \mid G = 0) = 0.05$).

3. **Conviction (V)**: Convictions were generated conditional on both guilt and confession status, with conviction being most likely for a guilty suspect who confesses ($P(V = 1 \mid G = 1, C = 1) = 0.90$) and least likely for an innocent suspect who does not confess ($P(V = 1 \mid G = 0, C = 0) = 0.03$)

A wrongful conviction (W) was defined as an innocent suspect who is convicted ($G = 0, V = 1$) and a false-confession wrongful conviction (FCWC) as an innocent suspect who confesses and is convicted ($G = 0, C = 1$, $V = 1$). Under this construction, falsity is determined by innocence status rather than inferred from the conviction outcome itself.

These parameters combine to produce an overall conviction rate across all interrogations. In the simulation, this resulted in $P(V = 1 \mid I) = 0.569$, meaning 56.9% of interrogations ended in a conviction.

From the simulated data, we can directly observe the known interrogation-conditioned quantity. In this simulated world, the simulation-defined interrogation-conditioned probability, i.e., the proportion of all interrogations that result in an FCWC is:

$P(FCWC \mid I) = 0.00248$ (0.248%), or approximately 1 FCWC per 403 interrogations.

### Comparing Alternative Estimands Within the Simulation

We then applied the competing formulations to the data generated by the model to examine how the resulting quantities differed relative to the simulation-defined interrogation-conditioned probability of 0.248%.

The published formulation used by Mourtgos and Adams (2026) estimates the conviction-conditioned quantity $P(FCWC \mid V = 1)$. In our simulation, this value is 0.437%. If this conviction-conditioned quantity is interpreted as an interrogation-conditioned probability, it exceeds the simulation-defined interrogation-conditioned quantity by 76% (0.437% vs. 0.248%). This occurs because the formulation conditions on conviction rather than on the broader population of interrogations.

By contrast, the earlier preprint method (Mourtgos & Adams, 2025) used the following factorization: $P(C \mid I) \times P(W \mid I) \times P(F \mid W, I)$, which yields an estimate of 0.102%. Relative to the simulation-defined interrogation-conditioned quantity, this formulation produces a substantially lower estimate (nearly 60%, 0.102 vs. 0.248%) because it combines the reference-class issue with unconditional wrongful-conviction probabilities that are not restricted to confession-based cases.

The interrogation-aligned decomposition: $P(C \mid I) \times P(W \mid C, I) \times P(F \mid W, C, I)$, recovers the simulation-defined interrogation-conditioned quantity: $P(FCWC \mid I) = 0.248\%$. This occurs because the factorization mirrors the data generating structure used to construct the simulation.

### Mapping Between Conviction-Conditioned and Interrogation-Conditioned Quantities

Our simulation confirms that a conviction-conditioned quantity cannot be substituted for an interrogation-conditioned one. The mathematical relationship between them requires explicit incorporation of the probability that an interrogation results in conviction: $P(FCWC \mid I) = P(V = 1 \mid I) \times P(FCWC \mid V = 1, I)$.

In our simulation, the conviction probability was 56.9%. Applying this mapping to the conviction-conditioned quantity yields: $0.569 \times 0.00437 = 0.00248$. This demonstrates why reference-class alignment matters. Because not every interrogation results in conviction, a conviction-conditioned quantity cannot be directly interpreted as an interrogation-conditioned probability without explicitly incorporating the conviction step.

Taken together, the simulations illustrate that different formulations answer different probabilistic questions. The published formulation estimates a conviction-conditioned quantity, whereas the interrogation-aligned decomposition estimates a probability defined on the population of interrogations. The simulations further show that reference-class specification and conditioning structure can influence resulting quantities in analytically distinct ways. These issues therefore need to be evaluated separately rather than treated as a single source of distortion.

**Extension: Modelling Technique Heterogeneity**

The preceding sections established that conviction-conditioned quantities cannot be substituted for an interrogation-conditioned probability without an explicit mapping via the probability of conviction. We now extend the generative model to address a second, closely related issue: whether pooled conviction-conditioned quantities can adequately characterize a technique-conditional interrogation-conditioned quantity, $P(FCWC \mid T)$.

This extension is motivated by the substantive debate surrounding interrogation practices (Kassin et al., 2010; Meissner et al., 2017). Concerns about false confessions are not evenly distributed across interrogation contexts but are instead concentrated around specific techniques that are theorized or empirically associated with heightened risk, such as minimization (Kassin & McNall, 1991) or theme-building (Leo, 2008). Any attempt to characterize the probability associated with a particular technique therefore presupposes a quantity that reflects the interrogation process itself and allows for heterogeneity across techniques.

***Model Extension: Introducing Technique Heterogeneity***

To clarify how reference-class choice affects the interpretation of technique-conditional quantities, we augmented the data-generating process by introducing an explicit indicator of interrogation technique. Specifically, we defined a binary variable, $T \in \{0,1\}$, where $T = 1$ denotes the use of a potentially problematic interrogation tactic and $T = 0$ denotes its absence. The technique was assigned independently of guilt and case characteristics across interrogations with probability $P(T = 1) = 0.25$, reflecting the assumption that most interrogations do not employ the tactic in question.

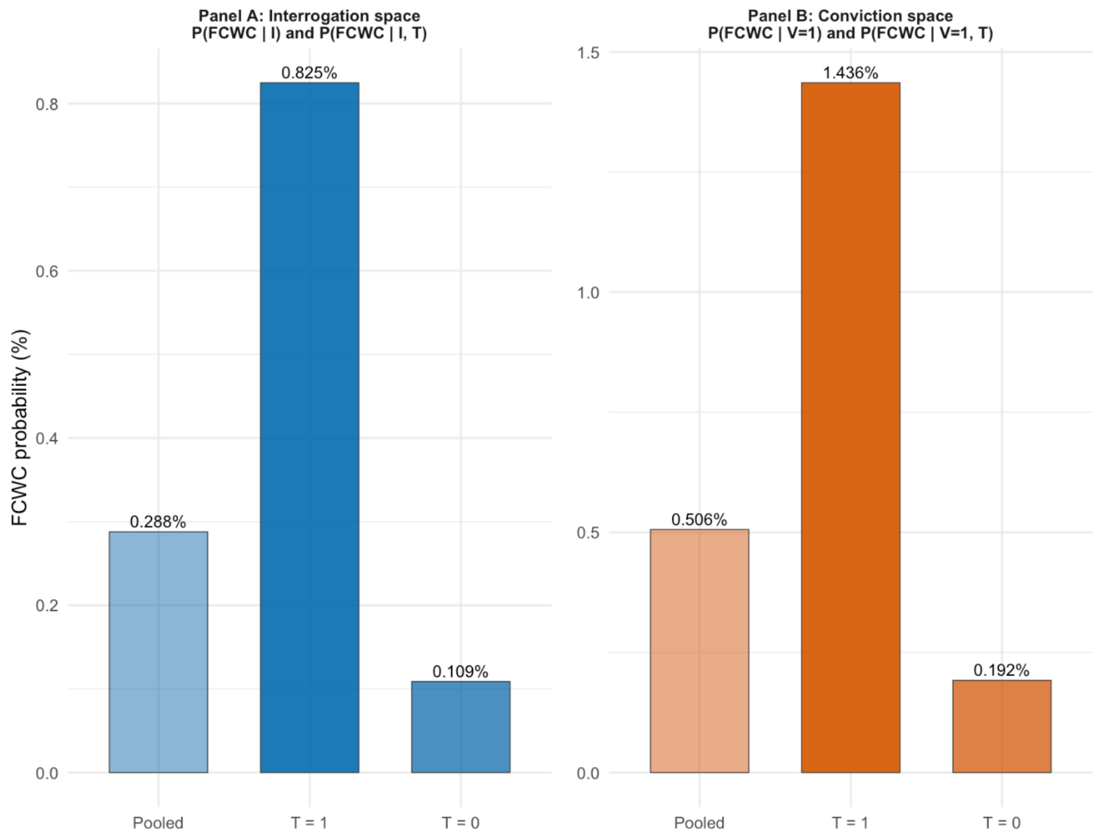
This assumption of independence is a diagnostic design choice intended to isolate the effect of reference-class and conditioning structure. Under independence, any discrepancy between interrogation-conditioned and conviction-conditioned quantities must arise from the structure of the estimand itself rather than from confounding.

Within the simulated process, the technique was specified as operating at the interrogation stage by increasing the probability that innocent suspects falsely confess and that such confessions contribute to conviction. Specifically, $T$ increased the probability that innocent suspects would falsely confess and increased the likelihood that such confessions would result in conviction. For simplicity, the simulated effect of the technique was restricted to innocent suspects. This choice was made to isolate the implications of the estimand structure rather than to model the full complexity of real interrogation dynamics. All other aspects of the data-generating process remained unchanged.

This extension allows us to compare two distinct quantities: the interrogation-conditioned probability conditional on technique use, $P(FCWC \mid T)$, and the conviction-conditioned proportion of convictions involving false-confession wrongful convictions, $P(FCWC \mid V = 1, T)$. Figure 2 contrasts these quantities directly, illustrating that while technique use may substantially increase the probability of FCWCs in both reference classes, the resulting probabilities answer different questions and are not interchangeable. In particular, conviction-conditioned estimates cannot be interpreted as interrogation-conditioned probability without explicit mapping through the probability of conviction, $P(V = 1 \mid I)$.

Our modeling choice contrasts with robustness analyses that treat interrogation tactics solely as a prevalence-weighted mixture (e.g., varying $P(T)$ while holding $P(FCWC \mid T)$ fixed). Such analyses assess how pooled prevalence estimates change under different assumed frequencies of tactic use, while holding technique-conditional quantities fixed. They therefore address the prevalence of tactics in the population rather than how tactics alter the interrogation-to-conviction process itself. By contrast, our extension specifies the technique as operating during the interrogation process itself by altering confession and conviction probabilities among innocent suspects and thereby permitting technique-conditional quantities to emerge from the simulated process, rather than being imposed solely through prevalence-weighted mixture assumptions.

**Figure 2.** ***Technique-specific FCWC quantities depends on the reference class***



*Note.* Panel A (Interrogation space) displays the probability that an interrogation results in a false-confession wrongful conviction, $P(FCWC \mid I)$, both pooled across interrogations and conditional on the presence ($T$ = 1) and absence ($T$ = 0) of a problematic technique.

Panel B (Conviction space) displays the proportion of convictions that involve a false-confession wrongful conviction, $P(FCWC \mid V = 1)$, again pooled and conditional on technique use. Blue bars correspond to interrogation-conditioned quantities (interrogation-conditioned quantity); orange bars correspond to conviction-conditioned quantities defined on the population of convictions. Although technique use substantially increases the probability of FCWCs in both panels, the quantities shown are not interchangeable. Conviction-conditioned estimates cannot be interpreted as interrogation-conditioned probabilities without explicit mapping via the conviction probability $P(V = 1 \mid I)$. The figure illustrates how failure to align the estimand with the interrogation process leads to distorted interpretation of technique-conditional quantities. Panels use different y-axis scales because the quantities are defined on different reference classes; values should be compared within, not across, panels. Percent labels provide the exact magnitudes.

### *Simulation-Defined Technique-Conditional Quantities*

The extended simulation generated distinct interrogation-conditioned quantities conditional on technique presence. In this extended model, the overall FCWC interrogation-conditioned quantity across all interrogations was: $P(FCWC \mid I) = 0.00288$ (0.288 %).

However, this pooled interrogation-conditioned quantity masks substantial heterogeneity across interrogation contexts. When conditioning on the interrogation technique, the simulation-implied interrogation-conditioned quantities differed substantially across technique conditions:

$P(FCWC \mid I, T = 1) = 0.00825$ (0.825%)

$P(FCWC \mid I, T = 0) = 0.00109$ (0.109%)

In other words, interrogations involving the problematic technique were associated with a 7.6-fold increase in the probability of producing a false-confession wrongful conviction relative to interrogations in which the technique was absent.

Despite this large difference in interrogation-conditioned quantities, the pooled FCWC interrogation-conditioned quantity remained low because the majority of interrogations did not involve the problematic tactic. This illustrates how pooled estimates can attenuate the apparent contribution of higher-risk interrogation contexts when those contexts occur relatively infrequently.

### *Implications for Technique-Conditional Estimates*

We next examined how these same data appear when expressed on the space of convictions. The conviction-level FCWC share was: $P(FCWC \mid V = 1) = 0.00506$ (0.506%), with corresponding technique-specific quantities:

$P(FCWC \mid V = 1, T = 1) = 0.01436,$

$P(FCWC \mid V = 1, T = 0) = 0.00192.$

Although these conviction-conditioned quantities preserve the rank ordering of the underlying probabilities, they remain conditional on conviction and therefore do not answer the policy-relevant question of how often the use of a particular technique produces a false-confession wrongful conviction during interrogation.

More importantly, approaches that begin with pooled conviction-conditioned quantities may attenuate differences between interrogation contexts when estimating technique-conditional quantities. Because pooled conviction-conditioned quantities average across heterogeneous interrogation contexts, they may attenuate differences between technique-specific interrogation-conditioned quantities, particularly when the technique is relatively uncommon.

#### *Summary of Generative Model Findings*

This extension of the generative model reinforces two conclusions. First, the relevant probability quantity for evaluating interrogation practices is most naturally defined on the interrogation process itself rather than on the subset of cases that result in conviction. Second, even a correctly interpreted conviction-conditioned probability is not a neutral prior for estimating technique-specific risk, because pooling across heterogeneous interrogation contexts can attenuate the risk associated with high-impact tactics.

Accordingly, conviction-conditioned pooled estimates and interrogation-conditioned technique-specific quantities should not be treated as interchangeable. The simulations suggest that pooling across heterogeneous interrogation contexts can attenuate apparent differences between techniques, particularly when higher-risk tactics are relatively uncommon within the broader population of interrogations.

### A critique of the Bayesian Workflow and Reproducibility

Mourtgos and Adams describe their component models in Bayesian terms, using hierarchical structures, weakly informative priors, and extensive MCMC sampling. The present critique does not concern the mathematical validity of Bayesian inference itself, but rather the relationship between the modeled quantity and the interrogation process the analysis is intended to represent. From the perspective of contemporary Bayesian workflow (Gelman et al., 2020), the primary concern is estimand alignment and the extent to which uncertainty about model structure is examined.

### Focus on Component Models over Structural Validity

The workflow places substantial emphasis on modeling uncertainty within the individual component quantities while devoting comparatively limited attention to whether the overall factorization corresponds to the interrogation-conditioned quantity discussed in the policy implications.

The core decomposition $P(FCWC) = P(W)\ x\ P(F|W)$ is mathematically coherent as a conviction-conditioned quantity. However, as discussed above, it does not by itself yield the probability that an interrogation produces a false-confession wrongful conviction.

Once the estimand is defined on convictions rather than interrogations, subsequent Bayesian refinement necessarily operates within that same reference class. In this sense, the principal issue is not computational sophistication, but whether the modeled quantity corresponds to the process-level question under discussion. The analysis devotes substantial effort to modeling the individual inputs but little attention to alternative decompositions of the target probability or to how different conditioning structures alter the interpretation of the resulting estimates.

#### *Structural Assumptions Implied by the Authors' Hierarchical Beta Model*

In addition to the reference-class issue, the Bayesian implementation itself imposes strong structural assumptions that are never examined. The FCWC "base rate" is modeled as the output of a hierarchical Beta distribution, implicitly treating false-confession wrongful convictions as draws from a single underlying Bernoulli process defined on a single reference class. This modeling strategy assumes that uncertainty primarily reflects incomplete knowledge about an underlying scalar probability parameter rather than systematic heterogeneity across interrogation contexts, interrogation techniques, evidentiary structures, or downstream conviction processes. Partial pooling then treats differences across studies largely as exchangeable variation around a common mean.

However, FCWCs arise through a sequence of conditional events involving interrogation, confession, conviction, and falsity, each operating on potentially different denominators and potentially varying across interrogation contexts. The concern raised here is therefore not that hierarchical Beta models are mathematically inappropriate in general, but that the resulting pooled quantity may not correspond to the interrogation-conditioned process discussed in the substantive interpretation of the findings.

Hierarchical pooling can improve stability and precision within a specified reference class, but it cannot by itself resolve uncertainty regarding whether the chosen estimand aligns with the interrogation-to-conviction sequence. Consequently, the posterior distribution may become highly precise for a quantity whose interpretation remains contingent on the underlying conditioning structure.

**Limited Exploration of Alternative Estimand Structures**

The reported sensitivity analyses focus on downstream parameters, such as the diagnostic properties of interrogation tactics or the share of police-induced FCWCs. However, the analyses do not substantially explore uncertainty arising from alternative decompositions of the target quantity itself. In a broader Bayesian workflow sense (Gelman et al., 2020), uncertainty can arise not only from parameter values within a model, but also from the specification of the estimand and the conditioning structure used to define it.

An informative extension would therefore involve comparing alternative probability decompositions explicitly, including formulations defined on interrogations rather than convictions, and examining how those alternative conditioning structures affect the resulting interpretation of FCWC quantities.

***Robustness Analyses Conditioned on a Fixed Estimand***

Mourtgos and Adams' conduct an extensive set of robustness checks in their supplementary materials, including alternative priors for sensitivity and specificity, multiple weighting schemes for study synthesis, leave-one-study-out analyses, and stress tests for underreporting. These analyses demonstrate that the posterior distribution of their FCWC base-rate estimate is stable conditional on the adopted factorization and reference class.

However, robustness to prior choice or weighting assumptions does not address uncertainty regarding whether the modeled quantity itself corresponds to the interrogation-conditioned process of interest. All reported robustness checks inherit the same conviction-conditioned FCWC quantity and therefore cannot evaluate how alternative reference classes would alter the substantive interpretation of the results. In a Bayesian workflow sense (Gelman et al., 2020), the analysis thoroughly explores uncertainty within a model but does not examine uncertainty about the model's structure itself.

This limitation is especially evident in the authors' robustness analysis varying the prevalence of potentially problematic interrogation tactics. Those analyses vary tactic prevalence while holding the underlying conviction-conditioned FCWC quantity fixed. As a result, they evaluate how pooled averages change under different assumed tactic frequencies, but they do not model how interrogation techniques influence the interrogation-to-conviction process itself.

Consequently, the analyses cannot assess whether specific interrogation techniques alter confession probabilities, conviction probabilities, or the relationship between confession and wrongful conviction. Nor can they evaluate how pooling across heterogeneous interrogation contexts may attenuate technique-conditional interrogation quantities.

### Reproducibility is Precluded by an Opaque Data Synthesis Process

A further limitation concerns reproducibility. A central component of Mourtgos and Adams' analysis is a meta-analytic synthesis of prior studies, yet the procedures underlying this synthesis are insufficiently specified. Although the authors note that studies are weighted by recency, sample size, and type of study, the exact weighting functions, parameter choices, and their justification are not fully reported. Without access to the underlying datasets, weighting schemes, or simulation code, independent verification of the reported base-rate distribution is not possible. This opacity makes it difficult to assess the extent to which the final estimates reflect empirical evidence versus subjective modeling choices.

### Superficial Reporting of Computational Details

The authors report MCMC settings like "four Markov chains with 251,000 iterations each," but these figures are insufficient to establish computational reliability. Iteration counts without convergence diagnostics (e.g., R-hat), effective sample sizes, and summaries of any divergences do little to validate the posterior distribution. This level of reporting demonstrates only that many samples were drawn from the chosen model, not that the model is appropriate or that the samples accurately represent the true posterior.

Importantly, however, the present critique does not depend on computational convergence issues. The central concern instead involves the interpretation of the modeled quantity and the relationship between the adopted factorization and the interrogation-conditioned process under discussion.

## Discussion

### Summary of the Core Argument

This analysis has identified an estimand-alignment problem in the probabilistic model used by Mourtgos and Adams (2025) to discuss the base rate of false-confession wrongful convictions (FCWCs). Specifically, we argue that the published formulation estimates a conviction-conditioned quantity, whereas many of the surrounding policy interpretations concern interrogation-conditioned quantities or probabilities.

An earlier preprint version additionally contained a redundant confession-prevalence term that mechanically attenuated the resulting estimate. Although this issue was corrected prior to publication, the published formulation still relies on a conviction-conditioned decomposition that differs from the interrogation-conditioned quantity examined throughout the present paper.

Our analysis identified two related issues:

1. the use of an overall wrongful-conviction quantity ($P(W)$) rather than a confession-conditioned quantity relevant to confession-based cases; and
2. the combination of quantities defined on different reference classes

including interrogations, convictions and exonerations.

The generative simulations clarify the central point of the present critique: different factorizations recover different quantities even when applied to identical underlying processes. In particular, conviction-conditioned quantities cannot be interpreted as interrogation-conditioned probabilities without explicit mapping through the probability of conviction.

More broadly, the findings illustrate a general methodological principle: increasing statistical sophistication within a model does not resolve uncertainty regarding whether the estimand itself corresponds to the underlying process being studied. Bayesian refinement can improve inference conditional on a specified structure, but it cannot by itself determine whether the selected conditioning structure aligns with the substantive policy question under discussion.

Finally, the present analysis also highlights the importance of transparency and reproducibility in probabilistic synthesis models. Although the published article provides a general description of the weighting and pooling framework, the absence of fully reproducible synthesis procedures and computational implementation details limits independent evaluation of how strongly the resulting posterior distribution depends on particular modeling assumptions.

**Independence Assumptions and Conservative Design**

A key feature of our generative simulation is the assumption that the use of a potentially problematic interrogation technique is independent of guilt status and other case characteristics. This assumption is not intended as a substantive claim about real-world practice. Rather, it is a deliberate design choice intended to isolate the consequences of alternative estimand structures.

By imposing independence, we eliminate confounding, selection effects, and strategic deployment of interrogation techniques as alternative explanations for discrepancies between conviction-conditioned and interrogation-conditioned estimates. Under this assumption, any distortion in estimated FCWC quantities must arise solely from differences in factorization and reference class, not from endogenous assignment of techniques.

This makes the simulations conservative in an important sense. In real-world settings, interrogation tactics are unlikely to be randomly distributed across cases. Potentially problematic techniques may be disproportionately used in difficult investigations, high-pressure contexts, or cases involving vulnerable suspects, conditions that could plausibly increase false-confession risk. Demonstrating substantial divergence between estimands even under independence therefore strengthens the broader argument regarding reference-class alignment.

At the same time, relaxing the independence assumption would not resolve the central issue identified in the present analysis. Even with perfect covariate adjustment, a conviction-conditioned quantity cannot be interpreted as the probability that an interrogation involving a specific technique produces a false-confession wrongful conviction unless the conviction process is modeled explicitly.

Recent demonstrations in which large language models generated false admissions under adversarial questioning further illustrate that interactive systems can produce unreliable admissions even in the absence of actual culpability. Although such systems differ fundamentally from human suspects, these demonstrations underscore the broader principle that confession-like outputs can emerge from interaction dynamics rather than underlying guilt.

**Limitations**

Several limitations of the present analysis should be acknowledged. Our critique focuses specifically on the probabilistic structure and interpretation of the published factorization and does not re-analyze the underlying empirical literature synthesized by Mourtgos and Adams (2026). The present critique therefore concerns estimand structure and interpretation rather than the empirical validity of any individual source study.

The purpose of our generative model is diagnostic rather than inferential: to assess how alternative probability decompositions behave under controlled assumptions where the underlying process is known by construction. In addition, the simulations necessarily simplify the interrogation-to-conviction process. Real-world interrogations involve substantial heterogeneity in evidentiary strength, suspect vulnerability, interrogation duration, legal safeguards, and downstream judicial decision-making.

The present simulations abstract away from many of these complexities in order to isolate the consequences of alternative conditioning structures. The paper does not attempt to estimate the actual causal effect of any specific interrogation technique on false-confession risk. The technique-heterogeneity simulations are intended to illustrate how pooled quantities can attenuate or obscure technique-specific differences under heterogeneous interrogation contexts, not to quantify the empirical risk associated with particular real-world interrogation practices.

These limitations, however, do not alter the central methodological argument of the paper: namely, that conviction-conditioned quantities and interrogation-conditioned quantities represent different estimands and therefore answer different substantive questions.

**Pathway to an Empirically Valid Estimate of *P(FCWC | I)***

A defensible empirical estimate of interrogation-conditioned FCWC risk would require data that links interrogation practices to downstream case outcomes across the full population of interrogations, not only exonerations or convictions. Achieving a valid estimate of $P(FCWC \mid I)$ therefore requires a comprehensive and methodologically rigorous approach:

1. **Comprehensive Data Collection:** Gather with the cooperation from law enforcement agencies a complete dataset of all interrogations conducted within a defined jurisdiction and time period. The use of real case data directly addresses the challenge of ecological validity, moving beyond laboratory paradigms to the institutional context where the risks materialize.
2. **Coding of Interrogation and Confession Content:**
    a. **Technique-Specific Data:** Record the interrogation technique or techniques used in each interrogation
    b. **Confession Reliability Analysis:** Record the presence of any confession in each interrogation.
        i. **Case Review:** Compare content from the interrogation with gathered and preferably verifiable case evidence.
3. **Systematic outcome tracking:** For the entire cohort of interrogations, track cases downstream to their final disposition. This requires systematically reviewing case outcomes, including tracking known exonerations and conducting independent reviews to

identify potential wrongful convictions among non-exonerated cases. By starting with the full population of interrogations and tracking all outcomes, this design eliminates the outcome-based selection bias inherent in analyzing only known false-confession cases.

4. **Statistical Modeling:** Use robust statistical models to estimate the two key probabilities:

    a. ***P(Uncorroborated Confession | T):*** The probability of an uncorroborated or false confession conditional on an interrogation technique

    b. **P(FCWC | I):** The probability of a false-confession wrongful conviction, conditional on an interrogation having occurred

5. **Transparency and Reproducibility:** Ensure all methodologies and non-sensitive data are transparent and available for independent verification.

By addressing these requirements, researchers can move towards producing a reliable estimate of $P(FCWC \mid I)$ that accurately reflects the risk inherent in interrogation practices, ultimately contributing to a deeper understanding of how false confessions lead to wrongful convictions.

Additionally, utilizing historical data could enhance this estimation process. For instance, one could sample around known exonerations by analyzing video-recorded interrogations from those cases and comparing them to other cases from the same jurisdiction and time period. This approach would allow researchers to identify patterns and contextual factors associated with false confessions, providing valuable insights into the conditions under which they occur. It may also identify specific interrogation techniques, such as the Reid technique's use of minimization and theme building that are more prevalent in FCWC cases, providing valuable insights for policy and practice reform.

**Implications for Policy and Practice**

Despite the above-mentioned limitations, our analysis has clear implications for policy, training and evidence-based investigative interviewing.

**1. The "Below One Percent" Claim Reflects a Conviction-Conditioned Quantity Rather than an Interrogation-Conditioned Estimate.**

Most critically, claims that lawful interrogation tactics produce FCWCs at a rate "below one percent" should be interpreted in light of the underlying conditioning structure of the model. Because the structure of the model defines the quantity on the space of convictions rather than interrogations, the resulting figure does not directly estimate the probability that an interrogation produces an FCWC.

In addition, the apparent precision of the Mourtgos and Adams' FCWC estimates may obscure the extent to which they depend on modeling assumptions rather than direct observation. The resulting quantity is shaped by explicit weighting decisions that assign comparatively less influence to older studies, despite the possibility that such studies may remain informative in an empirical domain where underlying interrogation practices and error mechanisms are not necessarily time-dependent. Such choices may be reasonable for some purposes, but they cannot substitute for models that align the estimand, denominator, and causal sequence with the interrogation process itself. Policymakers, courts, and practitioners relying on the authors' below one percent claim may draw unwarranted conclusions about the safety and reliability of certain investigative practices.

Notably, this conclusion holds even under conservative assumptions that treat technique use as randomly distributed across interrogations. In practice, techniques of concern are often deployed selectively in high-pressure or high-suspicion contexts, implying that conviction-conditioned estimates may fail to capture technique-specific risk to an even greater degree.

More fundamentally, the analysis highlights a difference in analytic perspective. Mourtgos and Adams frame interrogation techniques as probabilistic indicators that may coincide with FCWCs and evaluate risk retrospectively by conditioning on conviction-outcomes. Policy, training, and regulation, however, operate prospectively. Interrogation techniques are discretionary actions chosen by investigators, not passive markers observed after the fact. From a prevention and safety perspective, the relevant question is therefore not whether a tactic appears in a small proportion of FCWC cases, but whether its use increases the probability that an interrogation produces an FCWC relative to feasible alternatives. Framing technique evaluation around pooled conviction-conditioned quantities risks obscuring this distinction and normalizing avoidable sources of error.

Although the supplementary analyses show that conviction-conditioned FCWC estimates are stable under a wide range of assumptions, stability alone does not imply policy relevance. Training, regulation, and oversight operate at the interrogation stage, not at the subset of cases that ultimately result in conviction. Robustness within a conviction-space model therefore cannot establish the safety of interrogation practices without an explicit mapping to interrogation-conditioned quantities or probabilities.

**2. The Case for Reform is independent of a Specific Base Rate**.

The case for reforming investigative interviewing does not rest on the precise prevalence of false confessions, but on the reliability, evidentiary quality, and ethical defensibility of interrogation methods. A substantial body of research demonstrates that accusatorial interrogation methods are prone to eliciting unreliable information, can exacerbate investigator confirmation bias, and are inconsistent with modern scientific understanding of memory and cognition (Kassin et al., 2010; Gudjonsson, 2018). These methods do not represent a harmless status quo; they carry well-documented legal, ethical, and evidentiary risks (Gudjonsson, 2021).

Importantly, the empirical record of known exonerations is not neutral with respect to interrogation style. Evidence from documented exoneration cases indicates that false confessions are overwhelmingly associated with accusatorial and psychologically coercive interrogation practices, whereas information-gathering approaches are rarely implicated. Although exoneration data cannot establish population-level prevalence estimates of FCWCs, the asymmetric representation of interrogation styles in known failures provides a strong precautionary signal for policy and training (Leo & Ofshe 1997; Mindthoff & Meissner, 2023; Trainum, 2016). Even in the absence of precise base-rate estimates, this asymmetry supports a precautionary inference: techniques that are disproportionately represented in known failures warrant restriction or replacement when viable alternatives exist.

Accordingly, the policy question is not whether the average probability of an FCWC remains numerically small when high-risk techniques are used infrequently, but whether those techniques should be used at all given the availability of methods that achieve investigative goals with lower theoretical and empirical risk (Gudjonsson, 2021).

**3. Viable and evidence-based alternatives do exist.**

A growing body of research supports science-based interviewing methods that are more effective at eliciting information without the associated risk of false confession. Some of these include:

- The Cognitive Interview (Fisher & Geiselman, 1992)
- The Strategic Use of Evidence (SUE) approach (Hartwig et al., 2014)
- The Observing Rapport Based Interpersonal Techniques (ORBIT; Alison et al., 2013)

These methods have been validated in laboratory and/or field studies and adopted by major law enforcement organizations globally. They do not seek to generate confessions but to elicit accurate, verifiable information. Use of such information-gathering approaches has been found to elicit confessions at a similar rate to accusatorial methods without the associated risk of generating false confessions.

**4. The Premise of a "Tolerable" Error Rate Warrants Scrutiny.**

Finally, the premise that a low base rate of false-confession wrongful convictions represents a 'tolerable systemic cost' warrants scrutiny. This framing implicitly treats FCWCs as rare but unavoidable by-products of an otherwise effective interview process. However, interrogation is not merely an error-producing system; it is an information-production system, and policy choices necessarily involve trade-offs between information yield and the risk of serious downstream error.

From a prevention perspective, the relevant question is therefore not whether FCWCs are numerically rare on average, but whether particular interrogation techniques increase the probability that an interrogation produces unreliable or misleading information relative to available alternatives. Techniques that yield marginal gains in confessions or apparent cooperation but disproportionately elevate the risk of false or uncorroborated admissions, may be suboptimal even when the absolute incidence of wrongful conviction remains low. Framing acceptability around pooled base rates obscures this yield–risk trade-off and risks normalizing avoidable sources of error.

False confessions impose especially severe costs: they result in the wrongful incarceration of the innocent while allowing true perpetrators to remain at large, thereby posing an ongoing public safety threat. A more appropriate analogy than rare medical complications is to safety-critical systems such as aviation, where known, correctable sources of catastrophic failure are addressed proactively rather than accepted as an unavoidable cost of operation. In such domains, decisions are not guided by base rates alone. Instead, acceptable risk is evaluated relative to preventability, severity, and the availability of less harmful alternatives.

Applied to interrogation practice, this perspective implies that the continued use of techniques that are disproportionately represented in known failures demands justification not by their average error rate, but by their necessity. Where methods exist that achieve investigative goals with comparable or superior informational yield and lower theoretical and empirical risk, tolerating preventable error becomes difficult to defend—regardless of whether the aggregate base rate of FCWCs appears small.

## Conclusion

This paper has examined the mathematical foundation of a prominent quantitative claim used to defend the status quo in police interrogations. We have shown that the underlying probabilistic model is defined on a conviction-conditioned reference class rather than directly on the interrogation-to-conviction process discussed in many policy interpretations. The central issue identified in the present analysis is therefore one of estimand alignment and interpretation rather than computational validity. Correcting this issue is important not because it yields a definitive estimate of the prevalence of false-confession wrongful convictions, but because it removes a misleading quantitative anchor that has been used to argue that reform is unnecessary or disproportionate.

A central finding is that pooling risk across heterogeneous interrogation contexts systematically obscures substantial technique-specific differences in false-confession wrongful-conviction risk. Even when interpreted correctly, conviction-conditioned quantities cannot serve as neutral foundations for evaluating interrogation practices when high-risk techniques are diluted by aggregation. Low pooled conviction-conditioned quantities can coexist with sharply elevated interrogation-conditioned probabilities, particularly when problematic techniques are used infrequently but have disproportionate downstream consequences.

Importantly, our analysis does not depend on assumptions about the prevalence, strategic deployment, or effectiveness of particular interrogation techniques. Rather, it demonstrates a more general principle: no estimation strategy can recover technique-specific interrogation risk from conviction-conditioned quantities without explicitly modeling the interrogation-to-conviction sequence. More broadly, the paper illustrates how different factorizations can produce different substantive interpretations even when based on internally coherent probability calculations.

The broader implication is a shift in analytic perspective. Evaluating interrogation practices requires process-aligned risk metrics that reflect what techniques do during interrogations such as how they affect information quality, error generation, and downstream outcomes rather than post hoc summaries conditioned on conviction. From a prevention standpoint, the relevant question is not whether errors are rare on average, but whether they are avoidable given available alternatives that achieve investigative goals with lower risk.

Accordingly, the path forward lies not in defending legacy practices on the basis of pooled conviction-conditioned quantities, but in supporting the implementation of safeguards and methods grounded in modern investigative interviewing. These include treating confessions as hypotheses requiring independent corroboration, ensuring full audio-visual recording of interviews, and prioritizing the elicitation of verifiable information over confession production. Aligning evaluation metrics with the interrogation process itself is a necessary step toward reducing preventable error while improving the quality and reliability of investigative outcomes.

## REFERENCES


Alison, L. J., Alison, E., Noone, G., Elntib, S., & Christiansen, P. (2013). Why tough tactics fail and rapport gets results: Observing Rapport-Based Interpersonal Techniques (ORBIT) to generate useful information from terrorists. Psychology, Public Policy, and Law, 19(4), 411–431. https://doi.org/10.1037/a0034564

Fisher, R. P., & Geiselman, R. E. (1992). *Memory-enhancing techniques for investigative interviewing: The cognitive interview*. Charles C Thomas

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Harvard University Press. https://doi.org/10.4159/harvard.9780674060982

Gross, S. R., O'Brien, B., Hu, C., & Kennedy, E. H. (2014). Rate of false conviction of criminal defendants who are sentenced to death. *Proceedings of the National Academy of Sciences of the United States of America*, 111(20), 7230-7235. https://doi.org/10.1073/pnas.1306417111

Gudjonsson, G. H. (2018). The psychology of false confessions: Forty years of science and practice. John Wiley & Sons Ltd. https://doi.org/10.1002/9781119315636

Gudjonsson, G. H. (2021). The science-based pathways to understanding false confessions and wrongful convictions. Frontiers in Psychology, 12, Article 633936. https://doi.org/10.3389/fpsyg.2021.633936

Hartwig, M., Granhag, P. A., & Luke, T. (2014). Strategic use of evidence during investigative interviews: The state of the science. In D. C. Raskin, C. R. Honts, & J. C. Kircher (Eds.), *Credibility assessment: Scientific research and applications* (pp. 1–36). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-394433-7.00001-4

Jaynes, E. T. (2003). *Probability theory: The logic of science* (G. L. Bretthorst, Ed.). Cambridge University Press.

Kassin, S. M., Drizin, S. A., Grisso, T., Gudjonsson, G. H., Leo, R. A., & Redlich, A. D. (2010). Police-induced confessions: risk factors and recommendations. *Law and Human Behavior*, 34(1), 3–38. https://doi.org/10.1007/s10979-009-9188-6

Kassin, S. M., & McNall, K. (1991). Police interrogations and confessions: Communicating promises and threats by pragmatic implication. Law and Human Behavior, 15(3), 233–251. https://doi.org/10.1007/BF01061711

Kassin, S. M., Bogart, D., & Kerner, J. (2012). Confessions That Corrupt: Evidence from the DNA Exoneration Case Files. *Psychological Science*, 23(1), 41–45.

https://doi.org/10.1177/0956797611422918

Leo, R. A. (2008) Police Interrogation and American Justice. Harvard University Press.

Leo, R. A., & Ofshe, R. J. (1997). Consequences of false confessions: Deprivations of liberty and miscarriages of justice in the age of psychological interrogation. *The Journal of Criminal Law and Criminology*, 88, 429–496. https://doi.org/10.2307/1144288

Mindthoff, A., & Meissner, C. A. (2023). Interviewing and interrogations: From the third degree to science-based approaches. In E. Verona & B. Fox (Eds.) *Routledge Handbook of Evidence-Based Criminal Justice Practices* (1st ed.). Routledge. https://doi.org/10.4324/9781003219286

Mourtgos, S. M., & Adams, I. T. (2025). Reasonable Doubt: Quantifying the Risk of Police Interrogation Tactics in False Confession Wrongful Convictions [Preprint]. *CrimRxiv*. https://doi.org/10.21428/cb6ab371.1982e32d

Mourtgos, S. M., & Adams, I. T. (2026). Recalibrating the Risk of False Confession Wrongful Convictions: Interrogation Tactics and Inverse Probability. *Journal of Criminal Justice*, 103, Article 102600. https://doi.org/10.1016/j.jcrimjus.2026.102600

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Trainum, J. (2016). How the police generate false confessions: An inside look at the interrogation room. Bloomsbury Academic

Wickham, H. (2016). *Elegant Graphics for Data Analysis*. Springer-Verlag.

INVESTIGATIVE INTERVIEWING

# Research & Practice

The official journal of the International Investigative Interviewing Research Group

iIIRG

INTERNATIONAL
INVESTIGATIVE
INTERVIEWING
RESEARCH GROUP