

THE UNDEREXPLORED PSYCHOLOGY OF THE EXCITED UTTERANCE AND PRESENT SENSE IMPRESSION

Justin Sevier*

The present-sense impression (PSI) and excited-utterance (EU) hearsay exceptions are justified by two related psychological claims: that contemporaneity reduces fabrication and memory distortion, and that arousal stills the capacity for deliberate deception. Courts routinely rely on these claims, and the legal debate over these exceptions has largely relied on them as the organizing framework for empirical study. But these claims are incomplete in ways that the existing debate does not fully appreciate. Reliability, as it functions in proceedings involving PSIs and EUs, is multidimensional. It depends on what the declarant perceives and narrates (veracity), on how faithfully that narration is transmitted to the factfinder (fidelity), and on how accurately the factfinder weighs what she receives (calibration). The debate has focused substantially on the first stage, with notable gaps, and has paid little attention to the other two.

This Article maps that imbalance empirically. Drawing on a gap map of the typical sources cited in the major contributions to the empirical PSI-EU reliability debate, the analysis shows that all mapped sources address the veracity stage, while fidelity and calibration in the PSI-EU context receive almost no direct treatment. It then argues that this distribution explains why scholars have examined substantially the same empirical evidence and reached different conclusions—they focus primarily on veracity while often accepting the doctrine’s implicit assumption that securing the declarant’s statement secures the reliability of the evidence.

This Article then presents a research agenda comprising three pilot studies: one for each stage of the reliability chain. In the veracity pilot,

* Charles W. Ehrhardt Professor of Litigation, Florida State University College of Law. The author thanks the participants of the *Fordham Urban Law Journal* Symposium, “The Law’s Blind Spots: A Conversation on Outdated Legal Assumptions in Light of Cognitive Science” (Feb. 2026), and the editors of the *Fordham Urban Law Journal* for their helpful comments.

PSI- and EU-like narrators produced substantial inferential content even in immediate-narration conditions, and arousal amplified the confidence-accuracy gap rather than reducing it. The fidelity pilot suggests that the emotional charge of an excited utterance can systematically produce “hedge loss” and “certainty inflation” in recipients’ retellings, such that the emotionality that triggers the statement’s admissibility may also degrade the statement’s fidelity of transmission. The calibration pilot, in turn, suggests that factfinders may apply a blunt hearsay discount without sensitivity to the doctrinal distinctions courts use in making admissibility determinations. This Article concludes with reform recommendations and a research agenda which is better matched to the specific stages of the PSI-EU reliability problem.

Introduction	1190
I. Evidence Law Reliability Stories.....	1194
A. Two Rules, Twin Stories.....	1194
B. The Present Sense Impression: Timing as Proxy.....	1197
C. The Excited Utterance: Arousal as Proxy	1198
D. Empirical Assumptions	1199
II. The Empirical Debate	1201
A. The Current Literature	1201
B. Opposing Conclusions.....	1207
III. Hearsay Error Signatures.....	1209
A. Pilot Study: Veracity	1209
B. Pilot Study: Fidelity.....	1213
C. Pilot Study: Calibration.....	1214
IV. Legal Implications	1216
A. Asking Better Questions.....	1216
B. Improving Transmission.....	1217
C. Giving Factfinders Better Tools.....	1218
V. Empirical Implications.....	1220
A. Veracity: Ecologically Valid Designs	1220
B. Fidelity: Testing Correctives	1221
C. Calibration: Fine-Grained Weighting.....	1222
D. The Cascading Design: Tracing Distortion Across the Chain.....	1223
Conclusion.....	1224

INTRODUCTION

On the afternoon of March 19, 2012, three men were standing on the corner of 129th Street and St. Nicholas Terrace in Harlem when a

minivan slowly rolled past them—and a passenger climbed out. He pointed a gun at one of the three men and opened fire, wounding all of them before fleeing back into the minivan and speeding away.¹ One of the victims called 911, and, about twenty seconds into the call, a faint but audible voice was heard, belonging to an unidentified person standing in the gathered crowd.² The voice stated, “Yo, it was Twanek, man! It was Twanek, man!”³

That (presumptively) spontaneous and emotional statement became the centerpiece of the criminal prosecution of Twanek Cummings. The evidence against him was otherwise thin; the shooter had worn a hoodie, was unidentifiable on surveillance video, and none of the victims could identify him in a lineup.⁴ The person in the background of the 911 call who had named Cummings was never identified either—no one knew who he was or what he had seen.⁵ The prosecution’s identification evidence consisted almost entirely of those nine words and a fingerprint found on the passenger door of the minivan.⁶

The bystander’s statement was offered at trial as an excited utterance. Federal Rule of Evidence 803(2) and its state analogs provide that a statement relating to a startling event, made while the declarant is still under the stress of excitement it caused, is admissible as an exception to the rule barring hearsay evidence.⁷ The theory is that stress stills reflection and, therefore, the ability to lie.⁸ An excited person is presumed to be unable to easily fabricate, so their statement is presumed sincere.⁹ Courts have applied this logic for well over a century, and the “excited utterance” has become a fixture of criminal

1. See *People v. Cummings*, 31 N.Y.3d 204, 206 (N.Y. 2018). The factual account of the *Cummings* litigation is drawn from the New York Court of Appeals opinion, the record of proceedings below, and contemporaneous court documents. Citations to specific facts appear in the footnotes that follow.

2. See *id.* at 206 (“The exact timing is disputed, but somewhere between 2:29 p.m. and 2:32 p.m., Mr. Phillips called 911.”).

3. *Id.* at 207.

4. See *id.* at 211.

5. See *id.* at 207 (“The surveillance video does not show the scene of the shooting. Neither the shooter nor the person saying, ‘It was Twanek, man,’ can be identified from the video.”).

6. See *id.* (noting other circumstantial evidence, including cell-site data, a partial license plate number, and a getaway van description).

7. See FED. R. EVID. 803(2); see, e.g., FLA. STAT. §90.803(2) (2025) (Florida’s excited utterance hearsay exception).

8. See FED. R. EVID. 803(2) advisory committee’s note to proposed rules.

9. See *id.*

prosecutions, most commonly in the context of 911 calls made by victims and bystanders in the immediate aftermath of a violent crime.¹⁰

The court excluded the statement at Cummings's first trial,¹¹ which ended in a jury deadlock.¹² But at the retrial, the statement was eventually admitted as an excited utterance.¹³ The prosecution then, in its closing summation, relied heavily on the out-of-court speaker's nine words.¹⁴ Cummings was convicted of first-degree assault and related charges.¹⁵

The New York Court of Appeals reversed the conviction in 2018.¹⁶ The court held that the statement was inadmissible because the record contained no evidence from which a jury could have reasonably inferred that the unidentified declarant had personally observed the shooting.¹⁷ The Court of Appeals also held that the error was not harmless because, without the statement, the evidence of guilt was not overwhelming.¹⁸

*People v. Cummings*¹⁹ begins as a case about a foundational requirement of the excited-utterance hearsay exception: the declarant must have uttered the statement while subjectively under the stress of excitement from an event he personally witnessed. Although the prosecution did not do so, it also could have sought admission under the present-sense impression exception—a close cousin that admits a speaker's real-time factual description of an event perceived in the

10. See Steven Baicker-McKee, *The Excited Utterance Paradox*, 41 SEATTLE U. L. REV. 111, 115–20, 124, 126–27 (2017) (describing the genesis and history of the excited utterance exception); see also Timothy T. Lau, *Reliability of Excited Utterance Hearsay Evidence*, 87 MISS. L.J. 599, 640 (2018) [hereinafter Lau (EU)] (“The critics of the exception who nonetheless seek to abolish it will have to confront the practical reality that the exception is used to admit statements from victims of domestic violence who, because of their vulnerability, cannot or will not testify about what happened to them.”).

11. See *Cummings*, 31 N.Y.3d at 207.

12. See *id.*

13. The judge who initially presided at the retrial also excluded the statement before falling ill. Her replacement, however, admitted the statement when the prosecution renewed its application to use it. See *id.*

14. See *id.* at 213.

15. See *id.* at 207.

16. See *id.* at 213.

17. See *id.* at 212 (“Because there is no evidence from which a reasonable inference can be drawn that the declarant personally observed the incident, admission of the statement heard in the background of the 911 call was error.”).

18. See *id.* (describing the call records, the victims' lineup failures, and the fingerprint evidence). The court stated that the standard for harmless error is “whether proof of guilt was overwhelming, and whether there was no significant probability that the jury would have acquitted had the proscribed evidence not been introduced.” *Id.* The court concluded that the “standard is not met here.” *Id.*

19. 31 N.Y.3d 204, 207 (2018).

moment, subject to the same personal-knowledge requirement.²⁰ Both exceptions, in other words, rest on the same unstated behavioral assumption: that the speaker was actually perceiving and reporting accurately rather than inferring, confabulating, or miscommunicating. That is the critical assumption this Article addresses.

This unstated assumption raises several distinct questions for judges—who decide whether to admit these statements into evidence—and for jurors, who subsequently evaluate them on the merits if admitted. Was the speaker reporting a direct perception? Was she repeating something she had heard from someone else in the crowd? Or critically, was she inferring identity from partial information (for example, a half-glimpsed face) and reporting that inference with the confident certainty that arousal and social contagion can sometimes produce?²¹ If the out-of-court speaker is not simply lying, these questions require us to characterize the out-of-court statement in one of three ways: as (1) a snapshot of the speaker's clear perception and accurate memory, (2) the speaker's sincere but entirely mistaken belief, or (3) something in between. For example, does the statement reflect the kind of rapid, unintentionally biased, reconstructive account that people may form from incomplete or ambiguous perceptions—particularly during a fast-moving or emotionally charged event?²²

I argue that the law surrounding the present-sense impression and excited-utterance hearsay exceptions cannot truly address the possibility of sincere but inaccurate statements, in part because it was not designed to. The exceptions ask whether certain conditions, thought to reduce the likelihood that an out-of-court statement was fabricated, were present—but that is where they stop. Specifically, they treat contemporaneity and excitement as signals of reliability rather than as cognitive conditions that can exert complex effects on perception,

20. See FED. R. EVID. 803(1).

21. See generally Kenneth Deffenbacher et al., *A Meta-Analytic Review of the Effects of High Stress on Eyewitness Memory*, 28 L. & HUM. BEHAV. 687 (2004) (discussing the effects of stress on eyewitness memory); Elin M. Skagerberg & Daniel B. Wright, *The Co-Witness Misinformation Effect: Memory Blends or Memory Compliance?*, 16 MEMORY 436, 440–42 (2008) (discussing “memory blends”—the effects of social contagion on memory).

22. These kinds of accounts are shaped by selective attention, source confusion, and plausible inference rather than a literal record of what was observed. See DANIEL SCHACTER, *THE SEVEN SINS OF MEMORY: HOW THE MIND FORGETS AND REMEMBERS* 91–99, 163–65 (2001) (discussing source confusion, memory-binding errors, false recognition based on associative memory, and emotional-arousal effects on attention and recall); see also Ulric Neisser & Nicole Harsch, *Phantom Flashbulbs: False Recollections of Hearing the News About Challenger*, in *AFFECT AND ACCURACY IN RECALL: STUDIES OF “FLASHBULB” MEMORIES* 9–31 (E. Winograd & U. Neisser eds., 1992).

memory, and narrative ability. In doing so, they leave largely unexamined the possibility that a statement can be entirely sincere, made by a person who had no intention to deceive, and still be wrong in ways that neither the declarant, the in-court witness, nor the court can fully assess and appreciate.

This Article explores this issue in depth and proceeds in several parts. Part I discusses the cognitive reliability stories that underlie the present-sense impression and excited-utterance exceptions to the hearsay rule and that afford them favored evidentiary status. Part II examines the empirical debate surrounding those stories, focusing on what the debate has covered well and what is missing. Part III introduces three pilot studies that begin to fill those empirical gaps. Parts IV and V present targeted reform recommendations and an ecologically valid research agenda for policymakers and researchers, informed by the preliminary findings from the pilot studies. Part VI concludes by returning to the *Cummings* case and the cognitive questions it raises.

I. EVIDENCE LAW RELIABILITY STORIES

A. Two Rules, Twin Stories

Rules 803(1) and 803(2) of the Federal Rules of Evidence carve out from the general prohibition on hearsay two categories of out-of-court statements that the law has deemed sufficiently trustworthy to be admitted without the usual safeguards of sworn testimony and cross-examination. The first, the present-sense impression (PSI) exception, admits “[a] statement describing or explaining an event or condition, made while or immediately after the declarant perceived it.”²³ The second, the excited-utterance (EU) exception, admits “[a] statement relating to a startling event or condition, made while the declarant was under the stress of excitement that it caused.”²⁴ Although they have distinct doctrinal requirements, both exceptions rest on the same logic: that certain circumstances attending the making of a statement provide sufficient indicia of reliability to substitute for the procedural protections that cross-examination would otherwise supply.

The Federal Rules of Evidence Advisory Committee (Advisory Committee) notes articulate that logic in a pair of short passages. For the PSI, reliability derives from timing: the “substantial

23. FED. R. EVID. 803(1).

24. FED. R. EVID. 803(2).

contemporaneity of event and statement negative[s] the likelihood of deliberate o[r] conscious misrepresentation.”²⁵ For the EU exception, reliability derives from arousal: “circumstances may produce a condition of excitement which temporarily stills the capacity of reflection and produces utterances free of conscious fabrication.”²⁶ In each case, the exception relies on a psychological claim, about what proximity to an event does to the mind, or what excitement does to the impulse to deceive. These are empirical propositions about human cognition, and they are arguably stated with a confidence that the Advisory Committee notes cannot vigorously support or defend.

To be fair, that confidence is not unique to the hearsay exceptions. As Professor Edward Cleary observed decades ago, evidentiary doctrine has generally been constructed from “anecdote[s] and unsystematic observation, plus what hopefully passes for reason but could more honestly be labeled conjecture about human behavior.”²⁷ Professor Paul Meehl’s term “fireside induction”—a commonsense generalization accepted on the basis of cultural familiarity rather than systematic study—similarly captures the epistemology on which the PSI and EU exceptions were built.²⁸ Applying the fireside induction to evidence law more broadly, Professor Gregory Mitchell has stated that behind many evidence rules lurks “much armchair theorizing about human nature . . . and little careful study . . .”²⁹

Skeptical observations regarding the origin of the categorical evidentiary hearsay exceptions have also begun to appear in judicial opinions.³⁰ In perhaps the best-known example, Judge Richard Posner,

25. FED. R. EVID. 803 advisory committee’s note to Rule 803(1) (“The underlying theory of [the present-sense impression exception] is that substantial contemporaneity of event and statement negative[s] the likelihood of deliberate o[r] conscious misrepresentation.”).

26. FED. R. EVID. 803 advisory committee’s note to Rule 803(2) (“The theory of Exception (2) is simply that circumstances may produce a condition of excitement which temporarily stills the capacity of reflection and produces utterances free of conscious fabrication.”).

27. MICHAEL J. SAKS & BARBARA A. SPELLMAN, *THE PSYCHOLOGICAL FOUNDATIONS OF EVIDENCE LAW* 14 (2016) (quoting Edward W. Cleary, *Evidence as a Problem in Communicating*, 5 VAND. L. REV. 277, 279 (1952)).

28. Paul E. Meehl, *Law and the Fireside Inductions (with Postscript): Some Reflections of a Clinical Psychologist*, 7 BEHAV. SCI. & L. 521, 522 (1989) (“Analogously, by the language ‘fireside inductions’ I mean those commonsense empirical generalizations about human behavior which we accept on the culture’s authority plus introspection plus anecdotal evidence from ordinary life.”).

29. Gregory Mitchell, *Beyond FireSide Inductions*, 32 FLA. ST. U. L. REV. 315, 315 (2005).

30. See, e.g., *People v. Cummings*, 31 N.Y.3d 204, 215 (N.Y. 2018) (Rivera, J., concurring) (“It appears that only tenuous support exists for the proposition that a

concurring in *United States v. Boyce*, criticized the present-sense impression and excited-utterance exceptions sharply, describing the PSI exception as having “neither a theoretical nor an empirical basis” and the EU exception as resting “on no firmer ground than judicial habit.”³¹ He further argued that the law’s categorical approach to hearsay exceptions should be abolished in favor of a case-specific, discretionary approach under the hearsay rule’s residual exception,³² and expanded on his criticisms two years later in a provocative law review essay.³³

These criticisms go to the foundation of the present sense impression and the excited utterance. But—as Parts II and III suggest—the existing scholarship has addressed them only partially, and in ways that leave many of the most consequential questions unanswered. The purpose of this Part is to identify the psychological assumptions that underlie the PSI and EU exceptions with greater precision than the doctrine provides. Doing so can shed light on where the empirical literature should expand from here.

declarant’s event-concurrent statements should evade traditional evidentiary requirements, and thus for this judicially-created ‘excited-utterance’ exception. Science, fact, and common sense suggest that we should cabin, if not outright abandon, the exception.”); *see also* *Lust v. Sealy, Inc.*, 383 F.3d 580, 588 (7th Cir. 2004) (citing research at odds with the assumptions underlying the excited utterance exception and stating, “[a]s with much of the folk psychology of evidence, it is difficult to take this rationale entirely seriously, since people are entirely capable of spontaneous lies in emotional circumstances”).

31. *See* 742 F.3d 792, 801 (7th Cir. 2014) (Posner, J., concurring) (describing the present sense impression exception as having “neither a theoretical nor an empirical basis; and it’s not even common sense—it’s not even good folk psychology”); *id.* at 802 (characterizing the excited utterance exception as resting “on no firmer ground than judicial habit, in turn reflecting judicial incuriosity and reluctance to reconsider ancient dogmas”).

32. *See id.* at 802 (“What I would like to see is Rule 807 (‘Residual Exception’) swallow much of Rules 801 through 806 and thus many of the exclusions from evidence, exceptions to the exclusions, and notes of the Advisory Committee.”).

33. *See* Hon. Richard A. Posner, *On Hearsay*, 84 FORDHAM L. REV. 1465, 1471 (2016) (maintaining that although “Rule 807 should probably be tightened up, some of the hearsay exceptions should be abolished.”). For an article critiquing Judge Posner’s approach in *Boyce*, *see* Liesa L. Richter, *Posnerian Hearsay: Slaying the Discretion Dragon*, 67 FLA. L. REV. 1861, 1907 (2015) (“[T]his Article denounces a reactionary retreat from advances in hearsay doctrine made by the Federal Rules. . . . The Posnerian hearsay model represents . . . a poor design alternative.”). For a discussion of the debate between Judge Posner and Professor Richter and a suggestion for an alternative approach, *see* Justin Sevier, *On Hearsay Dragon-Slaying*, 67 FLA. L. REV. F. 269 (2016).

B. The Present Sense Impression: Timing as Proxy

The PSI exception, the younger of the two in the modern codification era,³⁴ focuses its reliability claim on a single variable: temporal proximity. The theory is that the close relationship between perception and statement creates two indicia of trustworthiness. The first is an anti-fabrication claim insofar as a declarant who describes an event while perceiving it, or immediately afterward, has no time to construct a false account.³⁵ The second is an anti-memory-contamination claim, because the declarant is narrating in real time, or nearly so, the statement is therefore not subject to reconstructive distortion.³⁶

These two claims are logically distinct. The anti-fabrication claim is about sincerity: it assumes that the conditions surrounding a PSI make deliberate deception unlikely or cognitively difficult. The anti-memory-contamination claim concerns accuracy: it assumes that those conditions make perceptual errors and reconstructive distortions less likely. A statement can satisfy one claim while failing the other. A declarant may be entirely sincere while still producing an account that becomes contaminated by rapid inference, schema completion, or perceptual error occurring at the moment of observation.³⁷ Yet the doctrine treats timing as a proxy for both claims simultaneously, as if the speed of a statement is sufficient evidence of both its sincerity and its narrative accuracy.

The “immediately after” language of Rule 803(1) is important doctrinally, and courts have interpreted it liberally.³⁸ In practice,

34. The present-sense impression exception was not universally recognized at common law and was absent from federal practice until the 1975 codification of the Federal Rules of Evidence. See Jeffrey Bellin, *Facebook, Twitter, and the Uncertain Future of Present Sense Impressions*, 160 U. PA. L. REV. 331, 340–41 (2012). The excited utterance traces its lineage to the *res gestae* tradition and was championed by Wigmore as early as the turn of the twentieth century. See 6 JOHN HENRY WIGMORE, EVIDENCE IN TRIALS AT COMMON LAW § 1747 (James H. Chadbourn ed., 1976).

35. See FED. R. EVID. 803 advisory committee’s note to Rule 803(1) (stating that “[t]he underlying theory of [Rule 803(1)] . . . is that substantial contemporaneity of event and statement negative the likelihood of deliberate o[r] conscious misrepresentation.”).

36. See *id.* (implicitly invoking this rationale by emphasizing that the statement is made while perceiving the event but focusing primarily on the anti-fabrication rationale); Edward J. Imwinkelried, *The Need to Resurrect the Present Sense Impression Hearsay Exception: A Relapse in Hearsay Policy*, 52 HOW. L.J. 319, 322 (2009) (“If the declarant made a statement about an event while he or she was describing the event, there could be little concern about the caliber of the declarant’s memory . . .”).

37. For a discussion of these concepts, see *infra* notes 68–69 and accompanying text.

38. FED. R. EVID. 803(1); see, e.g., *United States v. Blakey*, 607 F.2d 779, 786 (7th Cir. 1979) (potentially permitting a twenty-three-minute delay in reporting); see also *Miller v. Crown Amusements, Inc.*, 821 F. Supp. 703, 706 (S.D. Ga. 1993) (allowing an eight-minute delay).

statements made some minutes after a perceived event, with intervening activity or conversation, have been admitted as PSIs on the theory that the declarant “immediately” described what she perceived (or close enough to it).³⁹ This raises two psychological questions, one obvious and the other more subtle: (1) more obviously, how much time can “immediately” describe before the anti-fabrication and anti-contamination claims collapse; and (2) more subtly, even if the statement is uttered immediately, is the speaker’s cognitive machinery in that short interval between perception and narration nonetheless susceptible to predictable bias? Research on attention, dual-task narration, and the speed of inferential processing suggests that even a brief interval may be sufficient for the declarant’s statement to reflect a mixture of perception and interpretation.⁴⁰ That assumption has likely never been directly tested in a PSI-specific paradigm.

C. The Excited Utterance: Arousal as Proxy

The EU exception sets its reliability claim on a different variable: physiological and psychological arousal. Unlike with the PSI exception, the declarant need not speak contemporaneously with the event; she need only speak while still “under the stress of excitement” caused by the startling event.⁴¹ The exception’s broader temporal scope and looser content requirements reflect its different psychological foundation: although the PSI exception relies on speed, the EU exception relies on the stilling effect of excitement on a person’s ability for reflection.

As Professor Steven Baicker-McKee has argued in his analysis of the psychology of the EU exception, the exception actually rests on three bundled hypotheses not just one.⁴² The first—that startling events reliably produce arousal—he notes, is well supported by a century of

39. See FED. R. EVID. 803 advisory committee’s note to Rule 803(1); see also Bellin, *supra* note 34, at 339 n.31 (“The rule allows a ‘slight lapse’ following the event” for flexibility purposes); see also Douglas D. McFarland, *Present Sense Impressions Cannot Live in the Past*, 28 FLA. ST. U. L. REV. 907, 908 (2001) (arguing that many courts have used Rule 803(1)’s “immediately thereafter” language to stretch the temporal gap permitted by the exception beyond the trustworthiness rationale for the rule).

40. See *infra* notes 68–69 and accompanying text.

41. FED. R. EVID. 803 advisory committee’s note to Rule 803(2) (“The theory of [the excited utterance exception] is simply that circumstances may produce a condition of excitement which temporarily stills the capacity of reflection and produces utterances free of conscious fabrication.”). See 6 WIGMORE, *supra* note 34, § 1747 (describing the excited utterance as a “spontaneous and sincere response to the actual sensations and perceptions” of the declarant).

42. See Baicker-McKee, *supra* note 10, at 131–32 (identifying three implicit empirical hypotheses underlying the EU exception).

research on the fight-or-flight response. The second—that this arousal reliably interferes with the ability to deceive—is not supported.⁴³ The third—that arousal neither distorts perception nor impairs narration—is doubly unsupported.⁴⁴ This is the paradox that Baicker-McKee identifies within the excited-utterance exception: the conditions that the doctrine treats as guarantees of reliability are also the conditions most likely to produce sincere but inaccurate statements.⁴⁵

The EU and PSI exceptions both use a circumstantial proxy to stand in for a psychological reality that the proxy may not support. And like the PSI exception, the EU exception collapses two distinct reliability questions into one. The question, “Is the declarant telling the truth as she understands it?” implicates sincerity. The question, “Is the declarant’s understanding of the truth accurate?” implicates veracity in a deeper, more complex sense. Physiological arousal under the EU exception speaks imperfectly to the first question and it provides essentially no guidance on the second.

D. Empirical Assumptions

Both the PSI and EU exceptions treat the circumstances of a statement’s production as a proxy for its reliability, and both define “reliability” primarily in terms of sincerity (that is, the absence of deliberate fabrication). The Advisory Committee notes for both rules focus on conditions that reduce the opportunity or motivation to lie, not on conditions that guarantee the accuracy of what is reported.⁴⁶ This focus made sense given the epistemological landscape of mid-twentieth-century evidence law, which treated lying as the primary threat to hearsay’s utility.⁴⁷ But the psychological literature has since identified

43. See *id.* at 177 (“There is no scientific support for the exception’s fundamental premise that a startling event is likely to render a witness unable to fabricate.”).

44. See *id.* at 132–77 (reviewing the fight-or-flight literature and concluding that the science supports hypothesis one but not hypotheses two and three).

45. See *id.* at 175 (“[W]hen the levels of arousal are so high that an individual is unable to tell even a moderately detailed and plausible lie, his perception is likely so distorted that he makes mistakes in understanding reality and/or in communicating his impressions. But when arousal is low enough to facilitate more accurate perception and effective communication, it is also low enough to reduce the cognitive load that interferes with fabricating a lie.”).

46. See FED. R. EVID. 803 advisory committee’s note to Rule 803 (1)–(2).

47. See Edward J. Imwinkelried, *The Importance of the Memory Factor in Analyzing the Reliability of Hearsay Testimony: A Lesson Slowly Learnt—and Quickly Forgotten*, 41 FLA. L. REV. 215, 230 (1989) (“[T]he common law operated on the assumptions that insincerity was the primary hearsay danger and that defective memory was merely a secondary risk.”); see also Susan E. Provenzano, *Questioning Hearsay’s Formalism*, 76 FLA. L. REV. 913, 986 (2024) (criticizing the Federal Rules of Evidence’s “obsession with

a richer taxonomy of ways in which out-of-court statements can be unreliable, and insincerity accounts for only one of them.

Commentators Steven Baicker-McKee and Timothy Lau have been at the forefront of the contemporary empirical debate surrounding the PSI and the EU hearsay exceptions. Both have reached essentially the same broad conclusion: the existing science does not yield “firm and unequivocal conclusions” about the reliability of these exceptions, and the research base has not been designed with these legal constructs in mind.⁴⁸ The concern that stress might impair accuracy even while reducing deliberate deception is at least as old as the exceptions themselves: Hutchins and Slesinger raised it in 1928, noting that “[w]hat the emotion gains by way of overcoming the desire to lie, it loses by impairing the declarant’s power of observation.”⁴⁹ That proposition has never been fully tested, in part because the debate has been conducted at such a level of generality that it has prevented resolution. Critics and defenders argue over whether “stress” helps or harms reliability, but stress operates at multiple cognitive stages, with implications that may differ at each.

The reliability of a PSI or EU, as it ultimately functions in a legal proceeding, depends on a sequence of cognitive events. The declarant must first accurately perceive the event. She must then narrate it, translating perceptual experience into language under whatever cognitive constraints apply at the moment of speech. The in-court witness who repeats the statement must transmit it accurately, without the compressions, certainty inflations, and lost hedges that can accumulate in any act of retelling.⁵⁰ And the factfinder must ultimately weigh the statement appropriately, without over-crediting the vividness and confidence that arousal and immediacy can produce or underdiscounting the inferential leaps the statement may contain.⁵¹

insincerity in contrast to their relative neglect of witnesses’ memory and perception problems”).

48. See Timothy T. Lau, *Reliability of Present Sense Impression Hearsay Evidence*, 52 GONZ. L. REV. 175, 179, 195 (2017) [hereinafter Lau (PSI)] (“[A] search of the literature yields no research that directly and empirically tests the assumptions underlying the PSI exception. . . . [T]he research findings summarized above do not yield firm and unequivocal conclusions about the resistance of PSI hearsay evidence to the negative effects of fabrication and coaching.”); see also Lau (EU), *supra* note 10, at 602 (reporting a similar conclusion with respect to the EU).

49. Robert M. Hutchins & Donald Slesinger, *Some Observations on the Law of Evidence*, 28 COLUM. L. REV. 432, 439 (1928). The Advisory Committee’s notes to Rule 803(2) acknowledge this criticism but treat it as insufficient to preclude the exception. See FED. R. EVID. 803 advisory committee’s note to Rule 803(2).

50. See *infra* note 99 and accompanying text.

51. See *infra* note 74 and accompanying text.

The reliability rationale for the PSI and EU exceptions collapses this sequence into a single judgment.⁵² Both exceptions treat the circumstances of the statement's production as indicia of reliability across the entire chain. But each link in the chain is a distinct empirical question, and the existing literature, as Part II suggests, has addressed them unevenly. The lion's share of empirical attention has focused on fabrication at the moment of utterance. Almost no attention has been directed to what happens to accuracy at the moment of perception under the specific cognitive demands of PSI and EU contexts, to how statements change in transmission,⁵³ or to how factfinders calibrate their assessments when the admitted evidence carries the marks of excitement or immediacy.⁵⁴

II. THE EMPIRICAL DEBATE

A. The Current Literature

The existing empirical literature on the PSI and EU hearsay exceptions is substantial on its own terms. Across principal contributions to the debate—Baicker-McKee's *The Excited Utterance Paradox*, Lau's paired articles on the reliability of PSI and EU hearsay evidence, and the Federal Judicial Center memorandum that preceded them—the authors collectively identify nearly eighty major empirical sources.⁵⁵ Still, the literature is narrower than the source count

52. The law's tendency to treat fabrication and accuracy as a unified reliability question is clear in the structure of both rules. The EU exception does not separately require that the declarant's perception have been accurate—only that the statement have been made under the stress of excitement. See Lau (EU), *supra* note 10, at 600–02. Similarly, the PSI exception does not require corroboration of the declarant's account. Both treat the circumstantial conditions of the statement as proxies for the reliability. See FED. R. EVID. 803(1)–(2); FED. R. EVID. 803 advisory committee's note to Rules 803(1)–(2); see also Baicker-McKee, *supra* note 10, at 153–76. See generally Lau (PSI), *supra* note 48, at 176–78, 204; Lau (EU), *supra* note 10, at 600–02.

53. The “fidelity” stage, the transmission of a PSI or EU from declarant to in-court witness, is addressed in passing by Baicker-McKee but not developed into a systematic empirical inquiry. See Baicker-McKee, *supra* note 10, at 171–75. This Article identifies fidelity as a distinct stage of error with its own research questions and designs.

54. The “calibration” stage, the factfinder's weighting of admitted PSI-EU evidence, has been addressed in the broader hearsay literature but not in the PSI-EU context specifically. See *infra* note 74 and accompanying text.

55. The literature addressed here consists of the empirical works cited in the major contributions to the PSI-EU scientific debate: Baicker-McKee, *supra* note 10; Lau (PSI), *supra* note 48; Lau (EU), *supra* note 10; and the Federal Judicial Center memorandum. See generally TIMOTHY T. LAU, FED. JUD. CTR., MEMORANDUM: RELIABILITY OF PRESENT SENSE IMPRESSIONS AND EXCITED UTTERANCES (2016) [hereinafter LAU, FJC Memo]. A gap map was constructed by cataloging the empirical sources cited across these works and classifying each by the stage of reliability it addresses and by methodological setting.

suggests. To demonstrate this point, articles constituting the empirical literature were classified by the stage of reliability they address, using a scheme that distinguishes among the declarant's sincerity, the accuracy of the declarant's perception, the fidelity with which the statement is transmitted from declarant to in-court witness, and the calibration of the factfinder who ultimately evaluates it.

The evidentiary gap map below reveals a notable asymmetry.⁵⁶ All of the seventy-seven sources cataloged in Figure 1 below address some aspect of what this Article calls the veracity stage—the declarant's cognitive state at or near the moment of utterance. Thirty-six concern sincerity and fabrication: whether people can lie spontaneously, how cognitive load affects deception, and what the neuroscience of lying reveals about the capacity for deliberate deception.⁵⁷ Nineteen address stress and memory, twelve address perception and attention, and eight concern memory distortion and confabulation—the possibility that a sincere declarant produces a false account through the ordinary fallibility of perception and recall under stress.⁵⁸ Two additional sources address communication or narration under stress. No mapped source directly addresses fidelity—what happens to the statement between the declarant's lips and the courtroom—and no mapped source addresses calibration in the way this Article targets—how factfinders weigh PSI and EU evidence once admitted. Thus, the debate has been conducted almost entirely within a single link of a longer chain.⁵⁹

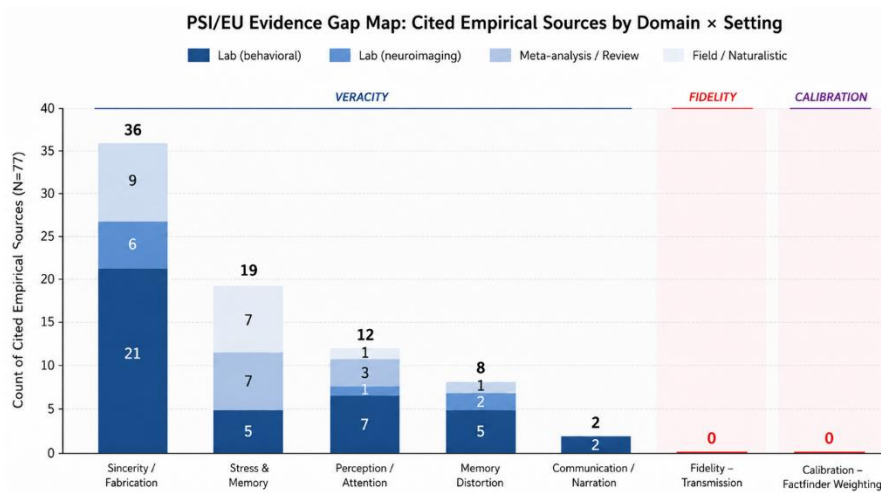
56. See Figure 1 on page 1202. A gap map (also called an 'evidence gap map' or EGM) is a systematic visual display of the available empirical evidence relevant to a specified research question, organized by domain or stage. It is used to identify where evidence is robust and where gaps persist. See, e.g., Birte Snilstveit et al., *Evidence & Gap Maps: A Tool for Promoting Evidence Informed Policy and Strategic Research Agendas*, 79 J. CLIN. EPIDEMIOLOGY. 120, 120–21 (2016); Howard White et al., *Guidance for Producing a Campbell Evidence and Gap Map*, 16 CAMPBELL SYSTEMATIC REV. 1, 2–3 (2020). The map used here adapts this method for a more limited purpose: it does not purport to be a full review of all psychological evidence bearing on PSI and EU reliability, but instead classifies the empirical and scientific sources cited in the principal PSI-EU reliability debate by the stage of the reliability chain they address.

57. See, e.g., Iris Blandón-Gitlin et al., *Cognitive-Load Approaches to Detect Deception: Searching for Cognitive Mechanisms*, 18 TRENDS COGNITIVE SCIS. 441, 442–43 (2014); ALBERT VRIJ, DETECTING LIES AND DECEIT: PITFALLS AND OPPORTUNITIES 365–72 (2d ed. 2008). An important addition to this discussion was published recently and is therefore not included in the gap map. See generally Chris Sundby, *The Neuroscience of Evidentiary Rules: The Case of the Present Sense Impression*, 78 SMU L. REV. 827 (2025) (applying neuroscience methods).

58. See, e.g., Charles A. Morgan III et al., *Accuracy of Eyewitness Memory for Persons Encountered During Exposure to Highly Intense Stress*, 27 INT'L J. L. & PSYCHIATRY 265, 266–77 (2004); Deffenbacher et al., *supra* note 21, at 691–704.

59. This is not a criticism of these scholars' selection of sources. The existing empirical base on lying, stress, and memory is relevant to the questions that the PSI and

Figure 1. PSI-EU Evidence Gap Map: Cited Empirical Sources by Domain and Methodological Setting⁶⁰



Note: Distribution of seventy-seven empirical sources cited across the principal PSI-EU reliability debate, classified by stage of reliability addressed and methodological setting.

This imbalance is understandable. Because both exceptions justify admissibility by referring to the declarant's cognitive state at the moment of utterance, scholars who evaluate those exceptions have naturally gravitated toward the psychological literature most relevant to that moment. And because the exceptions' primary claims are that the declarant's circumstances reduce the likelihood of fabrication, the question "Can people lie under these conditions?" became the debate's organizing principle. The sincerity question is, of course, important. But

EU exceptions pose, but it often supplies general guideposts instead of specific answers. The cited experimental paradigms were designed to test whether lying is cognitively demanding and whether stress and emotion affect perception, but in contexts that differ materially from the situation of a witness producing a spontaneous account in the immediate aftermath of a startling event and with stress manipulations that rarely approach the intensity or cognitive load that PSI and EU conditions create.

60. As a sensitivity analysis, the direct relevance core was coded more narrowly: sources that bear directly on the cognitive mechanisms that the PSI and EU exceptions invoke, rather than on general principles of memory, stress, deception, or cognition. See Julian P.T. Higgins et al., *Chapter 10: Analyzing Data and Undertaking Meta-Analyses*, in COCHRANE HANDBOOK FOR SYSTEMATIC REVIEWS OF INTERVENTIONS § 10.14 (Julian P.T. Higgins et al. eds., 2024) (describing sensitivity analyses as a way to assess the robustness of results to assumptions, borderline decisions, and inclusion criteria). That core contains fifty-three sources and follows the same distributional pattern: twenty sincerity and fabrication, twelve stress and memory, eleven perception and attention, eight memory distortion and confabulation, and two communication and narration. No source in the core directly addresses fidelity. The coding matrix is on file with the author.

the literature on this question, when carefully read, may not provide the level of support the exceptions need. Instead, it supplies findings that are individually suggestive but, in my view, collectively ambiguous.

On the anti-fabrication side, research on the cognitive demands of lying consistently shows that deception is a more taxing cognitive task than truth-telling, as it takes longer, recruits more neural resources, and degrades under conditions of high cognitive load.⁶¹ The problem, as Baicker-McKee documents extensively, is that the inference from these findings to the exceptions' reliability claims breaks down under scrutiny.⁶² The response-latency differences between truth-telling and lying are measured in milliseconds and are undetectable to ordinary observers.⁶³ People do lie under stress, sometimes spontaneously, and sometimes, with strategic intent.⁶⁴

On the accuracy side, the news is more troubling. Research on the fight-or-flight response consistently shows that arousal affects perception and memory in ways that can lead to sincere but inaccurate statements.⁶⁵ For example, attention narrows to salient features, threat assessments are amplified, confidence can substantially outpace accuracy, and encoding under extreme stress can produce memories that feel vivid and certain while being materially wrong.⁶⁶ The declarant who reports her genuine experience while excited is not necessarily reporting what actually happened.⁶⁷

61. See, e.g., Baicker-McKee, *supra* note 10, at 158–59, 162–65; see also Lau (EU), *supra* note 10, at 611–15, 617–21.

62. See Baicker-McKee, *supra* note 10, at 162–63.

63. See Baicker-McKee, *supra* note 10, at 162–63 (stating that response-latency differences are “typically measured in thousandths of a second”); see also Jeffrey J. Walczyk et al., *Advancing Lie Detection by Inducing Cognitive Load on Liars: A Review of Relevant Theories and Techniques Guided by Lessons from Polygraph-Based Approaches*, 4 FRONTIERS PSYCH. 1, 2–3 (2013) (discussing cognitive load theory more broadly); Giorgi Ganis et al., *Neural Correlates of Different Types of Deception*, 13 CEREBRAL CORTEX 830 (2003).

64. See Baicker-McKee, *supra* note 10, at 162, 163 (concluding that even under stress, “subjects were still able to lie, even if it took them longer,” and that response-latency differences are “measured in thousandths of a second”—undetectable to ordinary observers).

65. See, e.g., Deffenbacher et al., *supra* note 21, at 699–702.

66. See Lau (EU), *supra* note 10, at 627–29; Baicker-McKee, *supra* note 10, at 168–71 (cataloging perceptual errors including closure, weapon focus, attribution error, and cognitive dissonance).

67. See Baicker-McKee, *supra* note 10, at 143–45 (discussing flashbulb memories and the confidence-accuracy dissociation). The weapon focus effect illustrates the point as well: a bystander witnessing an armed robbery may have sharp recall of the gun while being unable to describe the perpetrator's face. See Jonathan M. Fawcett et al., *Of Guns and Geese: A Meta-Analytic Review of the “Weapon Focus” Literature*, 19 PSYCH., CRIME & L. 35, 36–37 (2013).

For PSI evidence, the accuracy concern takes a different form. The PSI exception's anti-memory-contamination guarantee assumes that real-time narration is closer to pure perception than retrospective testimony. But cognitive research on divided attention and concurrent task performance suggests that producing language while perceiving an event imposes additional cognitive demands that may affect perceptual processing and encoding.⁶⁸ A declarant who describes a scene as it unfolds is externalizing a percept while constructing a narrative in real time and under the dual demands of perception and speech.⁶⁹

These concerns about accuracy at the veracity stage may be underappreciated, but they are at least present in the scholarly record. What is almost entirely absent is any treatment of the two remaining stages at which reliability can degrade.⁷⁰ The fidelity stage is the gap between the declarant's statement and the in-court witness's repetition of it. As early as 1928, Hutchins and Slesinger stressed the importance of the in-court witness, whose testimony about the declarant's statement and surrounding circumstances could be tested by cross-examination.⁷¹ Baicker-McKee's article devotes a section to what he calls "miscommunications," identifying ways that the process could break down: specifically, the speaker's statement can be compressed and distorted, and the recipient's retelling can further distort it.⁷² But no author has yet developed this concern into a systematic empirical inquiry. A small verbal hedge ("I think he grabbed the bag") that appeared in the declarant's original statement might not appear in the witness's retelling ("she said he grabbed the bag"), which can transform a tentative observation into an asserted fact. The exception's

68. See Alan Baddeley, *Working Memory: Looking Back and Looking Forward*, 4 NATURE REV. NEUROSCIENCE 829, 830, 835 (2003) (discussing limited attentional capacity and simultaneous-task paradigms used to study divided attention); Lau (PSI), *supra* note 48, at 203–04 (reviewing research concerning attention and perceptual accuracy).

69. See Christian A. Meissner & John C. Brigham, *Thirty Years of Investigating the Own-Race Bias in Memory for Faces*, 7 PSYCH., PUB. POL'Y & L. 3, 10–11 (2001) (discussing "verbal overshadowing" in the context of eyewitness identification, in which "generating a verbal description of a face significantly impairs subsequent identification accuracy"). The tendency to organize ambiguous motion displays into coherent social narratives (what would now be called "schema completion") was classically illustrated by Fritz Heider & Marianne Simmel, *An Experimental Study of Apparent Behavior*, 57 AM. J. PSYCH. 243, 259 (1944) (illustrating observers' tendency to attribute motive, personality, and causal relations to ambiguous geometric motion).

70. See *supra* Part II.A.

71. Hutchins & Slesinger, *supra* note 49, at 439.

72. See Baicker-McKee, *supra* note 10, at 171–76. The communication psychology of retelling—hedge loss, certainty inflation, source confusion, temporal drift—has been documented in the broader eyewitness and narrative literature but not studied in the specific PSI-EU context.

admissibility rationale protects only the original statement; the transmission to court may be largely unprotected.⁷³

The calibration stage is the factfinder's problem, insofar as it requires juries to evaluate both the out-of-court statement and the in-court retelling. The broader hearsay literature—including prior work examining how mock jurors perceive hearsay exceptions across reliability and fairness dimensions—establishes that factfinders apply a general hearsay discount relative to live testimony, but it is not obvious that this discount reliably tracks fine-grained doctrinal distinctions within the hearsay categories.⁷⁴ For PSI and EU evidence specifically, there are at least three calibration concerns the existing debate does not address: (1) the possibility that the cues signaling admissibility function as amplifiers of credibility rather than as discounting cues; (2) the risk that simultaneous presence of “immediacy” and “excitement” leads factfinders to double-count these features either consciously or subconsciously; and (3) the risk that the judicial decision to admit carries an implicit credibility endorsement—an “admissibility

73. See *supra* note 99 and accompanying text.

74. See Justin Sevier, *Testing Tribe's Triangle: Juries, Hearsay, and Psychological Distance*, 103 GEO. L.J. 879, 914 (2015) [hereinafter Sevier, *Testing Tribe's Triangle*] (finding that mock jurors discount hearsay relative to live testimony); Justin Sevier, *Popularizing Hearsay*, 104 GEO. L.J. 643, 652–53, 678, 688 (2016) [hereinafter Sevier, *Popularizing Hearsay*] (finding that lay assessments of hearsay reflect both accuracy and procedural-fairness judgments); Justin Sevier, *Evidence-Based Hearsay*, 76 VAND. L. REV. 1799, 1810–23 (2023) [hereinafter Sevier, *Evidence-Based Hearsay*] (replicating the hearsay discount and fairness-reliability asymmetry across multiple studies). None of these studies examined PSI and EU evidence specifically. The question of whether the cues associated with PSI-EU admissibility—arousal, immediacy—alter factfinder weighting beyond the general hearsay discount has not been studied directly. A line of mock-juror studies has examined factfinder responses to excited utterance hearsay testimony in sexual assault and elder abuse contexts. See Jonathan M. Golding, Rebecca Polley Sanchez & Sandra A. Sego, *The Believability of Hearsay Testimony in a Child Sexual Assault Trial*, 21 L. & HUM. BEHAV. 299, 318–21 (1997); Jonathan M. Golding, Mary C. Alexander & Terri L. Stewart, *The Effect of Hearsay Witness Age in a Child Sexual Assault Trial*, 5 PSYCH., PUB. POL'Y & L. 420, 433–34 (1999); David F. Ross, R.C.L. Lindsay & Dorothy F. Marsil, *The Impact of Hearsay Testimony on Conviction Rates in Trials of Child Sexual Abuse: Toward Balancing the Rights of Defendants and Child Witnesses*, 5 PSYCH., PUB. POL'Y & L. 439, 448–54 (1999); Emily E. Dunlap et al., *Perceptions of Elder Physical Abuse in the Courtroom: The Influence of Hearsay Witness Testimony*, 19 J. ELDER ABUSE & NEGLECT 19, 35 (2007); Dylan R. Baker, Jonathan M. Golding & Mary M. Levi, *Juror Perceptions of Excited Utterance Hearsay Testimony in an Adult Sexual Assault Case*, 29 PSYCH., CRIME & L. 303, 314 (2023). These studies consistently find that EU hearsay can be as persuasive as live victim testimony. They do not, however, test whether factfinders calibrate their reliance to the actual reliability of the admitted statement, whether factfinders distinguish among EU predicates on doctrinal grounds, or how factfinders respond to PSI hearsay as distinct from EU hearsay.

laundering” effect—that inflates the weight factfinders assign to the evidence.⁷⁵

B. Opposing Conclusions

The asymmetric distribution of the empirical PSI and EU literature may explain why two sophisticated scholars have read substantially the same evidence and reached opposing conclusions. Baicker-McKee concludes that the EU exception is “both broken and irreparable” and should be abolished.⁷⁶ Lau concludes that the research is “not conclusive” but provides “some support” for the reliability premises of the exceptions and recommends against changes pending further study.⁷⁷ The Federal Judicial Center’s Advisory Committee accepted Lau’s more cautious framing and left the rules unchanged.⁷⁸

This disagreement may reflect a dispute about the appropriate default rule in the face of empirical uncertainty: should a legal exception be presumed valid until clearly disproven, or should it be required to justify itself affirmatively before imposing costs? Baicker-McKee’s abolitionism follows from treating exceptions as presumptively unjustified in the absence of empirical support. Lau’s conservatism follows from treating them as presumptively valid unless the empirical evidence against them is clear and compelling.⁷⁹

There is, however, a more specific explanation for the impasse. Both Baicker-McKee and Lau are arguing about veracity: whether the declarant’s statement, as produced, is likely to be accurate and sincere. Both share the doctrine’s implicit assumption that if the veracity stage is satisfied, the reliability of the evidence is satisfied.⁸⁰ That assumption,

75. The “admissibility laundering” concern has been noted in the evidence literature but not tested directly for PSI-EU evidence. See Nicholas J. Schweitzer & Michael J. Saks, *The Gatekeeper Effect: The Impact of Judges’ Admissibility Decisions on the Persuasiveness of Expert Testimony*, 15 PSYCH., PUB. POL’Y & L. 1, 11–13 (2009) (finding that admitting low-quality expert evidence can imbue it with undeserved credibility).

76. Baicker-McKee, *supra* note 10, at 111 (concluding that “the excited utterance exception is both broken and irreparable,” and recommending “abandoning the excited utterance exception altogether”).

77. Lau (PSI), *supra* note 48, at 206; Lau (EU), *supra* note 10, at 639–40.

78. The FJC Memo recommended against abolition pending further research. The Advisory Committee accepted that recommendation, and the rules remain unchanged. See LAU, FJC Memo, *supra* note 55, at 24.

79. This asymmetry tracks a methodological difference: Baicker-McKee treats the absence of direct support as sufficient reason for abolition, whereas Lau treats the same absence as warranting caution rather than elimination. Neither position is empirically determined.

80. The framing of the debate as “skeptics versus defenders” slightly oversimplifies a more nuanced landscape. Both scholars acknowledge the exceptions’ empirical

however, is in part what the gap map questions. Both scholars debate whether the reliability guarantees for the exceptions actually yield reliable declarant statements.⁸¹ Most importantly, neither is asking whether reliable declarant statements actually produce reliable evidence at trial. A PSI or EU that accurately reflects the declarant's perception can still be materially distorted by the time it reaches the factfinder, through other veracity failures not considered by Baicker-McKee and Lau, fidelity failures, calibration failures, or some combination of these failures.

The impasse between Baicker-McKee and Lau, in this sense, is an artifact of the literature's distribution; both scholars are doing important work within the existing research framework, but the framework itself is too narrow. The debate has been asking whether the Advisory Committee's expressed assumptions underlying the PSI and EU exceptions are justified. The more consequential question—whether those assumptions, justified or not, are sufficient to produce reliable evidence across the full chain from utterance to verdict—has not been asked at all.⁸²

Moreover, Lau identifies an additional important issue in the empirical debate over the justification for the PSI and EU exceptions. He discusses what might be termed the “laboratory-translation” problem, involving the difficulty in drawing firm inferences from the general psychology literature to the specific conditions under which PSI and EU evidence appear in actual proceedings.⁸³ Part III will argue that the correct response to Lau is to refrain from treating the existing literature as the (insufficient) last word with respect to the PSI and EU exceptions. Instead, behavioral scientists should design research that closes the gap. Those future research designs should therefore be aimed at the specific

weaknesses; they disagree about what to do with incomplete evidence. *See generally* Baicker-McKee, *supra* note 10; Lau (EU), *supra* note 10; Lau (PSI) *supra* note 48; LAU, FJC memo, *supra* note 55.

81. And even in this empirical space, their discussion is not quite complete, as this Article will discuss in Part III.

82. *See infra* Part V.D.

83. The laboratory-translation problem is familiar from other psycholegal literatures. *See* Gary L. Wells & Elizabeth A. Olson, *Eyewitness Testimony*, 54 ANN. REV. PSYCH. 277, 279, 290–91 (2003) (distinguishing estimator variables from system variables and noting that laboratory findings must be translated carefully to field conditions). The PSI-specific translation problem is particularly acute because the cognitive task most studied in the lying literature—deciding whether to make a false assertion in response to a direct question—does not resemble the cognitive task most relevant to PSI doctrine: producing a spontaneous, real-time narrative while simultaneously perceiving the event. *See* Lau (PSI), *supra* note 48, at 183–200.

cognitive demands those contexts create.⁸⁴ The three pilot studies discussed below offer a first step.

III. HEARSAY ERROR SIGNATURES

The gap map reported in this Article suggests that the existing empirical debate has concentrated almost exclusively on a single portion of a single stage of the PSI-EU reliability chain. The three pilot studies discussed in this Part begin to fill in the remaining gaps in that single stage and to investigate the other stages. They aim to demonstrate that those questions are testable and that the error patterns they predict are sufficiently important to warrant further investigation.⁸⁵ Each pilot targets a distinct stage: the first addresses veracity, the second addresses fidelity, and the third addresses calibration.

A. Pilot Study: Veracity

The first pilot study targets the veracity stage, but from an angle the existing literature has not taken. Where prior research has asked whether declarants can lie under PSI-EU conditions, the pilot asks what declarants actually say: what kinds of content they produce, how timing and arousal affect that content, and how quickly narration may shift from perceptual description into interpretation. The design involves three related paradigms, each addressing a different aspect of the veracity stage.

84. This is one of the gaps in the existing debate: a theoretical account of what cognitive processes are actually occurring when a declarant produces a PSI or EU, specified at a level of granularity sufficient to generate testable predictions. *See supra* Part II.A.

85. The three paradigms are pilot studies: they were designed to establish measurement feasibility, refine coding and stimulus materials, and generate preliminary estimates for future research. Results should be treated as directional, subject to replication with larger samples and more ecologically valid materials. *See* Andrew C. Leon, Lori L. Davis & Helena C. Kraemer, *The Role and Interpretation of Pilot Studies in Clinical Research*, 45 J. PSYCHIATRIC RSCH. 626, 626–29 (2011). The study materials, coding schemes, and preliminary statistics are available from the author upon request.

Figure 2. Representative Stills from Pilot Study 1 Stimulus Clips⁸⁶

Note: Study 1 used three short video clips to examine the veracity stage across three distinct paradigm conditions. Part A (PSI timing paradigm) depicts a person reaching through an open car window toward a bag on the passenger seat: a scene susceptible to both innocent and guilty interpretations, enabling coding of perceptual description versus interpretive elaboration. Timing of narration (immediate versus brief delay) was the main manipulation. Part B (EU arousal paradigm) depicts an elephant on a paved road with a truck approaching, selected to generate reliable variation in self-reported arousal without exceeding Institutional Review Board thresholds for permissible stress induction. Self-reported arousal was related to the accuracy and inferential content of participants' narratives. Part C (ambiguity paradigm) depicts two people in a tense interpersonal interaction; a matched neutral-interaction clip served as the comparison condition. The manipulation tested whether scene ambiguity drives inferential elaboration independent of timing and arousal.

The first paradigm targets the rationale for the PSI exception's timing. Participants watched a short video clip depicting a scene that could be interpreted as either innocent or guilty and were instructed to describe what they observed, as if reporting to a friend in real time over the phone.⁸⁷ Timing was the central manipulation. Half of the participants described the clip immediately, simulating the PSI's contemporaneous-

86. Part A (bag-from-car clip) serves the PSI timing paradigm: participants narrated the event either immediately or after a brief delay with an intervening distractor task. Part B (elephant clip) serves the EU arousal paradigm: the clip was selected to generate reliable variation in self-reported arousal without exceeding Institutional Review Board thresholds required for more ecologically valid EU stimuli. Part C (argument/chat clips) serves the ambiguity paradigm: participants in the conflict condition saw the scene depicted here; participants in the neutral condition saw a matched non-conflict interaction. All images are stock video stills.

87. Simultaneous-task paradigms are commonly used to study the limits of divided attention in the working-memory literature. See Baddeley, *supra* note 68, at 830, 835. The more specific question whether producing a real-time verbal narrative while simultaneously perceiving an unfolding event affects perception, coding, or narration has received little direct study in the PSI context. See Lau (PSI), *supra* note 48, at 179, 203–04 (noting the absence of research directly testing the assumptions underlying the PSI exception and reviewing research on attention and perceptual accuracy).

narration requirement; the other half experienced a brief delay with an intervening distractor task, simulating what happens arguably beyond the “immediately thereafter” threshold.⁸⁸ Participants’ open-ended narratives were coded along two dimensions: the proportion of perceptual description versus interpretive elaboration, and the accuracy of specifically verifiable claims relative to what the video actually depicted.⁸⁹

Several preliminary findings emerged. First, across both timing conditions, participants’ narratives contained substantial inferential content—more than the PSI exception’s “describing or explaining an event” language contemplates.⁹⁰ Participants honestly reported what they saw, but they also generated quick causal narratives, attributed motive to the actors, and used normative labels that went beyond the observable facts. This appeared in the immediate condition as well as the delayed condition, which suggests that the inference problem is not a function of delay but of narration itself.⁹¹

Second, the timing manipulation produced an asymmetry. Participants who immediately narrated showed evidence of a dual-task cost: their accounts were somewhat compressed, with fewer details overall. Delayed participants produced richer narratives in terms of total detail—but also showed a higher proportion of inferential content, consistent with the brief interval having permitted schema-completion processes to operate. Neither timing condition produced narratives that

88. The “phone call” framing approximates the conversational context of many PSI statements, including the 911 call scenario most common in PSI evidence. *See* Lau (PSI), *supra* note 48, at 179–82. The framing makes the narration task naturalistically motivated without requiring participants to be directly involved in the event.

89. The content coding scheme distinguished perceptual descriptions (directly observable features) from interpretive elaborations (attributed intention, inferred relationship, normative labels). This maps onto the PSI doctrine’s implicit assumption, that declarants ‘describe or explain’ perceived events, without addressing whether that description reflects pure perception or inference-laden interpretation. *See* FED. R. EVID. 803(1).

90. The bag-theft stimulus was deliberately ambiguous: capable of innocent and guilty interpretations. This enabled the experimental design to reveal whether participants stayed perceptual (“a man took a bag”) or went interpretive (“a man stole a bag”). The distinction is important because the rule arguably admits the statement without asking whether its content reflects perception or inference. *See* FED. R. EVID. 803(1).

91. Because these pilot studies serve as proof of concept for a new research paradigm, specific statistical analyses are omitted here but are on file with the author.

were purely perceptual. The doctrine's preference for immediacy may therefore reduce one form of error while leaving another undisturbed.⁹²

Arousal is the subject of the second paradigm. Participants watched one of two short video clips varied in their capacity to elicit self-reported arousal, described what they observed using the same phone call framing, and completed a self-report arousal measure.⁹³ The clip stimuli were deliberately modest in this pilot, with ecologically limited materials, but they nonetheless generated reliable individual variation in reported arousal, enabling within-condition analysis of the arousal-accuracy relationship.

Preliminarily, the results are consistent with the concern that the EU exception's reliability rationale gets the psychology backward, at least in part. Participants reporting higher levels of arousal produced narratives with a somewhat higher proportion of inferential elaboration and showed a larger gap between their expressed certainty and the verifiable accuracy of their accounts.⁹⁴ Thus, the EU exception treats the markers of arousal as reliability signals, but the preliminary data suggest that they may, at least in part, function as accuracy-degrading signals.

A third paradigm takes up something that the existing literature has almost entirely overlooked: the effect of social ambiguity on narrative content. Participants watched one of two clips depicting an interpersonal interaction—one framed as a potential conflict, the other as a clearly neutral exchange—and produced narratives using the same methodology.⁹⁵

The preliminary finding is that ambiguity affects the type of content participants produce more than the amount. Participants who watched the conflict-framed clip were more likely to attribute mental states to the actors, make relationship inferences, and use threat-coded

92. The inferential-elaboration finding is consistent with research on "verbal overshadowing" and its inverse: the conditions under which narrating an event shapes the narrator's perception of it. See Meissner & Brigham, *supra* note 69, at 10–11.

93. The design treats arousal as a continuous variable rather than a threshold admissibility condition.

94. The confidence-accuracy dissociation in high-arousal conditions is consistent with the flashbulb memory literature. See Andrew R.A. Conway et al., *Flashbulb Memory for 11 September 2001*, 23 APPLIED COGNITIVE PSYCH. 605, 605–06 (2009); Jennifer M. Talarico & David C. Rubin, *Confidence, Not Consistency, Characterizes Flashbulb Memories*, 14 PSYCH. SCI. 455, 459–60 (2003). For EU doctrine, this dissociation means the visible markers of excitement that courts use to establish admissibility may amplify factfinders' accuracy assessments without corresponding evidentiary support.

95. The argument-chat paradigm was designed to test whether participants 'describe or explain' in the PSI sense, or whether they rapidly generate interpretive narratives.

language—even when the clip depicted nothing that definitively established any of these characterizations.⁹⁶ The PSI and EU exceptions do not, in practice, neatly distinguish between a statement that describes what was perceived and one that describes what the declarant concluded about it. In many ambiguous social scenes that give rise to PSI and EU evidence, that distinction may be the most important one.

B. Pilot Study: Fidelity

The second pilot study shifts the perspective from the declarant to the transmitter. Where the first pilot asked what people say when they produce a PSI or EU, this pilot asks what happens to that statement when someone else repeats it. The design inverts the first pilot's protocol: here, participants receive a short audio recording of a statement describing an observed event—framed as a friend's call—and are subsequently asked to report what they were told, as if repeating it to another party.⁹⁷

Here, the manipulation is affective framing. In one condition, the statement is delivered in a calm, matter-of-fact register; in the other, it is delivered with audible emotional charge, including a faster pace, elevated affect, and vocal markers that courts would likely treat as evidence of the “stress of excitement” that the EU requires. The substantive content of the two recordings, however, is held constant. The question is whether the affective framing of the original statement, the central feature that triggers EU admissibility, affects what the recipient reports having heard.⁹⁸

Participants' retelling narratives were coded for several factors, including hedge loss (the disappearance of uncertainty markers present in the original), certainty inflation (the hardening of tentative claims into unqualified assertions), compression (the omission of details

96. The mental-state attribution and relationship-inference findings are consistent with research on social perception. See Fritz Heider & Marianne Simmel, *An Experimental Study of Apparent Behavior*, 57 AM. J. PSYCH. 243, 258–59 (1944); BERTRAM F. MALLE, HOW THE MIND EXPLAINS BEHAVIOR 1–4 (2004).

97. The fidelity paradigm is modeled on the “telephone” paradigm in serial reproduction research. See FREDERIC C. BARTLETT, REMEMBERING: A STUDY IN EXPERIMENTAL AND SOCIAL PSYCHOLOGY 118–120 (1932) (describing an experiment involving serial reproduction of information). It is adapted to the specific legal question of a single retelling from declarant to in-court witness.

98. This manipulation isolates whether emotional delivery affects transmission fidelity. Cf. Shinobu Kitayama, *Remembrance of Emotional Speech: Improvement and Impairment of Incidental Verbal Memory by Emotional Voice*, 32 J. EXPERIMENTAL SOC. PSYCH. 289, 289–90, 304–05 (1996); see also Annett Schirmer, *Mark My Words: Tone of Voice Changes Affective Word Representations in Memory*, 5 PLOS ONE 1, 6–8 (2010).

present in the original), and interpretive addition (the introduction of content absent from the original).⁹⁹

As one would expect from a pilot with a limited sample, the effect was directional and weak. Participants who heard the emotionally charged version showed somewhat greater hedge loss and certainty inflation in their retellings than participants who heard the matter-of-fact version. For example, in the latter condition, a statement to the listener that “I think he hit her” was more likely to be reported by the listener as “she said he hit her.”¹⁰⁰ The emotional intensity that makes a statement admissible as an EU may simultaneously make its transmission less faithful to what was actually said. If the effect replicates in larger studies, this suggests an empirical irony embedded in the exception’s design.

C. Pilot Study: Calibration

The final pilot study moves to the end of the chain: the factfinder. Participants read a short trial vignette describing a bar altercation in which the defendant claimed self-defense and received evidence that varied (1) in form (live testimony versus a PSI or EU repeated by a second witness) and (2) in how clearly it fit within the relevant exception. A second manipulation varied the substantive content of the hearsay statement from accusatory to a more neutral tone.¹⁰¹ Participants then completed measures of guilt, perceived reliability of the evidence, perceived fairness of its use, and satisfaction with the trial.

Two preliminary findings are worth reporting. The first is the expected hearsay discount: participants who received the critical

99. The coding categories draw on several adjacent literatures. On hedges and uncertainty markers, see George Lakoff, *Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts*, 2 J. PHIL. LOGIC 458, 471–83 (1973). On the tendency of hedged details to drop out in retelling, see Kris Liu & Jean E. Fox Tree, *Hedges Enhance Memory but Inhibit Retelling*, 19 PSYCHONOMIC BULL. & REV. 892, 895 (2012). On compression, rationalization, and addition in retellings, see, for example, GORDON W. ALLPORT & LEO POSTMAN, *THE PSYCH. OF RUMOR* 75–98 (1947) for a discussion of experimental results with respect to leveling and sharpening in the rumor context. For an evaluation of motivational factors affecting narrative retellings, see generally Nicole M. Dudukovic, Elizabeth J. Marsh & Barbara Tversky, *Telling a Story or Telling It Straight: The Effects of Entertaining Versus Accurate Retellings on Memory*, 18 APPLIED COGNITIVE PSYCH. 125 (2004).

100. This matters because cross-examination of the in-court witness may not recover uncertainty markers the witness no longer remembers.

101. The calibration paradigm used a bar-fight vignette in which the defendant claimed self-defense and the critical evidence was a bystander statement presented as live testimony or as a PSI-EU. The design follows methodology established in the empirical hearsay literature. See Sevier, *Testing Tribe’s Triangle*, *supra* note 74, at 904–10. The PSI- and EU-specific manipulations are new.

evidence as a PSI or EU assigned lower guilt ratings and reported lower perceived reliability than participants who received the same substantive information as live testimony.¹⁰² Factfinders are not naïve consumers of admitted hearsay.

The second finding is novel and more consequential for PSI and EU doctrine. Within the hearsay conditions, participants did not reliably distinguish between “clearly within” and “borderline” PSI or EU evidence. The elaborate multi-factor tests that courts apply did not translate into meaningfully different factfinder behavior.¹⁰³ The hearsay discount was applied with a blunt instrument: hearsay was treated as less than live testimony, but finer distinctions within hearsay were not tracked.¹⁰⁴

A parallel mediation analysis offers a preliminary account of why participants were less satisfied when hearsay was used at trial. Both perceived reliability and fairness declined when PSI or EU evidence was offered rather than live testimony, and both predicted a decline in satisfaction. But the fairness pathway accounted for a larger share of the decline in satisfaction than the reliability pathway.¹⁰⁵ Participants were less confident that the hearsay was accurate. They also felt that using it was, in some normative sense, not quite right. This is consistent with the broader finding in the empirical hearsay literature that lay attitudes toward hearsay reflect procedural-justice concerns about confrontation and participation, not only epistemic concerns about reliability.¹⁰⁶

Taken together, the pilots suggest, at least preliminarily, that each stage of the PSI-EU reliability chain is measurable, and that the predicted error patterns are detectable even with simple pilot stimuli. Veracity is not simply a function of timing or arousal. It also appears to be compromised by the inferential demands of narration itself.

102. This finding replicates and extends the pattern in prior empirical hearsay work. See Sevier, *Testing Tribe’s Triangle*, *supra* note 74, at 920.

103. The within-exception insensitivity—participants’ failure to track distinctions between “clearly within” and “borderline” PSI-EU evidence—is the more significant finding for this Article’s purposes.

104. The absence of evidence is not necessarily evidence of absence, of course. Replications of this paradigm with large samples and greater statistical power are necessary going forward.

105. The fairness pathway finding is consistent with the procedural-justice literature on hearsay. See Sevier, *Popularizing Hearsay*, *supra* note 74, at 664–76; Sevier, *Evidence-Based Hearsay*, *supra* note 74, at 1815–22. For PSI and EU evidence specifically, the fairness concern may be salient because both exceptions involve statements made in the absence of any opportunity for the opposing party to probe the declarant’s perception, sincerity, or narration.

106. See Sevier, *Evidence-Based Hearsay*, *supra* note 74, at 1815–20.

Similarly, fidelity is not guaranteed by the emotional salience of the original statement. Intriguingly, that salience may be what degrades the transmission. And finally, factfinder calibration may not be fine-grained; although factfinders apply a broad hearsay penalty for using the PSI and EU as evidence, they may not track the distinctions that courts use to police the boundaries of those exceptions.¹⁰⁷

IV. LEGAL IMPLICATIONS

The analysis in Parts I through III does not support the categorical abolition of the PSI and EU exceptions, and this Article does not advocate for such abolition. What it supports is a more honest accounting of what these exceptions' reliability actually guarantees—and a set of interventions at each link in the reliability chain that would bring the doctrine into closer alignment with the psychological evidence. The recommendations that follow are deliberately cautious. They are grounded in existing doctrinal tools, and most do not require legislative revision of the Federal Rules. They are offered as starting points for reform, not as a complete agenda, and they will be better supported as the research program described in Part V matures.

A. Asking Better Questions

Courts could act first, and most easily, on the gatekeeping inquiry that precedes admission of PSI and EU evidence. The existing multi-factor tests applied by courts across the Circuits ask, in various combinations, about the passage of time, the nature of the startling event, and the declarant's apparent state of mind or demeanor.¹⁰⁸ These factors are proxies for the psychological conditions the exceptions assume. As Parts I through III demonstrate, they are imperfect proxies.

The pilot findings suggest two additional considerations for the gatekeeping inquiry: (1) the ambiguity of the scene being described, and (2) the cognitive task the declarant is performing at the moment of

107. These are pilot results and should be interpreted cautiously as a proof-of-concept for future research designs. Appropriate next steps include replicating this work with larger samples, more ecologically valid materials, and (to be discussed in Part V) cascading designs that trace any distortion from the veracity stage through fidelity and into calibration within a single experimental framework.

108. The proposal does not necessarily require statutory amendment to Federal Rule of Evidence 803(1) or 803(2). Courts already apply multi-factor tests. *See* Baicker-McKee, *supra* note 10, at 125–28 (surveying seven distinct appellate approaches). The proposal is that courts incorporate two additional factors into existing gatekeeping analysis: scene ambiguity and cognitive task demands.

utterance. Consider ambiguity first.¹⁰⁹ The first pilot showed that inferential elaboration is not primarily a function of delay or arousal; instead, it is a function of narration itself, and it is most pronounced when the event being described is socially ambiguous. Courts should still ask how soon the statement was made and how excited the declarant appeared, but they should also ask how ambiguous the scene was that the statement purports to describe. A statement made immediately and excitedly about an ambiguous event may carry more inferential content than a statement made with slightly more delay about a clearly observable one.

Next, consider the cognitive task being performed. The PSI exception's requirement that narration be contemporaneous assumes that simultaneous perception and description is more reliable than narration from brief memory.¹¹⁰ The dual-task findings in the first pilot study complicate that assumption: declarants who described a scene while perceiving it showed evidence of cognitive compression. Courts that ask "How soon?" without asking "What was the declarant cognitively doing at that moment?" are missing a variable that the psychology literature suggests is significant.¹¹¹

B. Improving Transmission

Fidelity is harder to address at the admissibility stage because it occurs downstream in the interval between the declarant's utterance and the in-court witness's repetition of it. The second pilot's findings suggest that the emotional charge of the original statement—the very feature that triggers the admission of the EU—may degrade fidelity by amplifying hedge loss and certainty inflation in the recipient's retelling.¹¹²

The first more modest intervention is heightened attention to the in-court witness's basis for claiming to have accurately repeated the declarant's statement. Where the original statement was recorded—for

109. The ambiguity inquiry has analogues in existing doctrine. The personal-knowledge requirement already asks courts to evaluate the declarant's epistemic position. *See, e.g.,* *People v. Cummings*, 31 N.Y.3d 204, 210–11 (N.Y. 2018). The proposal here extends that inquiry: courts should ask not only whether the declarant observed the event but whether the event, as observed, was susceptible to interpretive inflation.

110. *See* FED. R. EVID. 803 advisory committee's note to Rule 803(1) ("[S]ubstantial contemporaneity of event and statement negative the likelihood of deliberate or conscious misrepresentation.").

111. *See* *United States v. Boyce*, 742 F.3d 792, 800–02 (7th Cir. 2014) (Posner, J., concurring) (calling for more rigorous scrutiny of the specific circumstances surrounding a statement's production).

112. *See supra* Part III.B.

example, on a 911 call or a body camera—courts and counsel should prefer admission of the recording over the witness’s paraphrase, because the recording is not subject to the fidelity distortions that accompany retelling. Where no recording exists, counsel might usefully explore on cross-examination whether the in-court witness recalls specific words or is reconstructing the substance of what was said—a distinction that tracks the hedge-loss problem the fidelity pilot identified.

The second intervention is more structural in nature and more ambitious: evidence rules committees might consider whether a limited corroboration or specificity requirement, analogous in spirit to the recorded-recollection rule’s requirement that the record accurately reflect the witness’s prior knowledge,¹¹³ would be appropriate for PSI and EU evidence offered through the testimony of a recipient witness, rather than through a contemporaneous recording. The current rules impose no such requirement, relying instead on the exceptions’ admissibility rationales to do the fidelity work that the rationales never purported to do.

C. Giving Factfinders Better Tools

Calibration points toward a reform that courts and rules committees can adopt without changing the substantive hearsay rules: improved jury instructions specific to PSI and EU evidence. Where the current pattern instructions exist at all, they provide factfinders with only occasional guidance about hearsay.¹¹⁴ This guidance does not tell factfinders anything about the specific conditions under which PSI and

113. The analogy to recorded recollection is instructive. Federal Rule of Evidence 803(5) conditions admission on requirements designed to ensure fidelity: the record must have been made when the matter was fresh and must accurately reflect the witness’s knowledge. *See* FED. R. EVID. 803(5). No comparable fidelity assurance exists for PSI or EU evidence admitted solely through a recipient’s testimony.

114. *See, e.g.*, NINTH CIR. MODEL CRIM. JURY INSTR. 2.12 (2018), “Evidence for Limited Purpose” (providing as an example in its official comment, “[I]n *United States v. Souza-Martinez*, 217 F.3d 754, 760 (9th Cir. 2000), the Ninth Circuit held the trial court ‘had no alternative’ but to give the jury a limiting instruction sua sponte when a testifying codefendant’s post-arrest statements were admitted as substantive evidence against her under Fed. R. Evid. 801(d)(2)(A) but were not admissible against another codefendant...”); *see also* THIRD CIR. MODEL CRIM. JURY INSTR. 4.22 (Apr. 2024), “Impeachment of Witness – Prior Inconsistent Statement for Credibility Only” (clarifying that prior inconsistent statements offered for the purpose of impeachment may not be used “as proof of the truth” of the witness’s earlier statement). Current pattern jury instructions do not provide factfinders with PSI-EU-specific guidance. The calibration pilot’s blunt-discount finding suggests that general limiting instructions are not doing the work the doctrine’s admissibility requirements contemplate.

EU statements are produced, the specific error patterns those conditions generate, or even the specific questions they should ask when evaluating such statements.¹¹⁵

A more specific instruction would explain, in accessible terms, what the PSI and EU exceptions are premised on and what their premises do and do not guarantee. It might convey that the contemporaneity of a PSI reduces some opportunities for deliberate misrepresentation but does not ensure that what the declarant described accurately reflects what was objectively observable, particularly when the scene was ambiguous or when the declarant was simultaneously perceiving and narrating under cognitive load. It might similarly note that the excitement of an EU may reduce the opportunity for deliberate fabrication but might also affect the speaker's perception, the selectivity of their attention, and the relationship between their expressed confidence and actual accuracy.

The fairness dimension of the calibration finding is harder to address through instruction reform alone. Factfinders' normative discomfort with PSI and EU evidence may reflect a concern about confrontation and participation that jury instructions cannot fully resolve. That concern points toward the Confrontation Clause doctrine governing when PSI and EU evidence is "testimonial" and therefore subject to the defendant's cross-examination right, and toward whether the current testimonial/non-testimonial constitutional line adequately captures the procedural-justice interests that factfinders are tracking.¹¹⁶ That is a question for another day—and another article.

These recommendations share a premise: that the exceptions would serve their purposes better if courts attended separately to veracity, fidelity, and calibration. The exceptions need not be abolished to be improved. But they need to be understood more carefully than the doctrine and the existing scholarship has yet allowed.¹¹⁷

115. See *supra* note 114 and accompanying text.

116. See *Crawford v. Washington*, 541 U.S. 36, 53–54 (2004) (holding that the admission of "testimonial hearsay" against a defendant without a meaningful opportunity to cross examine the speaker violates the Sixth Amendment's Confrontation Clause); *Davis v. Washington*, 547 U.S. 813, 822 (2006) (applying *Crawford* and distinguishing non-testimonial statements made to meet an ongoing emergency from testimonial statements made primarily to establish past facts for later prosecution).

117. See Sevier, *Evidence-Based Hearsay*, *supra* note 74, at 1823–25 (arguing for evidence-based hearsay reform rather than abolition). The modest reform agenda is also consistent with Lau's more cautious position, which recommends against categorical change. See Lau (EU), *supra* note 10, at 634–36. This Article identifies specific reforms using courts' existing gatekeeping discretion and authority over jury instructions.

V. EMPIRICAL IMPLICATIONS

The three pilots reported in Part III suggest that the three-stage framework generates useful research questions. This Part, intended for behavioral scientists, describes a research program that would begin to explore them more deeply, organized by stage of the reliability chain and focused on the specific design features that would make the resulting findings most useful for doctrinal purposes. Each suggestion addresses a gap identified by the gap-map analysis and may produce findings directly relevant to the reform recommendations of Part IV.

A. Veracity: Ecologically Valid Designs

The veracity pilot's stimuli were chosen for manageability rather than ecological validity, and that choice limits what the pilot can show.¹¹⁸ The most important next step for the veracity stage is to move toward stimuli that more closely approximate the conditions under which PSI and EU evidence is generated.

A first priority is to develop stimuli that vary along the three dimensions that the exceptions treat as reliability signals—timing, arousal, and scene ambiguity—while holding constant the objectively observable features of the event. A stimulus set with these properties would enable the first direct test of how each variable independently affects accuracy and inferential content and how the variables may interact.

Archival and field-based work using naturalistic PSI-EU evidence is the second priority, and the more promising one.¹¹⁹ Specifically, it could include body camera footage or 911 call recordings in jurisdictions

118. Veracity research is listed first because the existing veracity paradigms are most developed and most readily adapted to PSI-EU-specific designs. A well-characterized veracity stage is a necessary precondition for interpreting fidelity and calibration findings downstream.

119. The importance of ecological validity for PSI-EU research is worth contrasting with the laboratory-translation problem Lau identifies. *See* Lau (PSI), *supra* note 48, at 196 ("It is one thing to analyze the PSI hearsay exception using the general findings and observed trends from social science studies, which are actually quite consistent. However, the actual values from the studies about lying should not be taken out of the context of how those measurements were taken without an analysis of whether those results can meaningfully apply to real-world PSI statements."). The proposed research program addresses this problem by designing paradigms that approximate the specific cognitive and social conditions under which PSI and EU evidence is actually produced. Professor Ronald Allen has also criticized the debate's overreliance on out-of-legal-context laboratory studies. *See* Ronald J. Allen, *The Hearsay Rule as a Rule of Admission Revisited*, 84 FORDHAM L. REV. 1395, 1403 (2016) (expressly criticizing reliance on one psychology study on this ground).

where such recordings are publicly available.¹²⁰ The 911 call is perhaps the paradigmatic PSI-EU context, and, one that the existing laboratory literature has likely never directly studied. An archival program coding 911 call transcripts for perceptual description versus inferential elaboration, cross-referenced against available objective evidence, would supply an ecologically valid baseline that the pilot paradigms, although useful, can only approximate. Alongside this, experimental manipulations of accuracy incentives would help disentangle the structural (i.e., task-related) from the motivational (i.e., willingness to lie) contributors to inferential elaboration.

B. Fidelity: Testing Correctives

In the fidelity pilot, the emotional framing of an original statement shaped what recipients reported hearing, resulting in hedge loss and certainty inflation. The next step is to replicate and extend this effect with larger samples, a wider range of stimulus materials, and manipulations that test whether the effect can be attenuated.

Doing so requires at least two methodological additions. First, the audio recordings used as stimuli should be expanded to include a wider range of affective intensities and event types. Second, the coding scheme could be extended to capture source confusion—instances in which the recipient's retelling attributes to the declarant content that originated from the recipient's own inference.¹²¹

Two corrective interventions also invite testing. The first is the effect of a brief fidelity instruction to the recipient—a prompt emphasizing the importance of reproducing the declarant's exact words rather than the

120. 911 call recordings are a particularly promising, though jurisdictionally variable, source of naturalistic PSI-EU evidence. 911 calls frequently involve declarants reporting events close in time to perception, and courts often encounter them in PSI-EU litigation. See Baicker-McKee, *supra* note 10, at 124, 126–27; Lau (EU), *supra* note 10, at 640. Access to 911 recordings varies by state law. See *Open Government Guide: 911 Tapes*, REP. COMM. FOR FREEDOM OF THE PRESS, <https://www.rcfp.org/open-government-sections/3-911-tapes/> [<https://perma.cc/32T6-BRLN>] (last visited Aug. 19, 2026). Access to body-camera footage likewise varies by jurisdiction. See *Access to Police Body-Worn Camera Video*, REP. COMM. FOR FREEDOM OF THE PRESS, <https://www.rcfp.org/resources/bodycams> [<https://perma.cc/JA5E-8SBQ>] (last visited Aug. 19, 2026).

121. The serial-retelling design for fidelity research is well established. See BARTLETT, *supra* note 97, at 118–20. The present program adapts it to the specific legal question: a single retelling, from declarant to in-court witness, with fidelity outcomes coded at the level of specific linguistic features pertinent to the legal question of what the declarant actually said.

gist of what was said.¹²² If such an instruction reduces hedge loss and certainty inflation, it would suggest a practical intervention at the point of witness preparation that could be implemented without altering the admissibility rules. The second is the effect of cross-examination specifically directed at transmission fidelity.¹²³ If cross-examination improves calibration in the factfinder studies, that would support the existing adversarial structure. If it does not, that would strengthen the case for the structural reforms suggested in Part IV.

C. Calibration: Fine-Grained Weighting

The calibration pilot replicated the categorical-discount finding for hearsay evidence and the fairness-reliability asymmetry. The categorical-discount finding calls for an experimental test of the admissibility-laundrying hypothesis in the hearsay context.¹²⁴ Testing this hypothesis requires a design in which the admissibility ruling is separated from the evidence's substantive content. Participants would receive the same hearsay statement but would be told either that the judge admitted it as a PSI or EU, that the judge admitted it over objection, or that its admissibility status is undisclosed. Differences in credibility ratings and guilt verdicts across these conditions would isolate any laundrying effect.

Instruction-based interventions addressing the procedural-justice dimension of factfinders' hearsay assessments are the natural next step.

122. Significant scientific research suggests that people frequently reason and remember in terms of "gist" rather than verbatim. *See, e.g.*, Christopher R. Wolfe, Mitchell Dandignac & Valerie F. Reyna, *A Theoretically Motivated Method for Automatically Evaluating Text for Gist Inferences*, 51 BEHAV. RES. METHODS 2419, 2420 (2019) (discussing "fuzzy trace theory" as a "dual-process theory emphasizing meaning making and mental representation" in which people "simultaneously encode information into multiple representations that range from verbatim surface details to the bottom line meaning or gist").

123. Notable voices in the evidence literature have suggested that cross-examination is an imperfect corrective for a witness's sincere but inaccurate beliefs. *See, e.g.*, FED. R. EVID. ART. VIII advisory committee's introductory note to 1972 proposed rules ("The belief, or perhaps *hope*, that cross-examination is effective in exposing imperfections of perception, memory, and narration is fundamental.") (emphasis added); EDMUND M. MORGAN, FOREWORD TO MODEL CODE OF EVIDENCE 37 (1942) ("[A] cross-examiner frequently discloses imperfections as to observation and memory, while it is a rare case where he uncovers deliberate falsehood. The commonly repeated charge of rampant successful perjury in the courtroom is a demonstration of the futility of cross-examination to expose it . . .").

124. For a test of that hypothesis in the expert evidence context, see Schweitzer & Saks, *supra* note 75, at 4–13. A confirmed laundrying effect would significantly undermine the argument that judicial gatekeeping provides adequate protection against the risks the exceptions create.

Whether revised instructions actually improve calibration is an empirical question that the calibration research program can answer. If revised instructions fail to improve calibration, that would support a more structural response: either more rigorous admissibility gatekeeping or categorical reform of the exceptions.

D. The Cascading Design: Tracing Distortion Across the Chain

The individual-stage studies are necessary to further develop the empirical record regarding the psychology of the PSI and EU hearsay exceptions. But even more ambitious work can be done. Perhaps the most important single study in this research program would be a cascading design that traces the same event through all three stages within a single experimental framework.

The design could work as follows. A veracity participant observes an event and produces a PSI- or EU-like statement about it. That statement is then passed to a fidelity participant, who receives and retells it, approximating what an in-court witness would offer. That retelling is then presented to a calibration participant in the context of a trial vignette, who renders verdicts and assessments of reliability and fairness. By holding the underlying event constant across all three stages, the cascading design enables the first direct measurement of how distortion compounds as it moves through the chain.¹²⁵ This cascading design requires well-characterized and validated paradigms at each stage, which the individual-stage studies would supply, and stimuli that can be passed through the full chain without losing coherence. The pilots discussed here suggest that the constituent parts are individually feasible.

Field-based cascading design research would provide policymakers with value by focusing on PSI and EU evidence as they appear in real proceedings. Body-camera footage, paired with 911 call recordings and court transcripts, would provide the kind of real-world data that no laboratory paradigm can approximate. This approach would require institutional collaboration and thoughtful consideration of privacy and ethics concerns. But it would also yield findings on how the reliability chain actually operates, from the moment the declarant speaks to the moment the factfinder decides.

The research programs described in this Part ultimately serve to identify the types of empirical evidence the law should consider as it updates its understanding of the reliability of the PSI and EU exceptions.

125. In sum, the cascading design measures whether distortion compounds across stages rather than treating veracity, fidelity, and calibration as isolated patterns.

The existing debate has been conducted largely on borrowed science, insofar as it has relied on general psychological findings that speak to the veracity stage without being designed for it, and say little about the fidelity and calibration stages. The program proposed here could therefore fill the gap by providing evidence tailored to each stage of the PSI-EU reliability problem for policymakers to consider.

CONCLUSION

In the background of a 911 call made on a March afternoon in Harlem, an unidentified voice said nine words: “Yo, it was Twanek, man! It was Twanek, man!”¹²⁶ Those words were emotionally charged and were spoken immediately after a startling event, and so, on the surface, they satisfy everything the excited-utterance hearsay exception requires. And for the purposes of one trial judge and the jury she instructed, they were the most important evidence in a criminal prosecution.

The New York Court of Appeals reversed on narrow grounds: the declarant’s personal knowledge could not be established. But the deeper problem on which this Article has focused is that the personal-knowledge question, even if satisfactorily answered, does not resolve the tougher behavioral-science question: what was actually happening cognitively when the speaker named Twanek Cummings?

The current doctrine cannot fully answer that question, in part because it never set out to ask it. The PSI and EU exceptions were built on psychological stories about what timing and arousal warrant in terms of evidentiary reliability, and those stories—however intuitive—have not been subjected to the kind of empirical scrutiny that the stakes of criminal prosecution demand. This Article has argued that the scrutiny is overdue and the first steps toward it are feasible—and it has tried to show both what the debate has been missing and how to begin filling those gaps.

126. *People v. Cummings*, 31 N.Y.3d 204, 207 (N.Y. 2018).