Understanding Validity in Empirical Legal Research: The Case for Methodological Pluralism in Assessing the Impact of Science in Court

TENEILLE R. BROWN,* JAMES TABERY,** AND LISA G. ASPINWALL***

What makes a study valid or invalid? In 2013, the Hastings Law Journal *published a law review article by law professor Deborah Denno entitled* What Real-World Cases Tell Us About Genetic Evidence. *This article questioned the validity of an article that we published in* Science*:* The Double Edged-Sword: Does Biomechanism Increase or Decrease Judges' Sentencing of Psychopaths? *Denno's trenchant critique focused on our use of experimental, rather than archival, methodology, and revealed a misunderstanding of the diverse goals of empirical legal research. One study, which in our case investigated the impact of biological explanations of criminal behavior on sentencing, is not meant to answer the universe of potentially relevant questions. This is as true in science as it is in law. Rather, experimental and archival projects complement each other by asking and answering different questions aimed at different forms of validity. We describe archival and experimental research methods, and then explain how their design impacts external validity, including concerns of ecological validity, robustness, and generalizability; internal validity; and construct validity. We appreciate Denno's questions about external validity in particular, specifically asking how and under what conditions a particular set of experimental effects might occur in real court cases. However, the questions she poses do not challenge the internal validity of our study—that is, its ability to identify particular causal factors that influence judges' ratings and sentencing decisions in the particular set of conditions and case features we tested. By explaining the tradeoffs between different forms of validity, this brief article may serve as a helpful tool for scholars in law, psychology, and other social sciences, as well as attorneys and judges who rely on empirical legal research in their work.*

* Teneille R. Brown is a Professor of Law and a member of the Division of Medical Ethics and Humanities at the University of Utah.

** James Tabery is an Associate Professor in the Department of Philosophy and a member of the Division of Medical Ethics and Humanities at the University of Utah.

*** Lisa G. Aspinwall is Professor and Chair of Psychology at the University of Utah.

TABLE OF CONTENTS

INTRODUCTION

Scientific evidence concerning the potential biological causes of criminal behavior is becoming increasingly common in the courtroom.[1] Genetic tests for genes associated with antisocial personality disorder and neuroimages of purportedly psychopathic brains have been introduced to affect the sentencing of convicted criminals.[2] The prevalence of this genetic and neurobiological evidence has drawn the attention of legal scholars, as well as social scientists, philosophers, and practicing scientists outside of law, who are interested in understanding the impact of this

1. Paul S. Appelbaum et al., *Effects of Behavioral Genetic Evidence on Perceptions of Criminal Responsibility and Appropriate Punishment*, 21 PSYCHOL. PUB. POL'Y & L. 134 (2015).

2. Emiliano Feresin, *Lighter Sentence for Murderer with 'Bad Genes*,' NATURE.COM (Oct. 30, 2009), http://www.nature.com/news/2009/091030/full/news.2009.1050.html; Virginia Hughes, *Science in Court: Head Case*, NATURE.COM (Mar. 17, 2010), http://www.nature.com/news/2010/100317/full/464340a.html.

science in the courtroom. They seek to answer a variety of questions: How is this evidence being used? How (if at all) does this biological evidence affect evaluations of responsibility and punishment? Do different forms of evidence affect these evaluations differently? Do judges and jurors make these evaluations differently?

Scholars have traditionally employed two methodologies to answer such questions: the *archival approach* and the *experimental approach*. The archival approach reviews the written records of a subset of actual court cases where such scientific evidence has been introduced. The goal is either to describe how the evidence has been used or to assess how it might have been correlated with the outcome. The experimental approach, on the other hand, seeks to test specific theories about the impact of particular kinds of evidence on sentencing and related judgments, and to shed light on the decisionmaking process by which such effects might occur. Experiments do this by presenting research participants with a hypothetical or actual case and then systematically varying whether participants receive different types of scientific evidence or none at all. Participants are then asked to provide ratings on dimensions relevant to the case (such as guilt versus innocence, recommended sentence, and so on) and also to answer questions about how they assessed the case.

In comparing or evaluating what has been or could be learned from either approach, it is not possible to say that a single study conducted with either method is valid or invalid writ large. Rather, we must interrogate whether the inferences drawn from the research are valid, which depends upon the type of validity we are assessing. Researchers, and those seeking to interpret their results, often balance concerns for external validity with concerns for internal validity. *Internal validity* is the ability to draw causal inferences from the study, specifically to show that an independent variable *X* (for example, kind of evidence presented) makes a causal contribution to some effect or dependent variable *Y* (for example, judgments of responsibility). *External validity*, on the other hand, encompasses multiple concerns in the extrapolation of the findings, including *ecological validity* (the degree to which the situation tested matches real-world conditions), *robustness* (the degree to which the findings will generalize to other populations and settings), and *relevance* (the degree to which the findings will be useful to others).[3]

3. *See generally* Marilynn B. Brewer, *Research Designs and Issues of Validity*, *in* HANDBOOK OF RESEARCH METHODS IN SOCIAL AND PERSONALITY PSYCHOLOGY 4–12 (Harry T. Reis & Charles M. Judd eds., 1st ed. 2000) (discussing how "validity must be evaluated in the light of the *purposes* for which the research was undertaken in the first place"). While we are chiefly concerned with the comparison between internal and external validity, a third form of validity, *construct validity*, refers to the degree to which the particular operationalization of an independent variable or dependent variable reflects the theoretical concept it was intended to represent. *Id.* at 4. Construct validity is important to experimental and archival research alike, as it informs generalization from specific findings to the underlying ideas being tested. *Id.*

Archival research has a decided advantage over experimental research when it comes to one form of external validity, namely ecological validity, while experimental research has the advantage when it comes to internal validity.[4] With respect to external validity, the results of an experimental study might not extrapolate to real-world legal cases because the study cannot include all of the factors involved in an actual case and because participants know that there will be no real consequences of their decision: no one will go to jail, or be fined, or be exonerated. Conversely, with respect to internal validity, because the inclusion of different types of information cannot be systematically varied, the results of an archival study of legal cases will never be able to conclude causation from correlation.

In this Article, we will be exploring how these issues of comparative validity play out in the context of scientific evidence and criminal sentencing. Specifically, we will do so with reference to an experimental study we published in 2012.[5] In this study, we tested how the addition of genetic and neurobiological evidence concerning the biological causes—what we labeled "biomechanism"—of psychopathy affected evaluations of responsibility and punishment in a diagnosed psychopath. Our research participants were U.S. state trial court judges—the very experts tasked with making such evaluations. We will describe and discuss the experimental design and results in greater detail below.

Professor Deborah Denno, a legal scholar and practitioner of the archival approach, extensively criticized our experimental methods because they did not mirror the historical cases that she has documented in which genetic or neuroscientific evidence was introduced. She criticized, for example, our choice of psychopathy for a psychiatric diagnosis, our choice of a sentence involving prison time rather than capital punishment or life in prison, and our focus on genetic factors to the exclusion of environmental contributors to criminal behavior.[6] We start in Part I by reviewing the archival and experimental approaches, explaining how the design features of each lend themselves to inherent advantages and disadvantages regarding different forms of validity. In Part II, we turn to how the lessons outlined in Part I apply to our 2012 experimental study, showing the ways in which the experimental approach was the appropriate methodology for the question we posed. We further explain how Professor Denno's criticisms of our research reflect a misunderstanding of the decisions that must be made when designing an experiment. Finally,

4. Because experimental designs, underlying case facts, and related judicial reasoning possess considerable variations, the concerns of robustness and relevance are not inherently more likely to be present in archival or experimental research.

5. Lisa G. Aspinwall et al., *The Double-Edged Sword: Does Biomechanism Increase or Decrease Judges' Sentencing of Psychopaths?*, 337 SCI. MAG. 846 (2012).

6. Deborah W. Denno, *What Real-World Criminal Cases Tell Us About Genetics Evidence*, 64 HASTINGS L.J. 1591 (2013).

in Part III, we make the case for a pluralistic relationship between the archival and experimental approaches, proposing ways in which the results of each can be combined to draw on their relative benefits and to counterbalance the limitations of the other.

I. Costs and Benefits of the Archival and Experimental Approaches

In Part I, we introduce the archival and experimental approaches to studying the impact of biological evidence in criminal cases. Each approach has certain costs and benefits regarding validity. For example, the archival approach's strength is ecological validity, but it might pose problems for internal validity. The experimental approach's strength is internal validity, but it might pose problems with ecological validity. These validity tradeoffs arise out of the inherent design features of the archival and experimental approaches.

A. The Archival Approach

Numerous scholars employ the archival method. Indeed, it is the traditional approach for all legal research, including research on the impact of biological evidence in criminal cases. In addition to the thorough work of Professor Denno,[7] Nita Farahany and James Coleman's work demonstrates the value of collecting data on actual cases to determine how genetic information has been used in various contexts.[8] They do this by amassing and then reading the extensive case law and then categorizing opinions and their justifications into various groups. The classification system can be decided a priori or ex post, and can be done according to whichever variables the researcher finds interesting. These variables might include sentencing outcome, introduction of a particular type of evidence, or known demographics of the defendant. Researchers may also perform formal content analyses on opinions, where predicted text is given a "code" and then systematically reviewed to describe which types of evidence are used most often, by whom, and to what effect.

Archival research has the benefit of relying on actual decisions by judges in real-world cases. This grounds the data in an ecologically valid context. For example, because Professor Denno reviewed 800 published appellate cases over the course of two decades, she can describe what is common in real criminal cases. This permits her observation, at least in published, appellate cases, that the application of genetics evidence "is

7. Deborah W. Denno, *The Myth of the Double-Edged Sword: An Empirical Study of Neuroscience Evidence in Criminal Trials*, 56 B.C. L. Rev. 493, 493 (2015).

8. Nita A. Farahany & James E. Coleman Jr., *Genetics and Responsibility: To Know the Criminal from the Crime*, 69 Law & Contemp. Probs. 115, 116, 119–25 (2006); Nita A. Farahany, *Neuroscience and Behavioral Genetics in US Criminal Law: An Empirical Analysis*, 2016 J.L. & Biosciences 1.

likely to take place in the context of capital cases,"[9] and "neuroscience evidence is usually offered to mitigate punishments in the way that traditional criminal law has always allowed, especially in the penalty phase of death penalty trials."[10] An additional advantage of the archival approach is that it can describe multiple factors that may in combination be correlated with judicial outcomes (such as the age or ethnicity of the defendant or victim), or that may co-occur so regularly that they may rarely occur separately (such as the presentation of the defendant's neuropsychological testing and a history of her being physically abused).

1. The Problem of Trying to Discern What Occurred at Trial from the Appellate Record

Although archival research has clear advantages in terms of ecological validity, it also has some less frequently appreciated drawbacks. One limitation concerns the representativeness of the cases examined, which in turn compromises external validity, specifically robustness. Archival research relies almost entirely on appellate decisions, not decisions of the trial court. This is because the large majority of trial court decisions are not published online in searchable databases like Westlaw or LexisNexis, while most appellate decisions are. Federal trial courts and some state trial courts are beginning to publish opinions online, but those who do are still relatively few. To obtain complete state criminal trial records, archival researchers still need to visit the court and make hard copies of the transcripts that reside in physical files. These files are typically not searchable electronically, so in order to locate the case file researchers usually have to know the names of the litigants in advance. Also, as appellate courts defer to and summarize the factual findings of the lower trial courts, appellate decisions provide incomplete records of the evidence presented at trial and the trial court's reasoning about this evidence. Yet, this problem persists even when we do have access to trial court decisions. Because trial judges can choose which facts to highlight in their opinions, they may not even mention the scientific experts who testified at trial, or what types of exhibits were admitted along with their testimonies. This makes it very difficult to make comparisons between cases when the impact of the scientific evidence is precisely the thing you are interested in studying.

2. The Problem of Opaque Judicial Reasoning

In addition to providing an incomplete record of the evidence presented at trial, judicial opinions also provide incomplete records of the judges' reasons for their sentencing decisions. The full range of the

9. Denno, *supra* note 6, at 1604.

10. Denno, *supra* note 7, at 493, 544.

judges' mental deliberations will not be revealed by reading the archives; we only have access to what the judge chooses to highlight in her heavily edited published opinion. For some aspects of the case the judges' reasons are stated explicitly, but for others, no reason is provided. If we are interested in knowing *how a judge reasons* through her sentencing decision, trial opinions might say nothing about the judge's evaluation of recidivism data, the lack of remorse by the defendant, her arrogant demeanor, a particularly compelling witness, or some combination thereof. Alternatively, if we are curious about the judge's philosophy of punishment, we might not know whether she was motivated chiefly by retributivism, specific or general deterrence, or something else. A review of archival opinions cannot shed light on the judge's deliberative process in a systematic way, and to the extent the *trial* judge's reasoning is even included in the appellate record, the reasoning is typically heavily excerpted. Thus, this problem is compounded when the archive consists only of appellate cases.

The experimental approach allows researchers to gain additional information about the judges' reasoning by asking participants qualitative questions such as, "Why did you recommend the sentence that you did?" Asking judges such open-ended questions ensures that responses are generated by participants and not primed by the researchers. However, it has the drawback of not accounting for judges who in fact relied on this factor but never spontaneously mentioned it.[11] Further, in some cases, a judge might mention a mitigation argument without explicitly endorsing it in order to signal to reviewing courts that she at least considered arguments that might be constitutionally required. Conversely, a judge might *not* mention a reason that she would ordinarily mention in a published case, specifically because she knows that no appellate court will ever be reviewing her individualized reasons. Thus, even with the experimental approach, researchers still will never know exactly which judges *actually employed* specific reasons in their decisions.

Put simply, if you want to know more about what actually happened at trial—what evidence was presented and by whom, which specific reasons were considered by the judge, and which reasons were influential in the decision—you will often have a very hard time discerning this from the appellate record. In this way, external validity, or the confidence that you can generalize findings from the archival data you reviewed to a broader set of cases, is compromised when we rely only on the appellate record to discern what happened at trial.

11. For a discussion of the strengths and limitations of structured interviews and other qualitative methods versus survey instruments, see James C. Coyne & Benjamin H. Gottlieb, *The Mismeasure of Coping by Checklist*, 64 J. PERSONALITY 959 (1996).

3. The Problem of Systematic Selection Bias in the Appellate Record

A related drawback of archival research based on appellate decisions is that in criminal cases there is considerable and systematic selection bias. If a defendant is acquitted, the government cannot appeal the case. The only criminal cases that are appealed, and thus result in published opinions, are those where the defendant lost at trial. This presents an extremely biased view of the criminal case law, as the appellate record might not capture a significant amount of what is happening "on the ground" at trial. For example, biological evidence might have a different impact when the prosecution has a weak case or the defense has a strong one. If a defendant successfully introduces genetic or neuroscience evidence of psychopathy at the guilt phase, or the prosecution unsuccessfully does so, and the defendant is then acquitted, this use of genetics or neuroscience evidence will be *completely absent* from the appellate record. Unless we have access to trial records, the archival approach does not allow us to say anything about what happens in the many cases where the defendant was acquitted. Thus, reliance on archival research alone is likely to result in an incomplete and potentially misleading subset of cases.

4. The Problem of Using Retrospective Cases to Understand the Legal Impact of New Science

A fourth limitation of the archival approach concerns its inherently retrospective nature. The archival method can only assess the impact of actual evidence in actual cases that have actually transpired. It cannot prospectively consider the impact of new forms of scientific evidence that are being investigated but which have not gained acceptance in the courtroom. Specifically in behavioral genetics, there has been a transformation in the way that scientists have sought to explain behavior. Where human genetics used to describe genetic variations in populations, which were associated with behavioral traits, geneticists can now identify coding defects in specific genes in specific individuals that cause specific types of dysfunction. The latter type of personalized genomic evidence is rapidly developing and unlikely to be present in appellate cases. Neuroscientific evidence is also swiftly changing, with new imaging methodologies being developed every year. If we want to understand the impact of such evidence in the courtroom, then, it is important to keep in mind that what counts as "genetic" or "neuroscientific" evidence is evolving.[12] The archival approach can only assess what has come to pass, not what is here now (but has not yet entered a courtroom) or what is to come.

---

12. *See* JAMES TABERY, BEYOND VERSUS: THE STRUGGLE TO UNDERSTAND THE INTERACTION OF NATURE AND NURTURE (2014).

5. *The Problem of Poor Internal Validity*

Finally, even if we can access trial records and those records are complete and representative, the archival approach is dogged by poor internal validity. It cannot support causal conclusions about how particular kinds of scientific evidence considered in actual cases affected the actual outcomes. That is, researchers employing the archival approach can document things like who introduced the scientific evidence and what kind of scientific evidence was introduced. They can also search for and identify references to the scientific evidence in the decision. However, that scientific evidence would be embedded in a complex web of other scientific evidence and information about the crime, victim, and defendant, which makes it impossible to isolate the specific impact of one piece of evidence (or a key combination of two or more pieces) in that complex web. Crucially, this means researchers cannot establish whether the scientific evidence caused a guilty verdict, a mistrial, or a greater or lesser sentence in a specific case. Of course, with greater complexity in context and evidentiary sources, the archival approach provides superior ecological validity. What is gained in ecological validity, however, is lost in internal validity. Put another way, you might be able to describe events accurately, but you cannot isolate and quantify the effect of a variable on an outcome, or explain why the effect might have occurred.

B. The Experimental Approach

Experimental studies can make all aspects of the instructions to participants and all features of the case identical, such as the courtroom scenario and other evidence presented between experimental conditions, except those that are systematically varied between conditions as independent variables.[13] This permits researchers to isolate the effect of the presence or absence of the independent variable(s) and allows for an internally valid test of whether these variables have a corresponding effect on the verdict or sentencing decision. In this way, experimental studies can say something about causes and effects in a way that archival research cannot. However, just because a study is experimental, it is not immune to thoughtful criticism. Great care must be taken to reduce potential sources of bias in the selection of research participants, the operationalization of the independent and dependent variables,[14] and the creation of realistic and involving stimuli, among other factors.

13. For examples of this, see Paul S. Appelbaum & Nicholas Scurich, *Impact of Behavioral Genetic Evidence on the Adjudication of Criminal Behavior*, J. Am. Acad. Psychiatry L. 42 (2014); Johannes Fuss et al., *Neurogenetic Evidence in the Courtroom: A Randomised Controlled Trial with German Judges*, J. Med. Genetics 52 (2015); and, Nicholas Scurich & Paul Appelbaum, *The Blunt-Edged Sword: Genetic Explanations of Misbehavior Neither Mitigate nor Aggravate Punishment*, J.L. & Biosciences 1 (2015).

14. *See* Brewer, *supra* note 3, at 4–12.

Because well-designed experiments can systematically vary the presentation of scientific evidence, they can eliminate the source selection bias that is present—and amplified—when we focus solely on appellate decisions. While not unique to this domain, experimental studies allow researchers to draw causal inferences based on the findings and also follow up with questions designed to probe the participants' self-reported rationales.[15] In our study, this meant that we could supplement judges' ratings of responsibility and punishment with open-ended questions about *why* judges chose the sentence that they did and *why* they found the diagnosis with or without the scientific evidence to be either mitigating or aggravating. In other nonlegal settings, one might want to understand the participants' reasoning process by doing more than simply asking them about it.[16]

1. The Problem of Ecological Validity

Even though experimental methods permit the isolation of the causal impacts of specified types of scientific evidence, the experimental approach is typically compromised when it comes to ecological validity. Vignettes of cases that research participants read can convey information about a hypothetical or actual case, but that task is quite different from listening to the emotional testimony of a victim, watching the defendant's reaction to such testimony, and listening to cross-examination. Simulated cases can employ actors to play out such features, certainly increasing ecological validity. But ecological validity remains limited in even the most engaging simulations because the research participants watching the production know they are watching a play. That is, they are judging an actor playing a defendant, rather than an actual defendant who committed an actual crime and who will actually be punished. One strength of the archival approach is that many different kinds of cases and kinds of evidence can be represented in the analysis. However, some of that benefit is diminished when one considers that the rater or judge *also* varies across cases. This variation makes it impossible to determine which specific factors related to either the case or the judge account for the differences in outcomes, such as variation in sentencing.

2. The Problem of Construct Validity

Construct validity is a third form of validity that is separate from external and internal validity. As noted earlier, construct validity captures whether the particular operationalization or instantiation of the

15. Even so, there is a lively debate about the accuracy of self-reported reasoning. *See* Richard E. Nisbett & Timothy DeCamp Wilson, *Telling More than We Can Know: Verbal Reports on Mental Processes*, 84 PSYCHOL. REV. 231 (1977).

16. For example, researchers could use structured items asking about particular decisionmaking endpoints.

independent variables convincingly measures the construct it was designed to assess. Simply put, construct validity represents the judgment of the extent to which the independent variable is what you say it is. For example, if a researcher wants to study the effect of hunger on classroom performance, and she decides to make one classroom of kids wait in line for thirty minutes for a late lunch, she might have manipulated frustration instead of, or as well as, hunger, making the independent variable potentially confounded. The method is still an experiment in that the delayed lunch was applied to one group but not another, but not a valid one if what the researcher says she is measuring is hunger. In the present context, one could ask the degree to which the kind of neurogenetic or biomechanical evidence presented in our study effectively manipulated the effect of biomechanism on the sentencing of a psychopath. For example, is the biomechanism story we told the best one; that is, is it the one that most correctly captures the causal mechanisms of psychopathy? Our study selected a particular explanation of psychopathy that was based on the genetics of the monoamine oxidase A allele and the neurodevelopmental model proposed by Dr. James Blair.[17] While this represents a validated model of the development of psychopathy, a different explanatory model might prove to achieve greater construct validity.

3. The Problem of Generalizability of the Sample

Related to ecological validity is the problem of sampling in experimental research. Ideally, the experimental sample of participants will reflect the relevant community being investigated. So, for example, in research on courtroom decisionmaking, the respondents should be representative of the actual jury or judge population. This ideal is often not realized. Instead, undergraduate students or members of the online community traditionally serve as research participants. Undergraduate students over the age of eighteen and internet users do serve on juries, and are therefore useful in answering questions about jurors' decisionmaking. However, the demographics of undergraduate students will not reflect a representative cross section of jurors, given that they are younger, more educated, and may come from households with higher incomes than the typical members of a jury pool.[18] Thus, a common critique of experimental research is that there is something about the nonrepresentativeness of participants that affects the way that they respond to the experimental stimuli. For example, younger mock jurors

17. R. J. R. Blair, *The Emergence of Psychopathy: Implications for the Neuropsychological Approach to Developmental Disorders*, 101 COGNITION 414 (2006).

18. David O. Sears, *College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature*, 51 J. PERSONALITY & SOC. PSYCHOL. 515 (1986); *see also* Joseph Henrich et al., *The Weirdest People in the World?*, 33 BEHAV. & BRAIN SCI. 61, 63 (2010).

might be less retributive in their punishment philosophies or more receptive to biomechanism evidence. In our experiment we relied on state trial court judges as respondents, so this common critique about the generalizability of the sample is not as applicable to the study discussed herein. That said, there still might be something about the judges who responded that makes our sample unrepresentative of the entire state trial bench.

4. The Problem of Expense and Access to Resources

While it is not always the case in experimental research, legal experimental research is often more expensive and resource intensive. Legal academics receive unlimited access to Westlaw and Lexis, the two main databases for archival research. However, if one wants to construct an experimental study of legal research, additional funds might be needed to recruit and reimburse participants, and for software for stimulus presentation and data collection and analysis. In order to study legal decisionmaking with greater ecological validity, in essence by using actors and mock litigation, the expenses are greater, as the actors and mock jurors need to be paid, scripts have to be written, and courtroom space might need to be rented.

II. Aligning the Approach with the Chosen Research Question

Our experimental approach relied on actual trial judges reading a detailed, hypothetical case online in which genetic and neurobiological evidence were or were not presented. The research participants then answered a series of questions where they evaluated the free will, moral responsibility, legal responsibility, and suitable sentence of the convicted criminal. We sought to ask a very specific question: Would judges use the evidence of a biological mechanism of psychopathy (what we called the "biomechanism") to argue that because the defendant's violent behavior is "fixed" in her genes and brain, she had less control of her behavior and should receive a shorter sentence? Or, would judges use this same evidence to argue that because the behavior is "fixed" in the genes and brain, it is likely to recur and therefore the sentence should be greater to keep this convict off the streets? Or would both forms of reasoning be evident in some experimental conditions? If the additional biomechanism evidence increased punishment, we hypothesized this would be due to judges' reliance on a deterrence theory of punishment (the defendant poses a risk of offending again). If the biomechanism evidence decreased punishment, we hypothesized this would be due to judges' reliance on a retributive theory of punishment (the defendant had a harder time controlling her conduct). The ability for the same evidence to be used in opposite ways by the prosecution and defense had been referred to as the "double-edged sword," and so we set out to test which way the sword cut

when presented by either the defense or the prosecution. We also wanted to know, whichever way the sword cut, how the judges justified their sentences and evaluations in the presence or absence of information about a biomechanism of psychopathy—whether they appealed to retributive or to deterrent theories of punishment.

Because of these specific aims, and our desire to isolate the role played by the biomechanism, we employed the experimental rather than the archival approach. Our research questions could not be answered by looking at published cases that (1) do not systematically manipulate the presence or absence of different kinds of evidence; (2) only occasionally explain sentencing decisions in terms of retribution, deterrence, or notions of free will or responsibility; and (3) incorporate evidence that is solely statistical in nature as opposed to evidence about biological mechanisms. Even if the cases included references to these theories, we would not be able to say whether it was the biomechanism that affected the sentence, as opposed to concurrently presented information about the defendant's upbringing or immeasurable and selectively documented factors, like his demeanor with the judge. And even if all such concurrently presented information could be reliably assessed and documented, we would still only have an association between them and some set of outcomes. Such evidence could yield correlations consistent with a causal hypothesis, but not evidence for a causal relationship.

Moving to the design of our particular study, we prioritized participant generalizability and concerns of internal validity when we decided to opt for an experimental approach. One hundred and eighty-one judges (N=181) from nineteen states were randomly assigned to one cell of our two-by-two study design. This means they were presented with either information about the (1) presence *or* (2) absence of the biomechanism *as well as* the introduction of the biomechanism by either the (1) prosecution *or* (2) defense.[19] All judges read a hypothetical case, based loosely on *Mobley v. State*.[20] Judges either read (1) that the defendant was diagnosed with psychopathy, or (2) that the defendant was diagnosed with psychopathy, *and* her psychopathy was caused in part by genetic and neurological factors. This evidence was introduced either by the defense or the prosecution. We had two primary dependent measures: (1) ratings of the extent to which the evidence concerning psychopathy mitigated, aggravated, or had no effect on the punishment they would render to the defendant; and (2) the sentence they would render. The findings concerning the mitigating versus aggravating impact of biomechanical evidence were clear: whether the biomechanical evidence was introduced by the defense or the prosecution, it resulted in the evidence concerning psychopathy

19. *See* Aspinwall et al., *supra* note 5, at 846.

20. Mobley v. State, 455 S.E.2d 61 (Ga. 1995).

being rated as significantly less aggravating than in the absence of such information. Importantly, although sentencing outcomes were more complicated to interpret due to high variability by state and the unequal distribution of judges across states, the presentation of evidence concerning a biomechanical cause of psychopathy resulted in significantly lower sentences.

The same 181 judges were asked open-ended questions as to why they responded to each quantitative question concerning mitigation versus aggravation and sentencing in the way that they did. These reasons were coded using content analysis, yielding rich data on their justifications. Notably, when the defense introduced evidence of biomechanism in conjunction with the psychopathy diagnosis, sixty-six percent of judges mentioned a mitigating factor in their reasoning, compared to only thirty-two percent of the judges who read only about the psychopathy diagnosis. Thus, the biomechanism evidence doubled the proportion of judges who mentioned a mitigation factor in their reasoning. When the evidence of biomechanism of psychopathy was presented by the defense, forty-six percent explicitly mentioned the need to balance mitigation against aggravation factors. Of course, we do not know which of these stated reasons motivated answers to the questions about whether the evidence concerning psychopathy was aggravating or mitigating, or which of these reasons played a primary role or any role at all in sentencing decisions, only that the judges we studied volunteered these reasons as relevant to their answers. There are formal tests one could do to examine whether any of these intermediate outcomes mediated in full or in part the effects of the presentation of biomechanism information. We did not conduct these tests, in part because the publication's space limitations precluded reporting whether there were differences between experimental conditions in the particular mitigating versus aggravating reasons the judges mentioned. A second reason we did not conduct these tests was the difficulty comparing sentencing data between states with very different sentencing practices and guidelines.

A. Limitations of Our Experimental Approach to Understanding the Impact of Science in the Courtroom

In the brief discussion space allotted in our article, we acknowledged certain limitations of our study. First, we explained that the facts of the specific case we used both hindered and helped the inferences that we could draw. While the crime was violent and the defendant was portrayed as unsympathetic, this was not a capital case where the sentence could be reduced to life without the possibility of parole. Presenting a capital case would have aligned our vignette with the types of cases where this evidence is most likely to be introduced, but would have rendered arguments about recidivism moot. Additionally, it

would add complexities when we sought to replicate the study in mock jurors. For example, we would need to ensure that online respondents were "death qualified," in that they were not so opposed to the death penalty on principle that they would never sentence someone to death. We also emphasized that, in future studies, researchers could use a diagnosis other than psychopathy, for which there may be real potential for effective treatment. This would complicate the double-edged nature of the sentencing, and make future researchers' presentation of a third rehabilitative theory, with corresponding expert testimony, quite interesting. The judges were also not presented with cross-examination, having only received the expert scientific testimony from either the defense or the prosecution. Cross-examination of varying degrees could potentially increase or decrease the effect of the scientific evidence.

We recognize additional limitations of our study that relate specifically to the diagnosis of psychopathy and the scientific evidence we referenced in the expert testimony presented to judges. First, psychopathy is often mistakenly associated in popular culture with extreme violence, and as a result it is heavily stigmatized. Thus, these results might not generalize to other psychiatric diagnoses associated with antisocial behavior. Our account of the biomechanism of psychopathy was derived from Dr. Blair's neurocognitive model.[21] Employing a different causal description could alter the results. Finally, we combined psychiatric, genetic, and neurobiological science in constructing the expert testimony. Future research should examine the effects of separately introducing testimony on genetic penetrance and expression, neurodevelopment and plasticity, and probabilistic data indicating how much of the relative risk of developing the particular antisocial disorder can be attributed to the given biomechanism. This would allow researchers to test the boundary conditions of the effect, by more precisely pinpointing how the biomechanism might influence judicial reasoning in sentencing in combination with other kinds of evidence.

B. Responses to Professor Denno's Criticisms

Professor Denno takes issue with a number of features of our experimental study. At the heart of her critique, however lies her conclusion that our study is "significantly flawed" in large part because it is experimental rather than archival.[22] Her detailed criticisms suggest that

21. R. J. R. Blair, *Neurocognitive Models of Aggression, the Antisocial Personality Disorders, and Pyschopathy*, 71 J. Neurology, Neurosurgery & Psychiatry 727 (2001).

22. Denno, *supra* note 6, at 1593.

she discredits the value of experimental research in law.[23] We will consider and reply to each in turn.

1. An Atypical Diagnosis

We chose to diagnose our hypothetical defendant with psychopathy.[24] Professor Denno called this "[o]ne of the more inexplicable and questionable aspects" of our study.[25] However, there were in our view several very good reasons for choosing psychopathy. Because at present there is no effective treatment for psychopathy, this diagnosis provided an excellent means for testing which way the double-edged sword cuts. Further, the signature features of psychopathy, such as impulsivity and lack of empathy, could be seen as both mitigating and aggravating. Finally, perhaps because psychopathy's constellation of traits is so idiosyncratic, its genetic and neuroscientific biomechanism has been extensively studied. While we relied on these neurogenetic studies to provide a plausible causal account of psychopathy to the judges, the results of these studies are far from conclusive.

2. An Atypical Case

Professor Denno questioned our choice to make our hypothetical case one of aggravated battery, which would be punished with a prison sentence, rather than a capital case, which would be punished by either a death sentence or life in prison.[26] The reason, however, is quite clear, given our experimental design: A capital case would have dichotomized the sentencing option available to our participating judges—a binary choice between death and life in prison. By using a crime for which there would be much more variation in sentencing (albeit within state established guidelines), we were able to employ a more sensitive continuous measure of sentencing outcomes, thus increasing the statistical power to identify relations between independent and dependent variables.[27]

---

23. *Id.* at 1616 ("The sentencing study's authors may interpret the effects of genetics evidence in their single-hypothetical study as a double-edged sword, but it is not at all clear that there is any support for such a simplistic perspective in actual case law.").

24. Aspinwall et al., *supra* note 5, at 846.

25. Denno, *supra* note 6, at 1596.

26. *Id.* at 1602–03 ("It is unclear why the sentencing study's authors limited Donahue's crime to aggravated battery given that murder is the crime most commonly associated with the use of expert genetic testimony.").

27. *See* Jacob Cohen, *The Cost of Dichotomization*, 7 APPLIED PSYCHOL. MEASUREMENT 249 (1983). Cohen elegantly illustrates that dichotomizing one variable at the mean (not exactly what is happening with the sentencing dichotomy, but useful for illustrating the rationale for using continuous outcome measures) drops the observed correlation between two variables to .798 of its actual value if the underlying variable is actually continuous, with even greater decrements if the variable is dichotomized at more extreme points of its distribution. *Id.* at 249–50. The key issue is loss of measurement information that produces the drop in effect size: "To summarize, the cost in the degradation of measurement due to dichotomization is a loss of one-fifth to two-thirds of the variance

Of course, Profesor Denno is correct that genetic and neurobiological evidence are more likely to be introduced in capital cases; thus, extending this research to capital cases would increase its ecological validity.[28]

3. Atypical Presentation of Sentencing Evidence

Professor Denno questioned the form and manner of presentation of the genetic evidence we included in our hypothetical case.[29] We confined our evidence to a genetic test of the defendant alongside information from an expert scientist concerning what that gene does. Aside from this biomechanism evidence and its presentation by the defense or prosecution, everything else about the complex and detailed vignette was held constant.[30] In the actual cases that Professor Denno has documented, she notes that genetic evidence is often just one part of a more general profile of the defendant, which includes family history, behavioral history, medical history, and information about the environmental exposures of the individual.[31] However, as we have discussed above, and as Professor Denno herself acknowledges,[32] if we had presented the judges with more of the detail that would be found in real cases, such as an account of the defendant's upbringing, we would then not be able to determine whether it was the biological cause that was contributing to the reduced or increased sentence, or the other information. We also could not determine whether the biological causal evidence interacted with some of the other information, having an impact that it would not ordinarily have. Testing these different kinds of evidence in fully crossed experimental combinations (where each independent variable is systematically varied at each level of all other independent variables) is of course possible, but requires considerably larger sample sizes than those available to us. Furthermore, there are other tradeoffs involved in increasing the amount of materials presented to volunteer research participants. Given that our study was conducted online, we were already

---

that may be accounted for on the original variables, and a concomitant loss of power equivalent to that of discarding one-third to two-thirds of the sample." *Id.* at 253; *see also* David L. Streiner, *Breaking Up Is Hard to Do: The Heartbreak of Dichotomizing Continuous Data*, 47 CANADIAN J. PSYCHIATRY 262 (2002).

28. Denno, *supra* note 6, at 1603 ("[M]ost criminal cases addressing behavioral genetics involve capital crimes.").

29. *Id.* at 1610.

30. To review our experimental vignette and other materials that were too voluminous to include in the ultimate publication, please see the Supplementary Online Materials published along with our Article. Lisa G. Aspinwall et al., *Supplementary Materials: The Double-Edged Sword: Does Biomechanism Increase or Decrease Judges' Sentencing of Psychopaths?*, 337 SCI. MAG. 846 (2012), http://science.sciencemag.org/content/suppl/2012/08/15/337.6096.846.DC1.

31. Denno, *supra* note 6, at 1609–10.

32. *Id.* at 1606 ("[I]t can be challenging to isolate the effect of any one piece of mitigating evidence when it comes to interpreting the influences on death penalty sentences.").

demanding the judges' patience, and potentially compromising our response rate, with the lengthy and rich text that was provided. Fortunately, it is relatively straightforward, and potentially extremely useful, to develop testable hypotheses regarding how information about biomechanism might operate in the presence or absence of other kinds of information. Future studies might systematically vary the joint presentation of such information in order to maximize the ecological validity of the study in ways highlighted by Professor Denno's critique, and to identify whether these additional factors amplify or diminish the impact of biomechanical evidence. Thus, programmatically identifying and testing moderating or boundary conditions under which such effects are expected to apply more or less strongly, or even not at all, advances this research. Thus, our failure to present environmental or historical information, for example, is not an inherent limitation of the experimental approach of our study.

Conclusion: The Case for Methodological Pluralism

Our experiment had drawbacks, which we readily acknowledged in the 2500 words we were allotted for publication in the journal *Science*. But if we had completed the study using archival data, or if we had framed the vignette in the way Professor Denno would have preferred, then we would have been unable to answer the very questions that intrigued us. Our study was only the first attempt to isolate the role of biological cause in sentencing decisions by actual judges. While we presented the judges with a detailed and complex vignette, our hypothetical case did not contain or manipulate some other potentially important factors that are frequently found in actual cases. Rather than being limitations, the identification of these other factors in conjunction with careful theorizing about how they may interact, allows for programmatic testing and advancement of theory about the impact of different kinds of causal explanations on sentencing decisions and justifications. Future researchers may then extend this work by manipulating additional or different independent variables alone or in combination with other independent variables. Future researchers could also test whether the outcome is different when one adds genetic and environmental interactions, diagnoses the defendant with a different disorder, increases the severity of the crime, presents different defendant characteristics, and so on. In fact, this is exactly how both theoretical and empirical progress are made—not by seeing particular kinds of data or methods as in competition with one another, but by leveraging different research questions and modalities and recognizing the benefits and drawbacks of each.

It is rarely the goal of any single study to apply to all possible scenarios, or to cover all possible kinds of cases, defendants, diagnoses,

biomechanical evidence, and so on.[33] Thus, it is absolutely appropriate to raise questions—as Professor Denno did—about the generalizability of the findings to other kinds of cases and whether the experimental design of any single study reflects relevant aspects of actual cases. In fact, doing so provides a programmatic contribution to our understanding of how and under what conditions a particular set of effects might occur, which future researchers may use to test, replicate, or extend the original study. However, it should be noted from the outset that these questions Professor Denno poses do not affect concerns about the internal validity of our study—that is, its ability to identify particular causal factors that influence judges' ratings and sentencing decisions in the particular set of conditions and case features tested.

In the future, scholars would do well to distinguish different forms of validity when critiquing the validity of legal research, and resist making assumptions about which form of validity or research strategy is best. The research question often dictates the methodology, and not the other way around. It does little to advance our collective understandings of the value of empirical legal research to levy critiques that reflect confusion over the differences between archival and experimental methodologies. Rather than claiming that our chosen research design is valid and everything else is invalid, researchers should treat archival and experimental methods as iterative and reciprocal, providing complementary sources of information. The insights from archival research may be used to inform the replication and extension of experiments to more real-world conditions or combinations of conditions, and the insights from experimental research may be used to understand when and why a particular kind of scientific evidence may influence actual legal outcomes. As legal researchers continue to develop empirical projects, it is crucial to keep the productive interplay between different methodological approaches in mind. Archival and experimental methodologies are not in competition with each other, but rather present different advantages and disadvantages when it comes to distinct and highly important forms of validity.

33. Douglas G. Mook, *In Defense of External Invalidity*, 38 AM. PSYCHOLOGIST 379 (1983).

***