Dangerous, depraved, and death-worthy: A meta-analysis of the correlates of perceived psychopathy in jury simulation studies

Shannon E. Kelley¹ © | John F. Edens¹ © | Elyse N. Mowle¹ | Brittany N. Penson¹ © | Allison Rulseh¹,2

Abstract

Objectives: Experimental research suggests that legal defendants described as psychopathic are generally, although not uniformly, judged more negatively and punitively. Understanding the correlates of perceived psychopathy, regardless of exposure to mental health evidence, is an important step towards clarifying divergent findings.

Method: We conducted a quantitative synthesis of ten juror simulation studies (combined N = 2,980) examining the meta-analytic association between perceived defendant psychopathy and various psychologically important and legally relevant outcomes.

Results: Perceiving someone as being more psychopathic was associated with viewing that defendant as more dangerous ($r_W = 0.31$) and evil ($r_W = 0.44$). Moreover, perceptions of defendant psychopathy predicted greater support for more adverse consequences in terms of capital sentencing ($r_W = 0.22$) and sentence length ($r_W = 0.27$), although not perceived treatment amenability ($r_W = 0.09$).

Conclusions: These findings highlight the importance of including ratings of perceived psychopathy in experimental designs to identify the circumstances under which psychopathy evidence might prejudicially impact case outcomes.

KEYWORDS

forensic-clinical assessment, meta-analysis, psychopathic personality, stigma
Psychopathy, one of the most extensively researched forms of mental disorder (Patrick, 2018), is typically characterized as a constellation of maladaptive traits, including callousness, manipulativeness, egocentricity, and irresponsibility. Certain characteristics typically associated with psychopathy modestly to moderately predict various forms of criminal behavior, such as violent recidivism following release from incarceration (Douglas, Nikolova, Kelley, & Edens, 2015; Hawes, Boccaccini, & Murrie, 2013; Kennealy, Skeem, Walters, & Camp, 2010; McCulsh, Corrado, Hart, & DeLisi, 2015; Yang, Wong, & Coid, 2010). Largely because of this established relationship with aggression, clinical measurement instruments designed to assess psychopathic traits increasingly have been introduced into legal cases. For example, such instruments are frequently used by mental health expert witnesses and discussed in judicial opinions (Blais, 2015; DeMatteo et al., 2014; Viljoen, MacDougall, Gagnon, & Douglas, 2010; Viljoen, McLachlan, & Vincent, 2010). The disorder is even mentioned by name in some legislation, such as US civil commitment laws regarding sexually violent predators (Edens, Petrila, & Kelley, 2018). Importantly, the types of cases in which mental health experts testify about psychopathy can result in profound consequences for the defendant, involving significant loss of liberty (e.g., indeterminate detention) and even loss of life (i.e., capital punishment).

Perhaps not surprisingly, a growing body of literature suggests that the introduction of mental health expert evidence concerning psychopathic traits significantly impacts legal outcomes for criminal defendants and convicted offenders (e.g., Edens & Cox, 2012; cf. Boccaccini, Turner, Murrie, Henderson, & Chevalier, 2013). For example, in the USA, scores from the Hare Psychopathy Checklist-Revised (PCL-R; Hare, 2003) strongly predict clinicians’ perceptions of defendant dangerousness and statements they make regarding violence risk in court (DeMatteo et al., 2014). In Canada, experts’ PCL-R scores impact judges’ determinations as to whether a defendant will receive a legal designation as a “Dangerous Offender,” which results in an indeterminate period of incarceration beyond that typically given to other convicted felons (see Blais, 2015). In addition, mental health examiners’ scores from Factor 1 of the PCL-R, which taps personality traits such as remorselessness, manipulativeness, and grandiosity, predict whether California parole boards will release prison inmates back into the community (Guy, Kusaj, Packer, & Douglas, 2015).

Despite this widespread use of the PCL-R, there are reasons to be concerned about its introduction into clinical and legal decision-making. Perhaps foremost, a growing body of field literature (see Edens & Boccaccini, 2017) suggests that scores from this instrument are not particularly reliable in applied contexts in which mental health experts are introducing them. Although the professional manual (Hare, 2003) provides intraclass correlation (ICC) statistics across various offender samples in the 0.85 and above range, numerous field studies have suggested that it may be closer to 0.60 (or even lower) in adversarial forensic settings (e.g., Edens, Cox, Smith, DeMatteo, & Sörman, 2015; Miller, Kimonis, Otto, Kline, & Wasserman, 2012). Of particular concern, the aforementioned Factor 1 traits seem to be even less reliable, producing field results in which sometimes more than half the variance in scores is the result of some form of error (Miller et al., 2012). In terms of specific types of error, research indicates that adversarial allegiance is a significant concern, in that mental health experts who are retained by the prosecution produce higher scores than those retained by the defense (Murrie & Boccaccini, 2015; Murrie, Boccaccini, Guarnera, & Rufino, 2013). The individual examiner is important as well, in that some mental health experts on average produce much higher scores than others, regardless of who retains them (Boccaccini, Murrie, Rufino, & Gardner, 2014).

Because of the increased use of mental health experts to diagnose psychopathy (primarily using the PCL-R) in the legal system, since the early 2000s researchers have been investigating the potential stigmatizing effects of this clinical syndrome within civil and criminal justice proceedings. Such research typically examines the impact of a clinician’s diagnosis of psychopathy (and/or statements about the presence of psychopathic traits) by experimentally manipulating the presentation of this information in the context of a mock criminal or civil commitment trial. There are many similarities in the general design of these types of studies concerning
psychopathy, but there are also numerous variations in terms of case characteristics (e.g., capital murder, white collar crime, and juvenile offender), outcome measures (e.g., guilt/innocence verdicts, perceptions of dangerousness, and moral judgments about the defendant), and other important methodological features.

Narrative reviews of this body of literature (Douglas et al., 2015; Edens, Magyar, & Cox, 2013) have argued that when mental health experts describe defendants as being psychopathic, it generally results in mock jurors (a) having more negative attitudes toward these defendants (e.g., perceiving them as more dangerous and evil) and (b) supporting more punitive legal consequences for them (e.g., greater support for capital punishment). The results of the existing experimental studies do not always support such a conclusion, however, with some suggesting that the diagnostic label of psychopathy does not engender greater bias toward a defendant when compared with other conditions, such as no mental health diagnosis, or compared with some other diagnostic label (Boccaccini, Murrie, Clark, & Cornell, 2008; Murrie, Boccaccini, McCoy, & Cornell, 2007; Murrie, Cornell, & McCoy, 2005; see also Saks, Schweitzer, Aharoni, & Kiehl, 2014). Given the variations in the stimulus materials used across these simulation studies, it is perhaps unsurprising that results have not been uniformly clear regarding what impact mental health expert evidence about psychopathy might have on juror decision-making.

It is important to highlight that much of the experimental research on the effects of mental health expert evidence about psychopathy on case outcomes actually blurs two distinct and separable questions. One question is whether jurors' perception of a defendant as more psychopathic actually results in more adverse outcomes for that defendant. The second question is whether mental health expert testimony that a defendant is psychopathic can cause jurors to believe that a defendant actually is more psychopathic. If the answer to both questions is "yes," then one would expect that testimony about psychopathy would result in more adverse outcomes for defendants (for an illustrative example, see Edens, Colwell, Desforges, & Fernandez, 2005).

The first question noted above is the primary focus of this meta-analysis: does perceiving a criminal defendant as being more psychopathic actually result in more adverse outcomes for that defendant (e.g., increased support for a death sentence in a capital murder trial)? Several studies have directly assessed the extent to which participant attributions about a defendant's level of psychopathy—irrespective of any experimental presentation of mental health evidence—predict important criterion measures (e.g., Cox, Clark, Edens, Smith, & Magyar, 2013; Edens, Davis, Fernandez Smith, & Guy, 2013; Guy & Edens, 2006). Direct examinations of the correlates of a defendant's perceived level or degree of psychopathy are important to investigate for several reasons.

First, although "psychopathy" is a mental health concept that is typically introduced at trial to inform violence risk assessment (DeMatteo et al., 2014), research suggests that for laypersons the disorder calls to mind lurid and sensationalistic examples of ostensibly psychopathic figures. In fact, various community surveys indicate that the most commonly identified prototypical "psychopaths" are notorious real-world murderers (e.g., Ted Bundy, Charles Manson, and Jeffrey Dahmer) or fictional killers (e.g., Hannibal Lecter) and that psychopathic individuals tend to be characterized more generally as "monsters" prone to extreme acts of violence (Edens, Clark, Smith, Cox, & Kelley, 2013; Furnham, Daoud, & Swami, 2009; Helfgott, 1997; Keesler & DeMatteo, 2017; Rogers, Dion, & Lynett, 1992; Smith, Edens, Clark, & Rulseh, 2014). To the extent that psychopathy is at issue in a case because of its putative relationship with future violence or crime (e.g., in a capital murder trial), presenting evidence of a disorder known to elicit negative attributions of the defendant that go well beyond the legal question at hand (i.e., outside the scope of violence risk) arguably could be considered unduly prejudicial (DeMatteo & Edens, 2006; DeMatteo, Hodges, & Fairfax-Columbo, 2016). The potentially prejudicial impact of this evidence is further magnified when one considers the aforementioned concerns about poor field reliability for instruments such as the PCL-R. Unreliable psychological assessment data would only exacerbate concerns about undue prejudice caused by mental health expert testimony (i.e., not only is the defendant not similar to Ted Bundy or Hannibal Lecter, he or she may not even genuinely have a very high PCL-R score in the first place).

Another reason why it is informative to investigate this relationship is that it is important to identify the correlates of perceived psychopathy independent of why those perceptions exist. One obvious source contributing to perceptions of a defendant as psychopathic could be effective mental health expert testimony, but jurors clearly
have beliefs and perceptions concerning defendant characteristics (e.g., mental health status) that transcend expert evidence or testimony. Moreover, there is clear evidence that jurors are willing to discount evidence presented by mental health expert witnesses (Boccaccini et al., 2013; Kwartner & Boccaccini, 2008). As such, it seems plausible if not likely that many will disregard expert testimony that purports to tell them what sorts of personality traits a criminal defendant exhibits in any given case.

A third reason why such research is important relates back to the results of experimental simulation studies that have investigated the effects of expert testimony concerning psychopathy and failed to find much if any effect (e.g., Saks et al., 2014). If evidence clearly demonstrates that, in essence, thinking a defendant is a psychopath predicts negative attitudes about and legal consequences for him or her, then why would studies not show that expert testimony concerning psychopathy meaningfully influences these types of outcome measures? Interestingly, many experimental studies in this area of research do not assess directly jurors’ perceptions of the defendant’s degree or level of psychopathy. Rather, they only examine the effects of experimental psychopath/nonpsychopath manipulations on distal outcome measures, such as guilt/innocence verdicts or support for capital punishment.

In such studies, it is unclear whether the experimental manipulation had no effect on jurors or whether it had the intended effect (i.e., resulting in the defendant being perceived as more psychopathic in the experimental condition) but ultimately had no impact on the distal outcome measure (i.e., judging a more psychopathic defendant as more deserving of death). That is, jurors could either believe that (a) psychopathy is irrelevant to whether someone should live or die or that (b) psychopathy is relevant to whether someone should receive a death verdict, but the expert evidence was not sufficiently potent in context to persuade jurors that the defendant actually was more psychopathic (in the experimental condition).

Stimulus materials in, for example, capital murder simulations typically describe a defendant who has committed an egregious crime (or multiple crimes) and who may have a history of serious dysfunction, aggression, and maladaptive behavior. Similarly, stimulus materials for juvenile offender studies often include considerable background information suggestive of a seriously disturbed and aggressive youth. As such, it is plausible if not likely that expert mental health evidence may have relatively little impact in some studies because most jurors in the control conditions already believe a defendant to be a psychopath, without needing to review any expert testimony to confirm their impressions. That is, experimental manipulations could be relatively ineffective because mock jurors in every condition begin with the assumption or belief that criminal defendants are highly psychopathic, resulting in ceiling effects and range restriction on key variables of interest across control and experimental conditions.

1.1 | The present study

Although there have been several individual investigations of the impact of perceived psychopathic traits on juror perceptions of criminal defendants, as well as a few recent narrative summaries of this literature (e.g., Douglas et al., 2015; Edens et al., 2013), to date no quantitative reviews of this area have been undertaken. Accordingly, the present study synthesizes past research using meta-analytic techniques to examine the association between perceptions of a defendant’s level of psychopathy and various attitudinal variables (e.g., perceptions of how dangerous or evil the defendant is) and legal outcome criteria (e.g., sentencing recommendations).

2 | METHOD

2.1 | Study selection

As recommended by Lipsey and Wilson (2001), relevant studies were identified through multiple sources. A comprehensive search of the literature was conducted via PsycINFO, Web of Science, and Proquest Dissertation and Theses as part of a broader review of studies concerning psychopathy and stigma. The searches were
independently conducted by two advanced graduate students by combining terms related to psychopathy and related disorders (e.g., psychopath, psychopathy, conduct disorder, and antisocial personality disorder) with evaluative terms (e.g., jury or juror, pejorative, label*, stigma*, prejudice*). The search included peer-reviewed publications and unpublished theses and dissertations available as of July 2017. The reference sections of selected studies were also scanned to identify other potentially relevant studies. In addition, attempts were made to retrieve any unpublished manuscripts or conference presentations pertaining to the topic, in particular by reviewing the American Psychology-Law Society program for their annual research conference going back to the year 2000. Study authors were solicited for any analyses needed for this meta-analysis that were not provided in the original journal articles.

Following the preliminary search, 38 studies were retained and comprehensively assessed for eligibility. Studies were retained for coding if the participants: (a) reviewed any type of legal case vignette concerning a hypothetical defendant; (b) reported how psychopathic they perceived the defendant to be using some type of rating scale or assessment process, and; (c) provided some form of evaluative judgment (broadly defined, e.g., guilt/innocence verdicts, perceptions of future dangerousness, and sentence recommendations) concerning the defendant. This process ultimately led to the identification of 10 studies meeting inclusion criteria for coding, reflecting a combined sample size of 2,980. Two studies were removed as duplicates, and seven studies were deemed ineligible because they did not use legal case vignettes and/or specifically involve psychopathic personality. Most of the studies that were not retained (n = 19) were excluded because they did not directly measure juror perceptions of the defendant's level or degree of psychopathy, which is the predictor variable of interest in this meta-analysis (e.g., Boccaccini et al., 2008; Murrie et al., 2007). As noted in the introduction, many experimental studies in this area have not included any direct assessment of whether mock jurors perceived the defendant to be psychopathic and instead only focused on distal outcomes such as sentencing recommendations. Table 1 briefly summarizes key methodological features of the studies identified for inclusion. Of note, these studies represent publications by a relatively small and largely overlapping collection of researchers, despite a comprehensive search for relevant work across a variety of formats.

2.2 Coding

Retained studies were each independently coded by two graduate students according to a standardized system. After comparing these independent codings for reliability, a very small number of minor discrepancies were resolved by consulting with the senior author. Studies were coded for sample type (e.g., students, community members, and jurors); sample demographics (e.g., mean age, gender, political orientation, and years of job experience), data collection procedure (e.g., online or in-person, group or individual setting, and compensation for participation), case presentation method (e.g., written vignette, video, oral presentation, and trial transcript), and case material information (e.g., adult or juvenile defendant, crime type, and description of the defendant).

For each study, as many psychopathy predictor variables were coded as possible. Participant ratings of psychopathy across these studies took two basic forms. Across eight studies (e.g., Edens et al., 2005), after reading the case vignette and stimulus materials, participants rated the defendant on a series of 20 individual trait descriptors loosely based on the item content of the PCL-R using a 0 (absent), 1 (possibly present), or 2 (present) scale. The item ratings in the original studies were summed to create a global psychopathy score ranging from 0 to 40. Two distinct subscales were available across these studies as well, reflecting "affective/intpersonal" and "lifestyle/antisocial" features of psychopathy, analogous to the Factor 1 (personality)/Factor 2 (behavioral)

---

*The senior author (J.F.E.) first began investigating this line of research in the early 2000s, at which time an extensive and thorough literature review revealed that no previous studies, including conference presentations, had yet addressed perceptions of psychopathy using the type of manipulation check discussed in this report. Accordingly, there was no need to review conference abstracts before the year 2000 for potentially relevant studies.
distinction on the PCL-R (Hare, 2003). For these studies we coded an effect size for each predictor variable (i.e.,
global psychopathy, affective/interpersonal, and lifestyle/behavioral) separately in our data base.

Instead of, or in addition to, summary variables based on individual item-level ratings, four of the studies relied on
singular Likert-type ratings to assess psychopathy (e.g., Mowle, Edens, Clark, & Sörman, 2016), with participants instructed
to globally rate how psychopathic they perceived the defendant to be after reviewing the case vignette and stimulus
materials. These Likert-type ratings included anchor points ranging from not at all psychopathic to extremely psychopathic.

All available criterion measures regarding perceptions of or judgments about the defendant were coded. These
included attitudinal variables (e.g., Likert-type ratings of how evil and dangerous the defendant was perceived to
be) and legal outcomes such as recommendations regarding conviction (e.g., dichotomous guilt/innocence and civilly
commit/not civilly commit verdicts), sentencing (e.g., sentence length and death verdicts in capital cases), and
placement recommendations (e.g., intensive supervision, secure residential placement, treatment recommendation,
and transfer to adult court). Generally, criterion measures were interchangeable across studies; however, perceived
dangerousness was frequently reported as a composite variable averaging across items assessing the likelihood of
the defendant committing specific type of crimes and/or posing a risk of violence in specific settings.

Some studies were purely correlational (i.e., the same case vignette was presented to each participant), but
others examined correlates of perceived psychopathy within the context of experimental manipulations (see
Table 1). For these, relevant information regarding control and experimental conditions was coded. For the
purposes of this meta-analysis, we classified all cells in which the defendant was described by a mental health
expert as psychopathic as "experimental" conditions. Cells in which there was no information depicting the
defendant as psychopathic were classified as "control" conditions, which included a few cells ascribing other mental
health problems to the defendant (e.g., schizophrenia).²

Separate effect sizes were coded for the total sample and for individual control and experimental conditions
when applicable. The primary reason for examining experimental and control conditions separately was because of
concerns that there might be significant range restriction on psychopathy ratings in experimental conditions, which
might attenuate the magnitude of effects for the criterion measures. As noted in the introduction, it is plausible that
hearing a defendant described as being highly psychopathic in an experimental condition might result in ceiling
effects on juror ratings of psychopathy, which would suppress their degree of association with criterion measures.
Although this concern ultimately was not borne out, we report these results separately. We also report aggregated
analyses that combine all individual control and experimental conditions within each specific study.

3 | RESULTS

3.1 | Preliminary review and analytic plan

A total of 325 effect sizes were coded across all predictor and criterion measure combinations available across all
control, experimental, and combined conditions represented in these 10 studies. Preliminary investigation of these
effect sizes led us to several initial conclusions regarding how best to summarize these data. The first decision made
after completing a preliminary review of our data was to restrict the analyses of our criterion measures to only
those with enough independent effect sizes (from three or more separate studies) to justify the computation of a
weighted mean effect. Although this resulted in some theoretically interesting outcome variables being excluded (e.
ge.g., guilt/innocence verdicts and civil commitment recommendations), ultimately there was a sufficient number of
unique effect sizes to examine the relationship between perceptions of psychopathy and perceptions of how (a)
dangerousness and (b) evil the defendant was, as well as (c) recommended sentence length, (d) support for death
verdicts in capital punishment cases, and (e) perceptions of treatment amenability.

²Excluding the control conditions in which the defendant was described as experiencing (nonpsychopathic) mental health problems had essentially no
effect on the weighted mean effect sizes reported later in the Results section.
<table>
<thead>
<tr>
<th>Citation</th>
<th>Defendant</th>
<th>Case vignette</th>
<th>Participants</th>
<th>N*</th>
<th>Experimental design</th>
<th>Psychopathy rating method</th>
<th>Criterion measure*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox et al. (2013)</td>
<td>Adult murderer</td>
<td>Capital punishment</td>
<td>Jury panel community members</td>
<td>304</td>
<td>No</td>
<td>Affective/interpersonal</td>
<td>4</td>
</tr>
<tr>
<td>Cox, Edens, Rulseh, and Clark (2016)</td>
<td>White collar criminal</td>
<td>Criminal sentencing</td>
<td>Jury panel community members</td>
<td>268</td>
<td>No</td>
<td>Affective/interpersonal and Likert</td>
<td>2 and 3</td>
</tr>
<tr>
<td>Davis (2003)</td>
<td>Adult murderer</td>
<td>Capital punishment</td>
<td>College students</td>
<td>425</td>
<td>Yes</td>
<td>Affective/interpersonal</td>
<td>1 and 4</td>
</tr>
<tr>
<td>Edens et al. (2013)</td>
<td>Adult murderer</td>
<td>Capital punishment</td>
<td>Jury panel community members</td>
<td>286</td>
<td>No</td>
<td>Affective/interpersonal</td>
<td>1 and 2</td>
</tr>
<tr>
<td>Edens et al. (2005)</td>
<td>Adult murderer</td>
<td>Capital punishment</td>
<td>College students</td>
<td>203</td>
<td>Yes</td>
<td>Affective/interpersonal</td>
<td>1 and 4</td>
</tr>
<tr>
<td>Edens et al. (2003)</td>
<td>Juvenile murderer</td>
<td>Capital punishment</td>
<td>College students</td>
<td>360</td>
<td>Yes</td>
<td>Affective/interpersonal</td>
<td>1, 4, and 5</td>
</tr>
<tr>
<td>Edens et al. (2017)</td>
<td>Juvenile offender</td>
<td>Case disposition</td>
<td>Jury panel community members</td>
<td>326</td>
<td>Yes</td>
<td>Likert rating</td>
<td>1, 2, and 5</td>
</tr>
<tr>
<td>Guy and Edens (2006)</td>
<td>Adult sex offender</td>
<td>Civil commitment</td>
<td>College students</td>
<td>405</td>
<td>Yes</td>
<td>Affective/interpersonal</td>
<td>1 and 5</td>
</tr>
<tr>
<td>Mowle et al. (2016)</td>
<td>Violent offender</td>
<td>Verdict &amp; criminal sentencing</td>
<td>Jury panel community members</td>
<td>359</td>
<td>Yes</td>
<td>Likert rating</td>
<td>3</td>
</tr>
<tr>
<td>Rulseh, Edens, and Cox (2017)</td>
<td>White collar criminal</td>
<td>Criminal sentencing</td>
<td>MTurk workers</td>
<td>330</td>
<td>Yes</td>
<td>Affective/interpersonal and Likert rating</td>
<td>2 and 3</td>
</tr>
</tbody>
</table>

Note. Sample sizes reported here may differ from the total Ns in the original publications due to certain cells being irrelevant to the focus of this meta-analysis as well as small amounts of missing data on some criterion measures.

*1 = dangerousness/recidivism risk, 2 = evil, 3 = sentence length, 4 = death verdict, 5 = treatment recommendation.

*These two publications were based on the same data set but reported analyses for separate criterion measures.
For effect sizes meeting these inclusion criteria, we next addressed how to organize analyses given the availability of several predictor variables. When studies provided ratings of the defendant on the 20 items loosely based on PCL-R criteria, we chose to focus our analyses on affective/interpersonal ratings of psychopathy. This is generally consistent with the focus of the original studies published in this area of research (e.g., Edens et al., 2005) and aligns with layperson perceptions that these personality traits are most prototypical of a "psychopath" (Hoff, Rypdal, Mykletun, & Cooke, 2012; Smith et al., 2014; see also Rogers et al., 1992). These prototypicality studies have indicated that laypersons generally view affective and interpersonal features (e.g., remorseless, deceitful, and manipulative) as the most characteristic traits of psychopathy, and perceive behavioral aspects (e.g., irresponsible and impulsive) as much less salient to the construct. Moreover, global psychopathy ratings comprised of all 20 items generally produced effect sizes with criterion measures that were highly similar if not virtually identical to those for the affective/interpersonal ratings in isolation, indicating that these analyses were largely redundant with those we report below for the affective/interpersonal traits. Also, "lifestyle/behavioral" traits generally have produced weaker effect sizes compared with affective/interpersonal features in the prediction of relevant criterion measures (e.g., death verdicts and perceptions of "evil").

As mentioned, some studies incorporated single-item Likert-type ratings of psychopathy as predictor variables; however, we gave preference to effect sizes reported for affective/interpersonal features when available. To maximize the total number of samples available for analysis, we included effect sizes for Likert-type ratings when this item was the only assessment of perceived psychopathy available in a study. Given concerns about possibly combining "apples and oranges" in such analyses, we also computed results based only on one type of predictor (i.e., affective/interpersonal or Likert-type rating).

Coded effect sizes (r) were corrected for sample size bias according to formulas from Lipsey and Wilson (2001). Any effect sizes not reported as correlations (e.g., Area Under the Curve values) were converted to r using standard statistical formulas.

### 3.2 | Primary analyses

Table 2 summarizes the random-effects meta-analytic results of psychopathy ratings on the five primary criterion measures examined across this body of literature. To assess the consistency of effect sizes contributing to each analysis, we report Cochran’s Q. However, because our meta-analysis contains a relatively small number of studies, Cochran’s Q can suffer from insufficient power to detect true heterogeneity and may be misleading (i.e., a nonsignificant finding does not necessarily provide evidence of homogeneity). We therefore also report $I^2$ (Higgins, Thompson, Deeks, & Altman, 2003), which provides a relative measure of inconsistency that can be interpreted and compared between groups with a differing number of effect sizes. The $I^2$ statistic describes the proportion of variance across observed effects that is due to heterogeneity in true effects using a standard metric of 0–100%. That is, $I^2$ reflects the proportion of observed variance that would remain if sampling error were somehow removed. Categorization of the $I^2$ statistic as representing low (25%), moderate (50%), or high (75%) levels of heterogeneity is common; however, recent criticisms stress that $I^2$ is not an absolute value of how much effects vary (Borenstein, Higgins, Hedges, & Rothstein, 2017). To facilitate an approximate understanding of the actual dispersion of true effects, we thus provide the range of observed effects alongside the $I^2$ statistic (Table 2).

### 3.2.1 | Effect on perceptions of dangerousness

Table 2 provides a summary of the meta-analytic associations between perceptions of psychopathy and our five criterion measures. First, we obtained mean weighted effect sizes ($r_m$) for ratings of psychopathy (i.e., affective/
interpersonal features or single-item Likert-type ratings) on perceptions of defendant dangerousness. When exclusively analyzing control conditions, the mean weighted effect size was 0.35, with significant variability observed among effects. The mean weighted effect size was somewhat weaker (0.26) for experimental conditions, although these effects were considerably more homogeneous. Not surprisingly, the mean weighted effect size combining across groups was intermediate between these two (0.31) and findings again suggested a significant amount of heterogeneity among effects. We next conducted a meta-analytic analog to an analysis of variance (MetaF) to determine whether condition type explained the variability in effect sizes for perceived dangerousness. This analysis provided some modest support for a moderating effect of vignette condition, with $Q_4$ approaching significance ($3.16, p = 0.08$).

Given that effect size variability was not sufficiently explained by condition type, we investigated whether combining the two different methodologies for assessing perceived psychopathy (i.e., affective/interpersonal ratings and single-item Likert-type ratings) contributed to the observed heterogeneity. After removing the small number of effect sizes based on Likert-type ratings ($k = 4$), the mean weighted effect size for the remaining samples across condition type was largely unaffected (0.30); however, the composition of effects appeared relatively more homogeneous ($Q = 21.19, p = 0.17$; range = 0.02–0.72; $I^2 = 24.5\%$).

3.2.2 Effect on perceptions of being "evil"

For ratings of psychopathy on perceptions of the defendant as "evil" within control conditions, the mean weighted effect size was moderately strong (0.42) and observed variability among effect sizes was modest. Findings from experimental conditions were generally consistent, with a mean weighted effect size of 0.46, though slightly more heterogeneous. Overall, the mean weighted effect size across condition types was 0.44 and effect sizes were reasonably consistent. Given the apparent stability of findings, analysis of condition type as a moderating variable was unnecessary. However, it is worth noting that removing effect sizes based on Likert-type ratings ($k = 4$) from analyses modestly attenuated the mean weighted effect size (0.38), and slightly decreased variability ($Q = 7.22, p = 0.30$; range = 0.24–0.54; $I^2 = 30.8\%$). By contrast, using all the available effect sizes for Likert-type ratings

### Table 2: Association between perceived psychopathic traits and criterion measures

<table>
<thead>
<tr>
<th></th>
<th>Combined n</th>
<th>k</th>
<th>ES</th>
<th>95% CI</th>
<th>Q</th>
<th>$p$</th>
<th>Range</th>
<th>$I^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dangerousness/recidivism risk</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>973</td>
<td>12</td>
<td>0.25</td>
<td>0.25–0.44</td>
<td>21.83</td>
<td>0.03</td>
<td>0.08–0.72</td>
<td>49.6</td>
</tr>
<tr>
<td>Experimental</td>
<td>670</td>
<td>9</td>
<td>0.26</td>
<td>0.18–0.33</td>
<td>6.23</td>
<td>0.62</td>
<td>0.02–0.46</td>
<td>0.0</td>
</tr>
<tr>
<td>Combined</td>
<td>1,643</td>
<td>21</td>
<td>0.31</td>
<td>0.24–0.37</td>
<td>32.00</td>
<td>0.04</td>
<td>0.02–0.72</td>
<td>37.5</td>
</tr>
<tr>
<td><strong>Evil</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>908</td>
<td>7</td>
<td>0.42</td>
<td>0.34–0.51</td>
<td>8.54</td>
<td>0.20</td>
<td>0.29–0.58</td>
<td>29.7</td>
</tr>
<tr>
<td>Experimental</td>
<td>302</td>
<td>4</td>
<td>0.46</td>
<td>0.30–0.61</td>
<td>5.77</td>
<td>0.12</td>
<td>0.24–0.61</td>
<td>48.0</td>
</tr>
<tr>
<td>Combined</td>
<td>1,210</td>
<td>11</td>
<td>0.44</td>
<td>0.36–0.51</td>
<td>14.75</td>
<td>0.14</td>
<td>0.24–0.61</td>
<td>32.3</td>
</tr>
<tr>
<td><strong>Sentence length</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>631</td>
<td>5</td>
<td>0.24</td>
<td>0.07–0.41</td>
<td>14.63</td>
<td>0.01</td>
<td>0.03–0.39</td>
<td>72.7</td>
</tr>
<tr>
<td>Experimental</td>
<td>326</td>
<td>3</td>
<td>0.31</td>
<td>0.18–0.43</td>
<td>2.54</td>
<td>0.28</td>
<td>0.18–0.44</td>
<td>21.3</td>
</tr>
<tr>
<td>Combined</td>
<td>957</td>
<td>8</td>
<td>0.27</td>
<td>0.16–0.37</td>
<td>17.80</td>
<td>0.01</td>
<td>0.03–0.44</td>
<td>60.7</td>
</tr>
<tr>
<td><strong>Death verdicts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>687</td>
<td>9</td>
<td>0.26</td>
<td>0.14–0.36</td>
<td>18.70</td>
<td>0.02</td>
<td>0.02–0.45</td>
<td>57.2</td>
</tr>
<tr>
<td>Experimental</td>
<td>557</td>
<td>6</td>
<td>0.15</td>
<td>0.03–0.27</td>
<td>9.78</td>
<td>0.08</td>
<td>0.05–0.57</td>
<td>48.9</td>
</tr>
<tr>
<td>Combined</td>
<td>1,244</td>
<td>15</td>
<td>0.22</td>
<td>0.12–0.31</td>
<td>33.93</td>
<td>&lt;0.01</td>
<td>0.05–0.57</td>
<td>58.7</td>
</tr>
<tr>
<td><strong>Treatment amenability</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>614</td>
<td>7</td>
<td>0.12</td>
<td>-0.05–0.29</td>
<td>25.14</td>
<td>&lt;0.01</td>
<td>0.19–0.34</td>
<td>76.1</td>
</tr>
<tr>
<td>Experimental</td>
<td>462</td>
<td>5</td>
<td>0.05</td>
<td>-0.05–0.14</td>
<td>1.87</td>
<td>0.76</td>
<td>0.07–0.14</td>
<td>0.0</td>
</tr>
<tr>
<td>Combined</td>
<td>1,076</td>
<td>12</td>
<td>0.09</td>
<td>-0.01–0.19</td>
<td>29.15</td>
<td>&lt;0.01</td>
<td>0.19–0.34</td>
<td>62.3</td>
</tr>
</tbody>
</table>
(k = 10) instead of giving preference to affective/interpersonal ratings increased the mean weighted effect size to 0.54, with an appreciable corresponding decrease in variability (Q = 8.22, p = 0.51; range = 0.43–0.77; \(I^2 = 2.7\%\)).

3.2.3 | Effect on sentence length recommendations

We next meta-analyzed effect sizes for perceived psychopathy on sentence length recommendations. The mean weighted effect size for control conditions was 0.24, although substantial heterogeneity was observed. For experimental conditions, the mean weighted effect size actually improved modestly to 0.31, with the three effect sizes comprising this meta-analysis having the same positive direction and statistically appearing rather consistent. Combining across condition types, the mean weighted effect size was 0.27 and these aggregated effect sizes were largely heterogeneous. Findings from MetaF suggested that effect sizes for sentence length recommendations were not significantly moderated by condition type (\(Q_B = 0.40, p = 0.53\)). Therefore, we investigated whether effect sizes differed according to the psychopathy rating procedure. When only affective/interpersonal ratings (k = 6) were analyzed, the mean weighted effect size was 0.33, with effect sizes appearing generally homogeneous (Q = 5.17, p = 0.40; range = 0.18–0.44; \(I^2 = 22.6\%\)). The findings based on effect sizes for only Likert-type ratings (k = 8) suggested a weaker mean weighted effect size (0.24), although this was concomitant with significant variability (Q = 14.45, p = 0.04; range = 0.03–0.58; \(I^2 = 59.5\%\)).

3.2.4 | Effect on death penalty verdicts

With respect to dichotomous death penalty verdicts, effect sizes were only available using affective/interpersonal ratings as the predictor variable. The mean weighted effect size within control conditions was 0.26 (AUC = 0.65), although there was sizeable inconsistency among effects. This heterogeneity was similarly observed among effects from experimental conditions, which demonstrated a somewhat attenuated mean weighted effect size of 0.15. Overall, the mean weighted effect size across condition types was 0.22 (AUC = 0.63) and, as would be expected from the preceding analyses, effects were moderately heterogeneous. However, findings from a meta-analytic analog to an analysis of variance did not support condition type as a significant moderator of effect size (\(Q_B = 1.89, p = 0.17\)).

3.2.5 | Effect on perceptions of treatment amenability

Finally, we examined the mean weighted effect sizes for ratings of psychopathy on perceptions that the defendant was likely to benefit from treatment. For control conditions, the mean weighted effect size was small and nonsignificant (0.12). The Q and \(I^2\) statistics, however, were indicative of considerable variability among these effect sizes. The effect sizes from experimental conditions appeared substantially more homogeneous, yet paralleled the above mean weighted effect size in terms of overall nonsignificance (0.05). Combining across condition type, the mean weighted effect size was again negligible (0.09) and comprised of highly differing effect sizes. We investigated the influence of condition type on effect sizes using MetaF; however, this variable did not emerge as a significant moderator (\(Q_B = 0.66, p = 0.42\)). After removing effect sizes based on Likert-type ratings (k = 4), the mean weighted effect size across condition type remained nonsignificant (0.12) with no meaningful change in the extent of variability (Q = 18.73, p < 0.01; range = −0.07–0.34; \(I^2 = 68.0\%\)).

3.2.6 | Effect of publication bias

Despite searching for unpublished theses, dissertations, conference presentations, and manuscripts, our findings may be subject to publication bias. To address this concern, we computed the fail-safe N for each weighted mean effect size. These statistics support the stability of our findings, ranging from 27 to 1,179 using Rosenthal's (1979) early approach,
and from 15 to 46 using the more conservative Orwin (1983) method. For example, 44 studies with a mean effect size of zero would be required to reduce the association between perceived psychopathy and evilness to a negligible magnitude (0.10), and 979 would be necessary to completely nullify the effect. Based on these values, the significance of observed weighted mean effects is likely to be robust against potentially missing studies.

4 | DISCUSSION

The meta-analytic results across these 10 studies indicate that perceptions of psychopathy have a significant effect on ratings of how dangerous and evil a defendant is perceived to be and on recommendations for death verdicts and sentence length. There was appreciable heterogeneity across several weighted effect sizes but the effects were somewhat more homogenous after accounting for differences in measurement of our predictor variables (e.g., single-item Likert-type ratings vs. 8-item affective/interpersonal ratings). Condition type (control or experimental) somewhat surprisingly did not emerge as a significant moderator of effect sizes, indicating that the magnitude of this association was generally similar regardless of whether evidence of psychopathy was presented as part of the stimulus materials reviewed by the participants. As such, concerns about possible ceiling effects on the predictor variables in experimental conditions were not borne out. Of note, the aggregated effect size for treatment amenability was nonsignificant, although this finding is complicated by the considerable inconsistency observed (even after partitioning effects by predictor type), which may be attributable to the diversity of study characteristics for this criterion (e.g., offense type, juvenile vs. adult).

4.1 | Implications for experimental research on psychopathy

The fairly robust effect sizes across these studies for examined outcomes (with the exception of treatment amenability) raise important questions about the designs of previous studies that experimentally manipulated psychopathy evidence and found no significant effect on adverse outcomes for the defendant (e.g., Boccaccini et al., 2008; Murrie et al., 2007; Saks et al., 2014). Our findings support the possibility that previous studies that have failed to find significant differences in negative attitudes between psychopath/nonpsychopath conditions may not have sufficiently evoked differing perceptions of the defendant’s personality. Put differently, when mental health expert testimony does not contribute to adverse consequences, it may be because this type of evidence has not strongly altered perceptions of psychopathy for a given case, rather than because perceiving a defendant to be a psychopath does not evoke negative attitudes and adverse case outcomes among jurors.

Findings from the control conditions in our meta-analysis provide a particularly compelling argument for the need to assess perceived psychopathy in these studies. The aggregated effect sizes summarized in Table 2 clearly indicate that participants who think a defendant is psychopathic tend to have more negative and punitive attitudes toward that defendant than do participants perceiving said defendant as much less psychopathic, regardless of whatever mental health information they were presented. Our results highlight the importance of assessing the strength of manipulations in experimental studies to determine whether the control and various experimental conditions actually produce differences in how psychopathic the defendant is perceived to be. As such, it would be informative if studies that have produced null results for expert mental health testimony were replicated with the addition of a manipulation check to compare how different case vignette content affects participant perceptions of psychopathy. Including ratings of perceived psychopathic features in experimental designs also would provide the opportunity to examine mediational hypotheses for any effects of psychopathy manipulations on legally relevant outcomes. This procedure also may be valuable for research using experimentally manipulated vignettes to investigate the effect of other psychological diagnoses and characterizations on various outcomes (e.g., stigma surrounding depression).
Our results also provide some clarification concerning the influence of the diagnostic label of "psychopath" versus the personality features associated with psychopathy. The observed effects were evident regardless of whether participants indicated the extent to which the defendant was globally a "psychopath" or provided ratings on specific core psychopathic traits. There were modest differences in mean effect size magnitude between these two different methods of quantifying participant perceptions of psychopathy, which explained some of the heterogeneity among reported effect sizes; however, the same basic pattern of findings emerged whether participants provided ratings of the defendant on a single Likert-type question concerning global psychopathy or they rated the defendant on individual psychopathic traits that were not explicitly presented alongside this diagnostic term.

4.2 Implications for the use of psychopathy evidence in legal proceedings

Although the results of our meta-analysis indicate that perceptions of psychopathy are associated with negative attitudes towards a defendant and support for punitive legal outcomes, they do not necessarily support the argument that psychopathy evidence produces undue prejudice, in a legal sense. For example, under circumstances in which perceptions of future dangerousness and risk of recidivism inform decision-making, research indicates that it is not unreasonable to assume that a psychopathic defendant generally poses a greater risk of violence compared with a nonpsychopathic defendant (e.g., Blais, Solodukhin, & Forth, 2014; Yang et al., 2010).

That being said, to the extent that legal outcomes are being influenced by perceptions associated with defendant psychopathy beyond those empirically justified (e.g., heightened risk of violence in certain settings), ascribing these characteristics to a defendant could be unduly prejudicial. Concerns about stigmatization may be especially applicable to capital cases given that psychopathy testimony predicts death verdicts even after accounting for juror ratings of perceived dangerousness (Edens et al., 2005). This finding suggests that other characteristics associated with psychopathy, which may or may not be probative to the relevant legal issue, are contributing to more punitive sentencing. For example, modest to moderate associations with violent recidivism notwithstanding, the robust correlations between psychopathy perceptions and ratings of being "evil" are concerning and possibly contribute to undue prejudice. Equating "psychopathic" with being evil highlights longstanding concerns that psychopathy—particularly among laypersons—may be a "moral judgement masquerading as a clinical diagnosis" (Blackburn, 1988, p. 511; see also, Morse, 1978, 2014).

Although the mean effect sizes for correlates of perceived psychopathy do indicate that these perceptions negatively affect attitudes towards defendants and adversely influence sentencing decisions, whether psychopathy evidence should be inadmissible because of "undue prejudice" is a determination that ultimately lies in the hands of the courts. Despite the association between perceived psychopathic traits and adverse consequences, case law reviews (DeMatteo et al., 2014) indicate that the admissibility of expert mental health evidence concerning psychopathy is rarely challenged and, even when challenged, is rarely objected to on the basis of undue prejudice. Psychopathy evidence has been argued to have limited probative value for certain legal proceedings (e.g., capital sentencing), and defense attorneys may consider raising objections to mental health experts providing such evidence in these contexts given the empirically-supported possibility of stigmatization. Of course, the courts also should consider the expanding body of field research suggesting that mental health experts who are testifying about instruments such as the PCL-R produce highly unreliable scores, which raises additional concerns about their prejudicial impact.

Another important question to consider in interpreting the current results is whether findings are specific to mock juror perceptions of psychopathy or represent a more general pattern of stigmatization associated with mental disorder. Relatively few studies have investigated this question, but mock juror ratings of other forms of psychopathology (e.g., borderline personality traits and psychosis) do not appear to significantly predict the types of legal outcomes investigated in this meta-analysis. These other psychological characterizations have been found to be relatively mitigating or entirely unrelated to punitive decision-making, including perceptions of defendant
guilt, death verdicts, and sentence length (Cox et al., 2013; Edens et al., 2005; Gurley & Marcus, 2008; Mowle et al., 2016). It has been argued by some (e.g., Frick & Nigg, 2012) that psychopathic traits are not especially more stigmatizing than other types of mental health information, but the results of this meta-analysis in combination with findings from the individual studies noted above would seem to contradict such a claim. This is not entirely surprising, given that the component features of psychopathy (e.g., remorselessness) do little to elicit sympathy or alleviate perceived culpability, whereas symptoms such as hallucinations may make defendants seem at least somewhat less blame-worthy (but see Aspinwall, Brown, & Tabery, 2012).

4.3 Limitations and directions for future research

The conclusions reached about correlates of perceived psychopathy from these aggregated effect sizes are qualified by the usual limitations of mock jury studies. Some of the studies included in this review relied on college student samples (see Table 1). However, a recent meta-analytic review of the literature indicates that there are few differences in study results between student mock juror samples and more ecologically valid methodologies (Bornstein et al., 2017). Another limitation related to ecological validity is that the majority of studies included in this review used written stimulus materials (e.g., trial transcripts and brief descriptions of expert testimony) rather than more elaborate methods such as videotaped testimony. It is possible that using more elaborate presentations of mental health expert evidence to increase the resemblance of the study to actual juror experiences could influence participants’ perceptions of defendant psychopathy and/or the relative weight of these perceptions in relation to legally relevant outcomes. In addition, the nature of these studies precludes the ability to draw any conclusions about causal relationships. Other variables (e.g., authoritarianism) could conceivably account for both participants’ perceptions of psychopathy and participant attitudes and decision-making.

The ability to examine potential moderators was limited by the small number of studies available for many analyses. For some criterion variables, the particulars of stimulus materials were fairly similar (e.g., death penalty cases), whereas psychopathy was considered for very different crimes and sanctioning options among other groupings (e.g., dangerousness), which may account for heterogeneity among effect sizes. Whether the extent and impact of perceived psychopathy depend on contextual features of a case remains a question for future experiments or meta-analysis. Furthermore, the limited variability in how psychopathy was described in these studies prevented analyses of whether describing a defendant using the label of “psychopath” versus psychopathic traits impacted effect sizes. The differential implications of describing a defendant with psychopathic traits, a label of psychopathy, or a combination of both has been discussed at length in the literature focusing on experimental designs (e.g., Boccaccini et al., 2008; Filone, Strohmaier, Murphy, & DeMatteo, 2014; Murrie et al., 2007); however, the studies included in this review typically described the defendant using the psychopathy label and traits together, or intentionally omitted information about personality and mental health status. A few included studies using an experimental design contained conditions in which only psychopathic traits were ascribed to the defendant (e.g., Edens, Guy, & Fernandez, 2003; Edens, Mowle, Clark, & Magyar, 2017). We provide preliminary evidence based on comparisons of predictor variable rating methods that perceptions of global psychopathy and specific psychopathic traits have generally comparable effects; however, future research using experimental conditions that isolate the effect of psychopathic traits and the diagnostic label are needed before conclusions can be drawn about how the communication of psychopathy evidence influences perceptions of psychopathy or moderates their effect on criterion measures.

The possibility of a “file-drawer” problem is an important consideration when conducting meta-analyses and is particularly noteworthy given that we identified relatively few studies meeting selection criteria, and that these studies represent the work of a small and overlapping cluster of researchers. Fail-safe N analyses suggested that our findings were not solely an artifact of publication bias; however, certain consistencies within this study group (e.g., geographic location of participants and measurement approaches) may limit the generalizability of observed effects. For example, the operationalization of psychopathy for constructing experimental vignettes and assessing
participant perceptions was relatively narrow, only allowing for some comparison of global ratings and adapted PCL-R affective/interpersonal ratings. Future investigation is needed to establish whether findings extend to characterizations and measurement based on alternative models of psychopathy (e.g., triarchic model, Patrick, Fowles, & Krueger, 2009). Relatedly, the "scope of stigma" is not yet clear with respect to other variants of socially aversive personality (e.g., narcissism and Machiavellianism). Our hope is that this meta-analysis inspires independent and diverse groups to pursue further research in this area, such that future reviews may empirically determine any moderating effect of study authors, among other factors.

5 | CONCLUSIONS

In summary, the results of this meta-analysis reinforce concerns about the potentially stigmatizing effects of expert mental health evidence concerning psychopathy on legal decision-making and highlight uncertainties about the interpretation of previous experimental research investigating such effects. To the extent that jurors perceive a defendant as relatively more psychopathic when exposed to evidence of psychopathy than they otherwise would, this type of testimony will clearly be associated with more negative judgments and punitive outcomes for such a defendant. The inclusion of manipulation checks in future experimental mock juror studies of this nature is therefore essential to valid interpretation of diagnostic labeling effects and direct testing of mediation hypotheses. Given the gravity of legal situations in which psychopathy may be introduced, we strongly encourage replication efforts that measure perceived psychopathy and thus clarify when and how such mental health evidence influences juror attitudes and decision-making in a stigmatizing way.

ORCID

Shannon E. Kelley  http://orcid.org/0000-0001-6922-5505
John F. Edens  http://orcid.org/0000-0002-2218-0650
Brittany N. Person  http://orcid.org/0000-0002-5199-520X

REFERENCES


Rulseh, A., Edens, J. F., & Cox, J. (2017). Triarchic model personality traits and their impact on mock juror perceptions of a white-collar criminal. Triarchic model personality traits and their impact on mock juror perceptions of a white-collar...