The Depravity Standard III: Validating an evidence-based guide

Michael Welner[a,*], Kate Y. O'Malley[a], James Gonidakis[a], Alisha Saxena[b,1], Jada Stewart-Willis[a]

[a] The Forensic Panel, 224 West 30th Street, Suite 806, New York City, NY, 10001, United States
[b] Teachers College, Department of Clinical Psychology, 428 Horace Mann, New York, NY 10027, United States

ARTICLE INFO

ABSTRACT

Purpose: The Depravity Standard is an evidence-based guide developed to operationalize an approach to distinguish the worst of crimes in a consistent manner that minimizes bias. This phase of the research was designed to validate the Depravity Standard items and develop a scoring mechanism.
Methods and results: Inter-rater reliability was performed by two groups of trained raters, with each of the 25 Depravity Standard items finding high agreement. To distinguish the relative severity of each item as they may occur in a murder case, an online public survey was devised. U.S. participants (n = 1273) rated each item on a scale of 1–100 (100 = most depraved). The items were then applied to 770 case files of adjudicated murder convictions to establish content validity. 582 cases were retained for further analysis, and merged with survey data to establish a percentile scoring system.
Conclusions: The Depravity Standard is validated for application to murder cases to inform the presence or absence of the 25 items of depravity. It enables assessment of relative depravity of a perpetrator's intent, victim choice, actions, and attitudes. Application of the Depravity Standard relies on evidence, minimizes bias and prejudice, and promotes fairness in sentencing and release decisions.

1. Introduction

Codes of criminal law in America have established terms such as "heinous," "atrocious," "cruel," "depraved," "vile," and other analogues of evil to refer to the worst of crimes. The law uses these terms to distinguish those crimes that warrant more severe sentences. HAC (heinous, atrocious, cruel) aggravators, as they are commonly known in legal parlance, were specifically devised for application to murder cases.

Over the years, HAC aggravators have been consistently challenged in court for their vague terminology and resulting contribution to arbitrary sentencing. Higher court opinions, however, have upheld the use of HAC aggravators. At the same time, court decisions have struggled to bring clarity and consistency to these distinctions.

In *Gregg v. Georgia* (1976), the United States Supreme Court upheld the Georgia aggravator of "heinous, atrocious, and cruel" as constitutional but expressed concern that a jury would have difficulty deciding this issue. Writing for the majority, Justice Potter Stewart noted that the problem of jury inexperience could be "alleviated if the jury is given guidance regarding the factors about the crime and the defendant that the state, representing organized society, deems particularly relevant to the sentencing decision." (at 192). In *Godfrey v Georgia* (1980), the

Court determined that the death penalty must not be imposed by "standardless" sentencing, and that jurors require instruction in the otherwise ambiguous HAC language. Then, in *Walton v. Arizona* (1990), the Court clarified that aggravating factors needed to be identified through objective circumstances.

Even reliance upon objective circumstances, however, does not ensure that the interpretation of what is a depraved murder is not arbitrary or vague. *Norris v. State* (1999) involved a case of three bar patrons who were shot, one after the other, while sitting in a booth. The Alabama Court of Appeals was divided over whether the victims suffered "psychological torture" because they were aware they were going to die. The majority opinion ruled that the murders happened quickly enough that they did not fit the "torture" criteria. What is quickly enough? Without direction, the trier of fact must rely on visceral judgments which are vulnerable to bias and may provide different interpretations of the law from one case to the next. Specificity safeguards against the consequences of vagueness.

Courts' emerging recognition of the importance of substantive and evidence-driven arguments on HAC was illustrated in *Dixon v. Ryan* (2016). The District Court in *Dixon* upheld a finding of "cruel, heinous and depraved" because the prosecutor argued specific history and evidentiary findings at trial in support of the aggravator. However,

---

* Corresponding author.
  E-mail address: drwelner@forensicpanel.com (M. Welner).
[1] This work was carried out while the author was doing an internship at The Forensic Panel.

arguments highlighting particular evidence may still be specific to a given prosecutor; they are detailed, but not necessarily informative of any reliable definition of what a heinous murder is. In addition, an elective interpretation of "heinous," even if it provides an example that other prosecutors follow, still falls short of the *Gregg* aspirations of relying on developed societal standards.

Jurors and judges in American courts currently have no guidance to assess the level of heinousness or depravity of a crime. Inconsistency in defining the worst of murders continues to bedevil criminal casework. Moreover, the subjective interpretations of "depraved" and other HAC aggravators continue to fall short of accounting for societal standards.

In recent years, court decisions mandating large scale early release to relieve prison overcrowding has highlighted a related dilemma. How can early release decisions be rendered fairly among the sizeable numbers of prisoners of the same class of crime? In California, the state most notably associated with the magnitude of court-mandated release (see; *Brown v. Plata,* 2011), parole decision-making in murder cases has been criticized for the non-specific and widespread designation of first and second degree murders as "heinous, atrocious, and cruel" to avoid early release (Ball, 2009). When a court forces the issue, and demands that prisons be liquidated, what mechanism gives clarity to a parole board's newly mandated discipline for designating those prisoners never to be released?

How, for example, does one make a fair decision about which murder, attempted murder, or manslaughter defendant warrants early release? If both represent the same level of risk to the community, what makes one offender more suitable for release? An evidence-driven distinction of the worst of murders would better inform such decisions and promote public confidence in sensitive decisions.

The Depravity Standard is a 25-item evidence-based guide for the appraisal of criminal depravity in accordance with the goals set out by the Supreme Court in *Gregg, Godfrey*, and *Walton*. It has been developed to assist triers of fact to objectively assess a crime's relative severity in order to inform fair sentencing and release decisions. In an effort to minimize arbitrary sentencing decisions that result from inadequate scrutiny of relevant evidence, and from implicit and explicit biases, the Depravity Standard focuses investigation of depravity upon elements of the crime itself as opposed to the defendant's personal background.

In Welner, O'Malley, Gonidakis, Saxena, and Burnes (2018), the authors detailed the methodology of developing and refining the Depravity Standard in Studies 1–3 of this research. In Study 1, items of depravity were formed from review and thematic analysis of the specific rationale of 110 appellate court cases upheld as "heinous, atrocious, or cruel." Fifteen core elements of a perpetrators' intent, actions, and attitude were supplemented with data from 91 students and professionals in the industry, proposing aspects of crimes that rendered them "depraved." The efforts from these exercises resulted in 26 items for closer study.

In Study 2, the public opinion of 25,096 U.S. participants was captured to determine whether the items derived from Study 1 warranted inclusion in a final Depravity Standard. No items were flagged for exclusion from further study. The final development phase, Study 3, applied these items to actual closed felony murder cases. Through this review process, extensive qualifying and disqualifying definitions for each item were developed and refined.

In the course of the study, one item was dropped because of concerns that its overlap with other items would create vagueness in its definition, allowing for discrepant application. Twenty-five items remained for further study of the Depravity Standard (Table 1).

This article details the validation of the Depravity Standard (Studies 4 and 5) and its proposed application in U.S. court sentencing, in early release decision-making by parole officials, and for pardon decisions by elected officials.

2. Study 4 Phase A: inter-rater reliability

Each of the 25 items of the Depravity Standard include extensive definitions with various qualifying and disqualifying examples of the items' potential application in murder cases. This high level of detail minimizes arbitrariness and promotes consistent application in casework, however common or obscure the potential fact pattern. An important aspect of validating the tool, therefore, is ensuring inter-rater reliability. The data for this study is a subset of the data set reported in Study 3, described in depth in Welner et al. (2018).

2.1. Method

2.1.1. Sample

From the 770 cases reviewed as part of Study 3, a subsample of 250 cases were randomly selected for inclusion using a random sample generator (Haahr, 2006).

2.1.2. Procedure

Each of the 250 cases in Study 3 were rated twice by two independent groups of raters for the presence of the Depravity Standard items. As part of Study 3, each case was assigned an overall rating of present (Yes), absent (No) or insufficient data (ID) for each of the 25 items. The ID responses represented either a lack of information (i.e., in an autopsy report, the photo of the body was too overexposed to see any detail) or a rater's uncertainty about the information presented (i.e., the defendant's statement contradicts a co-defendant's statement and there is no way to determine which is the true account from the provided materials without speculation).

2.2. Data analysis

Data was entered into IBM SPSS version 22.0 for Windows (IBM Corp., 2013) for analysis. The data was screened for coding errors, and responses for the presence of each item were coded as $1 = $ Yes, $2 = $ No, and $3 = $ ID. Ratings of "No" and "ID" have the same functional significance in a criminal justice context – namely, that proof of guilt requires presence of evidence beyond a reasonable doubt. With the exception of prior convictions, a court may not use aggravating factors to impose a harsher sentence than usual unless the jury found those factors to be true beyond a reasonable doubt (i.e., *Cunningham v. California,* 2007). Therefore, absence of evidence equals evidence of absence. As such, the ratings of "No" and "ID" were combined for analysis and recoded to $1 = $ Yes and $2 = $ No/ID.

Upon preliminary analysis of Cohen's Kappa (Cohen, 1960) for inter-rater reliability, high agreement between raters yielded a low or negative Kappa value. Manual review of the 250 cases demonstrated that ratings of "No/ID," where the item is absent or cannot be determined with confidence, appeared significantly more frequently than ratings of "Yes," where the item is present without doubt.

The lower occurrence of "Yes" ratings can be attributed to the extensive development and validation phases of the research, where items were refined to describe specific criteria of depraved intent, actions, attitudes, and victim choice that reflect depravity, or the "worst-of-the-worst" murders. The thorough development process was described in Study 2 (Welner et al., 2018), and provided participant raters with specific criteria for each of the items to ensure that, when scored, the items reflected exceptional qualities in a crime. Raters therefore had the necessary guidance to highlight only a small subset of offenses for which any of the items were present. This meant that the worst, most depraved offenses only appeared as a small percentage of overall cases, and this discriminant sensitivity was the hypothesized result.

The AC_1 statistic (Gwet, 2008) was applied to each of the 25 items across the 250 random cases in IBM SPSS version 22.0 to determine the level of agreement between independent raters. The AC_1 statistic is noted by Gwet (2008) to measure 'true' inter-rater reliability in that it