Die Bedeutung von empirischen Befunden für die Strafzumessung in Deutschland

Methodische Bemerkungen anlässlich einer Studie zur Sanktionierung von Sexualdelikten (Ehlen/Hoven/Weigend, KriPoZ 2024, 16)

von Jessica Krüger, MPhil (Cantab)[*]

Abstract
In der juristischen und nicht-juristischen Öffentlichkeit ist in den letzten Wochen eine Debatte über die Strafzumessung bei Sexualdelikten angestoßen worden. Der Beitrag „Die strafrechtliche Sanktionierung von Sexualdelikten", veröffentlicht im ersten Heft der KriPoZ 2024, gibt an, empirische Befunde zu diesem Thema zu liefern. Aufgrund methodischer Schwachstellen sollten viele Ergebnisse des Beitrags jedoch nur zurückhaltend für Aussagen über die Strafzumessung in Deutschland oder für die Begründung kriminalpolitischer Forderungen herangezogen werden.

In recent weeks, a debate about sentencing in the case of sexual offenses has been initiated in the legal community and in public. The article "Die strafrechtliche Sanktionierung von Sexualdelikten", published in the first issue of KriPoZ 2024, claims to provide empirical findings on this topic. However, due to methodological issues, many of the article's findings should be used with caution when making statements about sentencing in Germany or criminal policy demands.

I. Einleitung

Eine Strafrechtswissenschaft, die ihre normativen Annahmen an der Wirklichkeit erprobt und aus empirischen Erkenntnissen neue Schlüsse zieht, hat viel zu gewinnen. *Ehlen*, *Hoven* und *Weigend* nutzen dieses Potenzial in ihrem jüngst erschienenen Beitrag (KriPoZ 2024, 16), indem sie, gestützt auf eine Auswertung der Strafverfolgungsstatistik, ein Umdenken in der Strafzumessung fordern – ein Anliegen, das auch durch Beiträge und Interviews in öffentlichen Medien[1] bereits vor Erscheinen der Studie für Diskussionen in der Strafrechtswissenschaft gesorgt hat.[2]

Zugleich zwingen die im Strafrecht regelmäßig erheblichen Rechtsfolgen im Besonderen dazu, die wirkungsmächtige Argumentation mit rechtstatsächlichen Erkenntnissen in methodischer Hinsicht hinreichend abzusichern. Insbesondere sollten die Grenzen der empirischen Methoden reflektiert und offengelegt werden. Hieran gemessen veranlassen Teile der publizierten Studie von *Ehlen*, *Hoven* und *Weigend* – ganz unabhängig von der Bewertung der Strafpraxis im Bereich der Sexualdelikte – zu einer kritischen Stellungnahme. Denn aufgrund methodischer Kritikpunkte, die im Folgenden dargestellt werden, sind viele Ergebnisse des Beitrags nur bedingt dazu geeignet, Aussagen über die Strafzumessung in Deutschland zu treffen oder kriminalpolitische Forderungen zu begründen und sollten entsprechend zurückhaltend verwendet werden.

II. Problemfeld 1: Erkenntnisinteresse und vermeintliche Erkenntnisse des Beitrags

Das Hauptanliegen von *Ehlen/Hoven/Weigend* ist es, anhand der Strafverfolgungsstatistik aufzuzeigen, dass sich die Strafzumessung für Sexualdelikte regelmäßig im unteren Drittel des gesetzlichen Strafrahmens bewegt[3] und darauf aufbauend zu begründen, weshalb diese niedrigen Strafhöhen problematisch seien.[4] Daneben werfen die Autoren in ihrer Einführung weitere Fragen auf, die mit ihren Untersuchungen beantwortet werden sollen, legen jedoch nicht dar, wie sie aus den präsentierten Ergebnissen der empirischen Untersuchungen die konkreten Schlussfolgerungen ableiten, die sie als Beantwortung dieser Forschungsfragen verstehen.

So schreiben *Ehlen/Hoven/Weigend* in der Einführung, dass sich die Frage stelle, „ob die Praxis der Strafzumessung bei Verletzungen von § 177 StGB diesem gewandelten Verständnis [Anm. d. Verf.: der Bevölkerung bzgl. der Gewichtung des Unrechts von Eingriffen in die sexuelle Selbstbestimmung] gerecht wird" und antworten sogleich: Die präsentierten Ergebnisse würden zeigen, „dass dies [Anm. d. Verf.: also eine Anpassung der Strafzumessungspraxis an die gewandelten Vorstellungen der Bevölkerung] weitgehend nicht der Fall ist, sondern dass die Strafzumessung traditionelle Maßstäbe aus vergangenen

[*] *Jessica Krüger* ist Wissenschaftliche Mitarbeiterin am Lehrstuhl für deutsches, europäisches und internationales Strafrecht und Strafprozessrecht, einschließlich Medizin-, Wirtschafts- und Steuerstrafrecht von *Prof. Dr. Karsten Gaede* an der Bucerius Law School in Hamburg. Für seine Unterstützung beim Verfassen des Manuskripts danke ich *Professor Dr. Kilian Wegner*.

[1] S. *Hoven/Rostalski*, Übergriffe härter bestrafen, FAZ.net v. 27.12.2023, online abrufbar unter: https://t1p.de/91tug (zuletzt abgerufen am 6.3.2024), sowie weiter *Garbe*, „Die Gesellschaft würde sexuelle Übergriffe anders bestrafen als die Justiz", Interview mit *Hoven*, Spiegel Online v. 31.1.2024, online abrufbar unter: https://t1p.de/w7wvb (zuletzt abgerufen am 6.3.2024).

[2] In chronologischer Reihenfolge: *Fischer*, Sollten Richter bei Sexualstraftaten härter urteilen?, LTO v. 6.1.2024, online abrufbar unter: https://t1p.de/jax1j (zuletzt abgerufen am 6.3.2024); *Hörnle*, Werden Vergewaltigungen und andere Sexualdelikte zu milde bestraft?, Spiegel Online, Gastbeitrag v. 20.1.2024, online abrufbar unter: https://t1p.de/2zgee (zuletzt abgerufen am 6.3.2024); *Fischer*, Strafzumessung per Umfrage? Spiegel Online, Kolumne v. 23.1.2023, online abrufbar unter: https://t1p.de/qtru9 (zuletzt abgerufen am 6.3.2024).

[3] *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (18).

[4] *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (23 ff.).

Zeiten vielfach ungeprüft anwendet".[5] Mit ähnlicher Stoßrichtung schreiben sie im letzten Absatz ihres Diskussion-Teils, die Untersuchung habe gezeigt, „dass die Verhängung milder Strafen im Bereich der Sexualdelikte Ausdruck einer gefestigten, seit vielen Jahren bestehenden Praxis" sei. Dabei bleibe unberücksichtigt, dass sich „gerade in diesem Bereich in den letzten Jahrzehnten ein erheblicher Wandel im gesellschaftlichen Bewusstsein vollzogen" habe.[6]

Mit Blick auf das Design der Studie verwundert die so umrissene Forschungsfrage ebenso wie die dazu präsentierte Antwort. Denn weder wird ein greifbares Maß für den Wertewandel in der Bevölkerung angeboten, noch wird im Verlauf des Beitrags ein konkreter (empirischer) Vergleich der Strafzumessungspraxis mit den Wertevorstellungen der Bevölkerung durchgeführt.

Auch die Frage, inwieweit sich die Strafzumessungspraxis im Bereich des Sexualstrafrechts über die Jahre verändert hat, wird in der Studie nur am Rande untersucht. Diese Frage werfen die Autoren zwar nicht explizit auf. Sie suggerieren jedoch an mehreren Stellen, sie untersucht zu haben: So stellen sie in der Einführung fest, ihre Untersuchung zeige, „dass die Strafzumessung *traditionelle Maßstäbe aus vergangenen Zeiten* vielfach ungeprüft *weiter anwendet*".[7] Zudem behaupten sie im Diskussions-Teil sowohl, dass die vorgenommenen empirischen Untersuchungen zu der Erkenntnis führe, dass die „*überkommene* Praxis in dem jeweiligen Gerichtsbezirk […] ein wesentlicher Faktor für die Bemessung der Strafe"[8] sei und dass die „Verhängung milder Strafen […] Ausdruck einer gefestigten, *seit vielen Jahren bestehenden* Praxis"[9] sei. Die empirischen Grundlagen für diese Aussagen werden aus der Studie jedoch nicht unmittelbar ersichtlich:

Im Studiendesign spiegelt sich das Interesse an der Entwicklung der Strafzumessung über die Zeit zwar insofern wider, als bei den Urteilsanalysen explizit auch Verurteilungen nach alter Rechtslage einbezogen werden, um diese mit Verurteilungen nach neuer Rechtslage vergleichen zu können.[10] Bei der Auswertung zu „regionalen Unterschieden" der Strafzumessung wird in Abb. 3 auch zwischen Verurteilungen nach alter und neuer Rechtslage unterschieden. Die Autoren widmen den in Abb. 3 dargestellten Unterschieden jedoch nur wenig Aufmerksamkeit und stellen fest: „Verurteilungen wegen Vergewaltigung mit Gewaltanwendung wiesen in Bayern und Sachsen im Durchschnitt wesentlich höhere Strafen auf als in Nordrhein-Westfalen und Hamburg. Dies gilt für Verurteilungen nach alter und neuer Rechtslage gleichermaßen".[11] Diese Interpretation übergeht, dass Abb. 3 für Verurteilungen nach neuer Rechtslage eine durchschnittliche Strafhöhe von 34 Monaten in Bayern und 39 Monaten in Hamburg anzeigt, für Hamburg also höhere Werte ausgewiesen werden als für Bayern. Die durchschnittlichen Höhen der in der Stichprobe verhängten Strafen nach neuem Recht liegen in diesen beiden Bundesländern mit fünf Monaten außerdem vergleichsweise nah beieinander (näher jedenfalls als die verhängten Strafen nach neuem Recht in Sachsen und Bayern, die elf Monate auseinander liegen). Nur bei Verurteilungen nach alter Rechtslage liegen Hamburg und Bayern 24,6 Monate auseinander.[12] Wollte man aus diesen Daten Erkenntnisse für das Bestehen (oder nicht-Bestehen) eines Wandels der Strafzumessungspraxis ableiten (gemessen an der Verhängung höherer Strafen für Vergewaltigung mit Gewaltanwendung)[13], hätte man zudem darauf eingehen können, dass bzw. warum die in Abb. 3 abgebildeten Daten in drei von vier Bundesländern eine niedrigere durchschnittliche Strafhöhe der ausgewerteten Fälle bei Verurteilungen nach neuer Rechtslage anzeigen als bei Verurteilungen nach alter Rechtslage, wobei Hamburg die Ausnahme darstellt. In Bayern liegen die im Sample gemessenen Strafhöhen nach neuer Rechtslage gar 18,1 Monate unter denen nach alter Rechtslage. Dies wird im Beitrag jedoch nicht angesprochen.

In den beiden anderen Teil-Untersuchungen wird von vornherein kein Vergleich zwischen Urteilen nach alter und neuer Rechtslage vorgenommen: Die Untersuchung zu den „Strafhöhen" beschränkt sich auf Urteile nach neuer Rechtslage. Bei der Untersuchung zu strafzumessungsrelevanten Faktoren wird nicht zwischen alter und neuer Rechtslage differenziert. Auch der Strafverfolgungsstatistik werden keine Erkenntnisse zu Bestehen oder Fehlen eines Wandels der Strafzumessungspraxis entnommen, da die Autoren die Strafverfolgungsstatistik nicht über die Zeit hinweg auswerten. Die Gruppengespräche schließlich bringen Erkenntnisse dazu, dass die Teilnehmer der Gespräche Schwierigkeiten mit der Interpretation der neuen Gesetzeslage hatten und dass regionale Gewohnheiten bei der Strafzumessung zum Wohnungseinbruchsdiebstahl existieren. Die präsentierten Erkenntnisse bieten aber weder einen Aufschluss darüber, ob sich die Praxis der Strafzumessung über die Zeit gewandelt hat, noch darüber, wie lange bestimmte regionale Gewohnheiten speziell bei Sexualdelikten schon praktiziert werden.

III. Problemfeld 2: Teilweise fehlende Begründung der Wahl der konkreten Forschungsmethode

Im empirischen Teil ihres Beitrags werten die Autoren zunächst die Strafverfolgungsstatistik aus, nehmen sodann eine quantitative Analyse einer Stichprobe von 86 Urteilen vor und präsentieren abschließend Ausschnitte aus

5 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (17).
6 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (25).
7 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (17), Hervorhebung durch d. Verf.
8 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (23), Hervorhebung durch d. Verf.
9 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (25), Hervorhebung durch d. Verf.
10 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (19 oben).
11 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (19).
12 Der Grund für diese Unterschiede liegt vermutlich nicht in einer drastisch geänderten Strafzumessungspraxis in Bayern oder Hamburg, sondern wohl daran, dass das für Abb. 3 verwendete Subsample möglicherweise klein und nicht repräsentativ ist, s. dazu unten IV. und VI.
13 Solche Schlussfolgerungen sollten allerdings angesichts der geringen Stichprobengröße, der nicht gewährleisteten Repräsentativität des Samples und ohne Überprüfung der Signifikanz nur mit Vorsicht gezogen werden, s. unten VI.

drei Gruppengesprächen mit Richtern und Staatsanwälten. Im Rahmen ihrer quantitativen Urteilsauswertung erschließt sich jedoch nicht stets, welches konkrete Erkenntnisinteresse mit der jeweiligen Teilstudie verfolgt wird und weshalb die Methode der jeweiligen Teilstudie besonders geeignet sein soll, einen Beitrag zur Befriedigung dieses konkreten Erkenntnisinteresses zu leisten:

Unklar bleibt, welchen Mehrwert die Autoren aus ihrer Auswertung der verhängten Strafhöhen in der kleinen, nicht repräsentativen Stichprobe an Urteilen gegenüber der Auswertung der Grundgesamtheit, also der verhängten Strafhöhen aller Urteile in Deutschland, ziehen. Üblicherweise wird mit Stichproben gearbeitet, wenn die Auswertung der Grundgesamtheit zu schwierig, zu aufwendig oder zu teuer ist.[14] Gegenüber einer Auswertung der in der Strafverfolgungsstatistik enthaltenen Daten der Grundgesamtheit bietet die Auswertung des Samples in Abb. 2 lediglich insofern ein Mehr an Detailtiefe, als die Autoren die in der Strafverfolgungsstatistik angegebenen Zeiträume „1-2 Jahre“ und „2-3 Jahre“ in ihrer Abbildung auf jeweils zwei Subkategorien aufteilen sowie Mittelwert und Median der verhängten Freiheitsstrafen ausweisen.[15] Allerdings machen die Autoren diese Werte für ihre weitere Diskussion nicht fruchtbar.

Bei der Auswertung der Urteile nach strafzumessungsrelevanten Merkmalen schließlich stellt sich die Frage, weshalb hier ein quantitativer und kein qualitativer Ansatz gewählt worden ist. Mit qualitativen Analyse-Methoden wie einer qualitativen Inhaltsanalyse[16] hätten z.B. Begründungsmuster, Argumentationsstruktur, die Beziehung der einzelnen Strafzumessungsfaktoren untereinander und deren Relevanz greifbar und in Detailtiefe herausgearbeitet werden können. Der von den Autoren gewählte quantitative Ansatz ist nicht in der Lage, vergleichbar detaillierte Erkenntnisse zu liefern.[17] Der Vorteil eines quantitativen Ansatzes gegenüber einem qualitativen Ansatz hingegen ist, dass die Ergebnisse unter bestimmten Bedingungen generalisierbar sind, also auf die Grundgesamtheit übertragen werden können. Aus den präsentierten Informationen geht aber nicht hervor, dass die Voraussetzungen vorliegen, um die Ergebnisse der quantitativen Analyse zu generalisieren (näher dazu unten VI.). Die potenziellen Vorteile des quantitativen Ansatzes gegenüber qualitativen Methoden werden so nicht fruchtbar gemacht.

IV. Problemfeld 3: Die Darstellung der verwendeten Daten

Ehlen/Hoven/Weigend fassen die Informationen über die in ihrem Datensatz enthaltenen Urteile in einer guten Spalte Text zusammen.[18] Über den Datensatz erfährt man, dass 86 Urteile mit 97 Taten ausgewertet wurden, davon 49 nach alter und 48 nach neuer Rechtslage abgeurteilt. Eine detaillierte Aufschlüsselung, wie genau sich die Urteile auf die einzelnen Jahre, Bundesländer und Delikte verteilen, fehlt dagegen. So kann der Leser nicht nachvollziehen, auf welcher Datenbasis die im Beitrag erfolgte Analyse der Strafzumessung nach Bundesländern[19] beruht und kann die Aussagekraft der gewonnenen Ergebnisse nicht vollständig beurteilen:

Die Auswertung der regionalen Unterschiede in der Strafhöhe bezieht sich nur auf Verurteilungen wegen „Vergewaltigung mit Gewaltanwendung“. Es ist daher nur ein Teil der 97 insgesamt ausgewerteten Taten, die auch andere Deliktsvariationen umfassen, in dieser Unterstichprobe enthalten. Wie viele dies genau sind, weist der Beitrag jedoch nicht aus. Geht man davon aus, dass die Stichproben „Aburteilung nach alter Rechtslage“ (49 Taten) und „Aburteilung nach neuer Rechtslage“ (48 Taten) zu je einem Viertel Fälle zu „Vergewaltigung mit Gewaltanwendung“ enthalten[20], ergibt das 12,5 bzw. 12 Fälle von Vergewaltigung mit Gewaltanwendung pro Unterstichprobe. Bei angenommener gleichmäßiger Verteilung auf die Bundesländer bleiben ca. 3 Fälle pro Bundesland nach alter Rechtslage und 3 Fälle nach neuer Rechtslage, die in den Vergleich eingestellt werden konnten.

V. Problemfeld 4: Wahl, Anwendung und Erläuterung der statistischen Methoden sowie die Interpretation der ermittelten Werte bei der Auswertung der Strafzumessungsumstände

Auch die Auswertung der Urteile nach Strafzumessungsumständen wirft Fragen auf, und zwar bezüglich der Wahl, der Anwendung und der Erläuterung der statistischen Methoden sowie der Interpretation der präsentierten Werte. So scheinen in einem Fall, in dem die Autoren den Spearman-Korrelationskoeffizienten berechnet haben, die Voraussetzungen für die Berechnung dieses Koeffizienten nicht vorzuliegen (1.). Die angegebene Definition für „Eta-Quadrat“, die Hinweise zur Interpretation der präsentierten Werte und die Interpretation der Werte selbst sind falsch bzw. ungenau (2.). Zuletzt erschwert die knappe Auseinandersetzung mit den präsentierten Ergebnissen dem mit Statistik nicht vertrauten Leser, die angegebenen Werte einzuordnen (3.).

1. Wahl und Anwendung der statistischen Methoden (Spearman)

Die Entscheidung, welchen statistischen Test man verwendet, hängt u.a. entscheidend von der Art der Variablen ab.[21] Man kann grundsätzlich zwischen nominalskalierten, ordinalskalierten und intervallskalierten Variablen

---

14 *Weisburd/Britt*, Statistics in Criminal Justice, 4. Aufl. (2014), S. 127.

15 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (19).

16 S. dazu sowie zu weiteren qualitativen Auswertungsmethoden *Bortz/Döring*, Forschungsmethoden und Evaluation, 4. Aufl. (2006), S. 331 ff.

17 Zur beschränkten Aussagekraft der berechneten Zusammenhangswerte s. auch V.3.

18 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (18 f.).

19 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (19).

20 Dies ist mangels konkreter Angaben im Sachverhalt ein reiner Schätzwert. Da vier unterschiedliche Arten von Delikten (sexuelle Übergriffe mit und ohne Gewaltanwendung und Vergewaltigung mit und ohne Gewaltanwendung) aufgenommen wurden, habe ich die Zahl der Urteile durch vier geteilt.

21 *Gau*, Statistics for Criminology and Criminal Justice, 3. Aufl. (2019), S. 204.

unterscheiden: Nominalskalierte Variablen können entweder dichotom sein (dies ist der Fall, wenn die Variable nur zwei Ausprägungen hat, z.B. Vorstrafen vorhanden [1] oder nicht vorhanden [2]) oder mehrere Ausprägungen umfassen (z.B. in der Variable „Bundesland" die Bundesländer Bayern [1], Hamburg [2], NRW [3] oder Sachsen [4]). Die einzelnen Kategorien einer nominalskalierten Variable stehen untereinander in keinem Rangverhältnis. Variablen, deren Werte Zahlen sind, die zueinander in einem Rangverhältnis stehen und deren Abstand zueinander stets gleich groß ist (z.B. die Höhe der Freiheitsstrafe gemessen in Jahren: ein Jahr, zwei Jahre, drei Jahre usw.), werden als intervallskalierte oder metrische Variablen bezeichnet. Eine dritte Art von Variablen sind ordinalskalierte Variablen. Hier werden die einzelnen Werte in eine Rangfolge gebracht, die aber nicht zwingend dem tatsächlichen Abstand der gemessenen Werte entspricht. Beispiel: Die Haftstrafe in Urteil A beträgt sechs Monate, in Urteil B zehn Monate und in Urteil C 24 Monate. Die Länge der Haftstrafe ist damit zunächst eine intervallskalierte Variable. Möchte man diese intervallskalierte Variable auf einer Ordinalskala abbilden, würde man der Haftstrafe in Urteil A den Wert 1, in Urteil B den Wert 2 und in Urteil C den Wert 3 geben. So bleibt zwar die Rangfolge der Werte gleich (1<2<3), über den Abstand zwischen den Werten kann aber keine Aussage mehr getroffen werden.[22]

Für die Berechnung der Relevanz der einzelnen Strafzumessungsfaktoren haben sich die Autoren für den Spearman-Koeffizienten („Vorstrafen Ja/Nein" und „Anzahl der Vorstrafen"), für den Pearson-Koeffizienten („Höhe der Vorstrafen") sowie für „Eta-Quadrat" (die restlichen Variablen) entschieden.[23] Die Pearson-Korrelation beschreibt die Stärke der linearen Beziehung zwischen zwei intervallskalierten Variablen. Auch die Spearman-Korrelation beschreibt die Stärke einer Beziehung zwischen zwei Variablen, die sich (ähnlich wie bei Pearson) gemeinsam in eine Richtung verändern. Spearman setzt aber – anders als Pearson – keine Intervallskalierung voraus, sondern kann auch für Variablen, die mindestens ordinalskaliert sind, aussagekräftige Ergebnisse erzielen.[24] Vor diesem Hintergrund verwundert es, dass für die Berechnung des Zusammenhangs zwischen dem Vorhandensein von Vorstrafen und der Strafhöhe den Spearman-Koeffizienten gewählt wurde. Die Variable „Vorstrafe Ja/Nein" ist eine dichotome nominalskalierte Variable, da sie nur zwei Antwortmöglichkeiten offenlässt: Ja oder Nein. Sie ist keine ordinalskalierte Variable. Dies wäre allerdings eine Voraussetzung, um einen aussagekräftigen Spearman-Koeffizienten zu berechnen.[25]

2. Ungenauigkeiten bei der Definition von „Eta-Quadrat", den Interpretationshinweisen und der Interpretation einzelner Werte

In Fn. 22 übernehmen die Autoren die in *Janssen/Laatz*, Statistische Datenanalyse mit SPSS, 9. Aufl. (2016), S. 280 angegebene Definition von „Eta", um „Eta-Quadrat" zu definieren. „Eta" ist ein Koeffizient, der das Maß des Zusammenhangs zwischen einer nominalen Variable und einer intervallskalierten Variable angibt.[26] Eta-Quadrat wiederum wird aus dem Quadrat von Eta berechnet und gibt den Anteil der Varianz der abhängigen Variable (Strafhöhe) an, der durch die unabhängige Variable (Strafzumessungsfaktor) erklärt wird,[27] wobei „Varianz" beschreibt, wie stark die einzelnen Werte einer Gruppe (hier: die Kategorien eines Strafzumessungsfaktors) vom Mittelwert der Gruppe abweichen.[28] Mit anderen Worten: Eta-Quadrat zeigt an, welcher Anteil der Unterschiede der Strafhöhen in den Gruppen der nominalen Variable (z.B. Geständnis grundsätzlich [ja] und [nein]) sich durch den untersuchten Strafzumessungsfaktor erklären lässt.[29] Je größer Eta-Quadrat ist, desto mehr (oder besser) werden die Unterschiede in den Strafhöhen der beiden Gruppen durch den untersuchten Strafzumessungsfaktor erklärt.[30] Eta und Eta-Quadrat sind unterschiedliche Werte: Hat Eta beispielsweise einen Wert von 0.2 so beträgt Eta-Quadrat $0.2^2=0.04$ (dies ist der für „Geständnis grundsätzlich" angegebene Wert in Abb. 5). Geht man also davon aus, dass die in der rechten Spalte von Abb. 5 angegebenen Werte Eta-Quadrat-Werte und keine Eta-Werte sind, kann man sagen, dass 4 % der beobachteten Varianz in den Strafhöhen der beiden Urteilsgruppen „Geständnis grundsätzlich: Ja" und „Geständnis grundsätzlich: Nein" durch das Ablegen eines Geständnisses erklärt werden kann. Anders als Pearson oder Spearman, die Werte zwischen -1 und +1 annehmen können und damit auch Aussagen über die Richtung des bestehenden Zusammenhangs treffen,[31] kann Eta-Quadrat nur Werte zwischen 0 und 1 annehmen und damit keine Aussage über die Richtung des Zusammenhangs treffen.[32]

Um dem Leser zu ermöglichen, diese Werte selbst einzuordnen, geben *Ehlen/Hoven/Weigend* dem Leser in Fn. 24 Hinweise zur Interpretation der Werte. Dabei präzisieren

22 S. zu den unterschiedlichen Arten von Variablen instruktiv *Gau*, Statistics for Criminology and Criminal Justice, S. 21 ff.

23 Alles *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (20).

24 S. für einen Überblick über verschiedene Zusammenhangsmaße (darunter *Pearson* und *Spearman*) und deren Voraussetzungen hinsichtlich der Art der Variablen *Janssen/Laatz*, Statistische Datenanalyse mit SPSS, 9. Aufl. (2016), S. 268.

25 *Janssen/Laatz*, Statistische Datenanalyse mit SPSS, S. 275 f.

26 *Janssen/Laatz*, Statistische Datenanalyse mit SPSS, S. 280: „Eta ist ein spezieller Koeffizient für den Fall, dass die unabhängige Variable auf Nominalskalenniveau gemessen wurde, die abhängige aber mindestens auf Intervallskalenniveau. Er zeigt an, wie sehr sich die Mittelwerte für die abhängige Variable zwischen den verschiedenen Kategorien der unabhängigen unterscheiden. Unterscheiden sie sich gar nicht, wird eta 0. Unterscheiden sie sich dagegen stark und ist zudem die Varianz innerhalb der Kategorien der unabhängigen Variablen gering, tendiert er gegen 1."

27 *Bortz/Döring*, Forschungsmethoden und Evaluation, 4. Aufl. (2006), S. 615.

28 *Weisburd/Britt*, Statistics in Criminal Justice, S. 104 ff. Liegen beispielsweise in einer fiktiven Gruppe A alle in den Urteilen gemessenen Strafhöhen eng beieinander (alle verhängten Strafen liegen im Bereich zwischen 20 und 24 Monaten) ist die Varianz geringer als in einer fiktiven Gruppe B, in der die gemessenen Strafhöhen deutlich unterschiedlicher ausfallen (die verhängten Strafen liegen zwischen 6 und 40 Monaten).

29 *Janssen/Laatz*, Statistische Datenanalyse mit SPSS, S. 280.

30 *Weisburd/Britt*, Statistics in Criminal Justice, S. 329.

31 So auch erläutert von *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 in Fn. 22.

32 *Weisburd/Britt*, Statistics in Criminal Justice, S. 329.

sie jedoch nicht, dass die von ihnen angegebene Klassifikation der Effektgrößen bei *Bortz/Döring*, auf die sie in Fn. 24 explizit Bezug nehmen, nur für Eta-Quadrat, nicht aber für Pearson oder Spearman gilt. Für Pearson und Spearman geben *Bortz/Döring* folgende Klassifikation der Effektgrößen an: Klein: 0.10 – Mittel: 0.30 – Groß: 0.50.[33] Während eine mittlere Effektgröße bei Eta-Quadrat also bereits bei einem Wert von 0.1 gegeben ist, liegt ein mittelgroßer Zusammenhang bei Pearson oder Spearman erst bei 0.3 vor. Die so formulierten Interpretationshinweise bergen die Gefahr, dass mit Statistik nicht vertraute Leser die ermittelten Zusammenhänge nach Pearson oder Spearman einerseits überschätzen und andererseits verwundert sein dürften ob der (nach *Bortz/Döring* zutreffenden) Beschreibung der Autoren auf S. 20, der ermittelte Wert von 0.316 (Pearson) für „Höhe der Vorstrafen" zeige einen „mittelgroßen Zusammenhang". Wenig konsequent erscheint zudem, dass der ermittelte Wert von nur 0.231 (Spearman) für „Anzahl der Vorstrafen" entgegen der bei *Bortz/Döring* angegebenen Klassifikation ebenfalls als mittelgroßer Zusammenhang eingeordnet wird.[34]

3. Die gezogenen Schlussfolgerungen

Ehlen/Hoven/Weigend schließen aus den von ihnen präsentierten Zusammenhangsgrößen, ihre Ergebnisse zeigten, „dass die ausdrückliche Nennung eines Strafzumessungsumstandes keineswegs zwingend bedeutet, dass er auch tatsächlich Einfluss auf die Strafhöhe hat."[35] Ähnlich formulieren sie in einer ihrer zusammenfassenden sechs Erkenntnisse auf S. 22 f.: „Als strafzumessungsrelevante Faktoren werden insbesondere die psychischen Folgen beim Opfer und das Geständnis des Täters genannt. Ihr tatsächlicher Einfluss auf die Strafhöhe lässt sich allerdings nicht belegen".[36] Allerdings haben die Autoren in Abb. 5 für viele Strafzumessungsumstände Werte ausgewiesen, die nach der Klassifikation, auf die sie sich beziehen, immerhin als kleine oder mittlere Effekte einzuordnen wären. Dies gilt auch für die explizit erwähnten Gründe „Geständnis grundsätzlich" (0.04) und „Psychische Folgen der Tat für Opfer" (0.015). Angesichts dessen stellt sich die Frage, ob diese knappen Aussagen die gefundenen Ergebnisse hinreichend differenziert zusammenfassen und die angegebenen Interpretationshinweise konsequent umsetzen.[37]

Unabhängig davon sollten die berechneten Effektstärken nur zurückhaltend für Schlussfolgerungen über die tatsächliche Auswirkung von einzelnen Strafzumessungsgründen auf die Strafhöhe herangezogen werden. Denn die angegebenen Werte können mögliche Ausgleichs- und Verstärkungseffekte bei einem Zusammentreffen mehrerer Strafzumessungsgründe in einem Urteil nicht offenlegen. Nimmt man beispielsweise an, dass bei Ablegen eines Geständnisses die Strafhöhe tendenziell geringer ausfällt, das Vorliegen von psychischen Schäden beim Opfer hingegen zu einer höheren Strafe führt, liegt es nahe, dass sich straferhöhender und strafmildernder Effekt in einem Urteil, in dem beide Strafzumessungsgründe genannt werden, zu einem gewissen Grad ausgleichen. Je nachdem, wie sich derartige Ausgleichseffekte in der untersuchten Stichprobe verteilen, kann dies Einfluss auf die berechneten Effektgrößen für einzelne Strafzumessungsfaktoren haben. Hat beispielsweise in den neun ausgewerteten Fällen, in denen das Gericht als (strafschärfenden) Strafzumessungsumstand die demütigende, erniedrigende oder ekelhafte Behandlung des Opfers festgehalten hat, der Täter deutlich häufiger ein Geständnis abgelegt als in den Fällen, in denen eine solche Behandlung nicht in den Urteilsgründen dokumentiert wurde, könnte die im Sample gemessene Effektstärke für den Strafzumessungsfaktor der demütigenden Behandlung niedriger ausfallen, weil der strafschärfende Effekt überproportional oft durch den gegenläufigen strafmildernden Effekt eines Geständnisses ausgeglichen wurde.[38]

VI. Problemfeld 5: Zur Aussagekraft und Interpretation der gewonnenen Ergebnisse

Ehlen/Hoven/Weigend behaupten in ihrem Beitrag nicht explizit, dass ihre Ergebnisse der quantitativen Urteilsauswertung auf die Strafzumessung in Deutschland übertragen werden können. Die Autoren stellen in ihrem Beitrag jedoch auch nicht klar, dass die von ihnen präsentierten Daten nicht ausreichen, um die Ergebnisse der (quantitativen) Urteilsanalyse verallgemeinern zu können. Das ist insofern unglücklich, als dass sich an zwei Stellen der Stu-

33 *Bortz/Döring*, Forschungsmethoden und Evaluation, S. 606, Tab. 9.1, „2. Korrelationstest, ρ ".

34 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (20).

35 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (20).

36 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (23).

37 Ähnlich missverständlich erscheinen vor diesem Hintergrund einige Ausführungen *Hovens* im Interview mit Spiegel-Online (Fn. 1). Dort erklärt sie: „(…) in einer statistischen Analyse konnten wir zumindest bei den von uns untersuchten Urteilen nicht feststellen, dass sich die Folgen für die Opfer tatsächlich auf das Strafmaß ausgewirkt haben. Schweres Leid durch sexuelle Gewalt führt nicht zwingend zu härteren Strafen." Auf die Nachfrage „Was wirkt sich denn auf die Strafe aus?" erläutert sie: „(…) Es wirkt strafmildernd, wenn der Täter nicht vorbestraft war oder wenn er bei der Tat alkoholbedingt enthemmt war". Für die Variable „alkoholbedingt enthemmt" wird bei *Ehlen/Hoven/Weigend* ein Eta-Quadrat-Wert ausgewiesen, der kleiner ist als der für „psychische Folgen beim Opfer", nämlich 0.006. Für die Aussage, es wirke strafmildernd, wenn der Täter nicht vorbestraft ist, bietet die hier thematisierte Studie keine Datengrundlage, da die Voraussetzungen für den durchgeführten Spearman-Test nicht vorlagen, der in der Studie angeführte Wert also nicht aussagekräftig ist (s. oben V.1.).

38 Dieses Beispiel soll lediglich illustrieren, wie sich das Zusammenwirken unterschiedlicher Strafzumessungsgründe auf die gemessenen Effekte auswirken kann – in welchem Maße das Zusammentreffen von Strafzumessungsgründen zu solchen oder ähnlichen Effekten führt, kann mangels näherer Kenntnis des Datensatzes nicht beurteilt werden.

die Formulierungen finden, die beim Leser Missverständnisse hinsichtlich der Übertragbarkeit dieser Ergebnisse auslösen könnten.

Eine wichtige Information, um die Übertragbarkeit von Daten aus Stichproben auf die Grundgesamtheit beurteilen zu können, ist die statistische Signifikanz. Statistische Signifikanzwerte zeigen – vereinfacht gesagt – an, wie wahrscheinlich es ist, dass ein im Datensatz gefundener Effekt (z.B. eine Effektgröße) nicht nur das Ergebnis von Berechnungen mit dem Datensatz der Stichprobe ist, sondern der Effekt (also die gleiche Effektgröße) auch bei Berechnungen mit der Grundgesamtheit errechnet werden würde.[39] Anders gewendet: Ohne Signifikanzwert bedeuten die errechneten Werte zur Effektstärke nur, dass innerhalb der untersuchten Stichprobe von 86 Urteilen der Zusammenhang zwischen einem Strafzumessungsfaktor und der Strafhöhe eine bestimmte Größe hat. Der Leser kann aber nicht beurteilen, ob bzw. mit welcher Wahrscheinlichkeit dieser Zusammenhang auch in der Grundgesamtheit nachweisbar wäre, also in einem Datensatz, der alle Urteile zu sexuellen Übergriffen und Vergewaltigungen in Deutschland enthielte. Gleiches gilt für die ermittelten Unterschiede in der Strafhöhe zwischen den Bundesländern: Ohne eine Berechnung der Signifikanz sind keine Rückschlüsse darauf möglich, ob diese Unterschiede so auch außerhalb der Stichprobe existieren.[40]

Zusätzlich zur Signifikanz ist auch die Art der Stichprobengewinnung wichtig, um die Generalisierbarkeit von Ergebnissen beurteilen zu können. So wird in einschlägigen Statistik-Lehrbüchern stets betont, dass eine Generalisierung von statistischen Werten nur bei Zufallsstichproben zuverlässig möglich ist.[41] *Ehlen/Hoven/Weigend* haben ihre Stichprobe ermittelt, indem sie zunächst eine „qualitative Urteilsanalyse" vorgenommen und sodann auf Grundlage des Prinzips der maximalen strukturellen Variation „mehrere Fallauswertungen" durchgeführt haben, wobei sie auf besonders große regionale Varianz der Gerichte geachtet haben.[42] Bei der beschriebenen Methode handelt es sich um eine nicht-zufällige Stichprobenziehung. Mit dieser Methode soll abgesichert werden, dass bei einer *qualitativen* Auswertung Strukturen und Zusammenhänge möglichst umfassend erfasst werden können um neue Erkenntnisse zu gewinnen.[43]

Die Autoren schreiben zwar knapp, Ziel des gewählten Sampling-Verfahrens sei es gewesen, „den notwendigen Grad der Generalisierung rekonstruktiver Analyseergebnisse zu erreichen", zudem sollte „die relative Verallgemeinerbarkeit der Analyse gewährleistet werden."[44] Sie ordnen jedoch nicht näher ein, welche Ergebnisse der nachfolgenden *quantitativen* Erhebung sie zu den rekonstruktiven Analyseergebnissen zählen, die generalisierbar seien, welche Anforderungen sie im Kontext einer quantitativen Analyse an den („notwendigen") Grad der Generalisierung stellen und worauf sich eine nur relative Verallgemeinerbarkeit im Kontext der erfolgten quantitativen Auswertung bezieht. Auch wird nicht deutlich, woher die Urteile konkret stammen, also ob sie beispielsweise durch eine Suche in öffentlich verfügbaren Datenbanken oder durch Anfragen an Gerichte oder Verteidiger gewonnen wurden und welche möglichen Verzerrungen mit der jeweiligen Art der Stichprobengewinnung einhergehen könnten. So bleibt unklar, was die Art der Stichprobenziehung für die Relevanz der gefundenen Ergebnisse konkret bedeutet. Denn einerseits scheint es nicht das Ziel der Studie zu sein, die Übertragbarkeit der Ergebnisse der quantitativen Auswertung der Urteile auf die Strafzumessung in Deutschland abzusichern, da dafür jedenfalls Signifikanzwerte nötig gewesen wären. Andererseits scheinen die Autoren mit ihrer Art der Stichprobenziehung ein gewisses, wenn auch nicht konkretisiertes „Mehr" an Übertragbarkeit gegenüber anderen Arten der Stichprobenziehung gewährleisten zu wollen.

Da sie in der Diskussion aber nicht mehr einordnen, welche Bedeutung sie ihren Ergebnissen über die von ihnen untersuchte Stichprobe hinaus zumessen,[45] bergen die zitierten Formulierungen die Gefahr, einen mit Statistik wenig vertrauten Leser zu dem Schluss zu verleiten, die im Anschluss präsentierten Ergebnisse seien generalisierbar bzw. verallgemeinerbar. Ein ähnliches Problem stellt sich, wenn die Autoren im Methoden-Teil zu den Gruppengesprächen explizit herausstreichen, dass *diese* (qualitative) Art der Forschung „nach ihrem Design nicht in der Lage" sei, „quantitativ-repräsentative Ergebnisse zu liefern". Zum einen wird so die Schwäche der gewählten qualitativen Methode explizit in Kontrast mit den Vorzügen der „Repräsentativität" quantitativer Ansätze gesetzt, zum anderen wird durch den Bindestrich eine enge Verbindung von quantitativen Methoden mit „repräsentativen" Ergebnissen hergestellt, obwohl eine solche Verbindung keinesfalls zwingend ist. Dies geschieht unmittelbar nach der Präsentation von quantitativen Ergebnissen, die nicht gesichert übertragbar sind, was mE folgende Gefahr birgt:

---

39 Ausführlich *Weisburd/Britt*, Statistics in Criminal Justice, S. 135 ff.

40 Die Schwelle, ab der herkömmlich von einem „signifikanten Ergebnis" ausgegangen wird, liegt bei p=<0.05, wobei bei entsprechender Begründung davon abgewichen werden kann, *Weisburd/Britt*, Statistics in Criminal Justice, S. 136 ff.; *Fields*, Discovering Statistics Using IBM SPSS Statistics, 5. Aufl. (2018), S. 102, 111 f. Die Autoren setzen sich jedoch nicht mit der Bedeutung von Signifikanz oder bestimmten Schwellenwerten für Signifikanz auseinander, indem sie bspw. begründen, weshalb im konkreten Fall die Effektstärke (möglicherweise) trotz eines Wertes von p>0.05 eine interessante Erkenntnis liefert. In jedem Fall hätten aber auch in diesem Fall Signifikanzwerte ausgewiesen werden müssen.

41 *Paternoster/Bachman*, Essentials of Statistics for Criminology and Criminal Justice, 2018, S. 9; *Czaja/Blair*, Designing surveys. A guide to decisions and procedures, 2011, S. 124 ff.; Ausführlich *Bortz/Döring*, Forschungsmethoden und Evaluation, S. 335 f. (zur Problematik von Generalisierungen auf Basis nicht-zufallsgeleiteter, qualitativer Stichprobengewinnung generell) und S. 398 (zur Zufallsstichprobe).

42 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (18).

43 *Kleining*, Kölner Zeitschrift für Soziologie und Sozialpsychologie 1982, 224 (234 ff.).

44 *Ehlen/Hoven/Weigend*, KriPoZ 2024, 16 (18).

45 S. *Bortz/Döring*, Forschungsmethoden und Evaluation, S. 480 für ausführlichere Hinweise zu den Punkten, die bei der Arbeit mit einer nicht-probabilistischen Stichproben explizit angesprochen werden sollten – darunter die Frage, inwieweit die „Generalisierung der Ergebnisse durch Besonderheiten des Auswahlverfahrens eingeschränkt ist" – und den Vorteilen einer solchen Diskussion.

Der statistisch oftmals nicht geschulte Strafrechtler verbindet gedanklich die zuvor präsentierten quantitativen Ergebnisse fälschlicherweise mit dem Prädikat der „Repräsentativität“ und überschätzt sodann die Aussagekraft der quantitativ gewonnenen Ergebnisse oder hält sie möglicherweise gar für generalisierbar. Die im Beitrag präsentierten Informationen bieten aber keine belastbare Grundlage für eine Generalisierung.

VII. Zusammenfassung

Damit lassen sich die wesentlichen Kritikpunkte an der hier untersuchten Studie wie folgt zusammenfassen:[46]

1. Die in der Einführung formulierte, studienübergreifende Forschungsfrage, ob die Rechtspraxis der Strafzumessung bei Verletzungen von § 177 StGB den gewandelten Vorstellungen der Bevölkerung von der Schwere eines Eingriffs in die sexuelle Selbstbestimmung gerecht wird, wird in der Studie nicht beantwortet. Gleichzeitig soll die Studie zeigen, dass die derzeitige Strafzumessungspraxis „traditionelle Maßstäbe aus vergangenen Zeiten vielfach ungeprüft anwende“. Inwiefern sich aus den präsentierten Ergebnissen der empirischen Untersuchungen eine solche Schlussfolgerung ableiten lässt, erschließt sich nicht. Gleiches gilt für die wenig differenzierte Interpretation der in Abb. 3 abgebildeten regionalen Verteilung der Strafhöhen.

2. Bei der quantitativen Auswertung der in der Studie untersuchten Urteile nach „Strafhöhen“ ist unklar, welchen Erkenntnisgewinn die Auswertung des Samples gegenüber der Auswertung der Strafverfolgungsstatistik bringen soll. Bei der quantitativen Auswertung der Urteile nach Strafzumessungsgründen bleibt die Frage offen, weshalb die Autoren einen quantitativen anstatt eines qualitativen Ansatzes gewählt haben, da die Vorteile des quantitativen Ansatzes (Generalisierbarkeit der Ergebnisse) ungenutzt bleiben.

3. Die knappe Darstellung der verwendeten Daten erschwert es, die Aussagekraft der Ergebnisse zur regionalen Verteilung von Strafhöhen zu beurteilen, da dort nur ein Teil des beschriebenen Datensatzes an Urteilen analysiert wird, aber Angaben dazu fehlen, wie viele Fälle dieser Sub-Datensatz enthält und wie sich diese regional verteilen.

4. Bei der statistischen Auswertung der Strafzumessungsfaktoren liegt in einem Fall die Vermutung nahe, dass die Voraussetzungen für den durchgeführten statistischen Test nicht vorlagen, der errechnete Wert also nicht sinnvoll interpretierbar ist. Den Autoren unterläuft ein Fehler bei der Definition für „Eta-Quadrat“. Die Interpretationshinweise in Fn. 24 sind missverständlich formuliert, die Interpretation der Werte selbst nicht konsequent. Mit Statistik wenig vertraute Leser hätten zudem von einer differenzierteren Einordnung der präsentierten Daten profitiert.

5. Die Limitationen des für die Urteilsauswertung gewählten Forschungsdesigns werden nicht in der gebotenen Klarheit analysiert und herausgestellt. Dies birgt die Gefahr, bei der Leserschaft Missverständnisse bezüglich der Aussagekraft und Übertragbarkeit der gewonnenen Ergebnisse hervorzurufen.

[46] S. zu weiteren Kritikpunkten auch *Kölbel/Linder*, StV 2024, 322 (i.E.).