Juristische Entscheidungsfindung von Laien, Noviz:innen und Expert:innen im strafrechtlichen Ermittlungsverfahren: Psychologische Einflussfaktoren und individuelle Unterschiede

Inauguraldissertation
zur
Erlangung des Doktorgrades
der Humanwissenschaftlichen Fakultät
der Universität zu Köln
nach der Promotionsordnung vom 18.12.2018
vorgelegt von

Merle Ruppenthal
aus
Koblenz
im Studiengang Psychologie

Erstbetreuerin: Prof. Dr. Ellen Aschermann
Zweitbetreuer: PD Dr. Heinz Zimmer

Diese Dissertation wurde von der Humanwissenschaftlichen Fakultät der Universität zu Köln im Juli 2023 angenommen.

Abstract

Prosecutorial decisions made during the preliminary proceedings affect the following trial substantially. However, in many legal situations, there is not only one right decision, especially if the evidence is ambiguous. Models in decision-making, such as dual-process theories, indicate factors which influence the decision-making process. The aims of this study were to investigate how participants with different levels of expertise handle ambiguous evidence in a criminal case and to identify process-dependent as well as person-dependent factors contributing to this decision. A mixed sample of laypeople, legal novices (students and legal trainees), and experts took part in this questionnaire-based quasi-experiment. After being randomly presented with a vignette describing a case of either assault or theft, participants were required to make a decision from a prosecutorial perspective about any existing ground for suspicion and to provide information about the decision-making process as well as the handling of evidence. A randomized subsample temporarily experienced experimentally induced time pressure which did not significantly affect the decision-making process in most analyses. The type of crime, however, partly played a significant role. Cognitive reflection and the need for cognition as person-dependent factors showed small and partly non-significant effects. The level of expertise was a significant factor in most analyses to the extent that laypeople and experts even gave contrary responses. The results confirm the applicability of the newly-constructed vignettes, demonstrate the willingness of legal samples to contribute to empiric studies, and imply ideas of reform for legal education and professional practice.

Kurzzusammenfassung

Im Ermittlungsverfahren, das durch die Staatsanwaltschaft geleitet wird, werden Entscheidungen getroffen, die sich auf den gesamten Strafprozess auswirken. In vielen juristischen Situationen gibt es aber nicht die eine richtige Entscheidung, insbesondere wenn keine eindeutige Beweislage vorliegt. Anhand von Entscheidungsmodellen wie Dual-Prozess-Theorien lassen sich Faktoren ableiten, die den Entscheidungsprozess beeinflussen. Diese Studie hatte zum Ziel zu untersuchen, wie Menschen unterschiedlicher Expertise eine nicht eindeutige Beweislage in einem Kriminalfall handhaben und welche prozess- und personenbedingten Faktoren sich darauf auswirken. Für dieses fragebogenbasierte Quasi-Experiment wurde eine gemischte Stichprobe aus juristischen Laien, Noviz:innen (im Studium oder Referendariat) sowie Expert:innen rekrutiert. Nach der randomisierten Präsentation einer Fallvignette, die entweder das Delikt der Körperverletzung oder des Diebstahls beinhaltete, galt es, sich aus staatsanwaltschaftlicher Perspektive für oder gegen das Vorliegen des Tatverdachtes zu entscheiden und Angaben zum Entscheidungsprozess sowie zum Umgang mit den Beweismitteln zu machen. Eine randomisierte Teilstichprobe wurde stellenweise unter Zeitdruck gesetzt. Das Erleben von Zeitdruck wirkte sich nur geringfügig und größtenteils nicht signifikant auf den Prozess aus. Der behandelte Delikttyp spielte in einigen Betrachtungen wider Erwarten eine kleine, aber bedeutsame Rolle. Als personenbedingte Faktoren wurden die kognitive Reflexion sowie das Kognitionsbedürfnis erhoben, die kleine, teils nicht signifikante Effekte ausübten. Die Expertise erwies sich in den meisten Analysen als signifikanter Faktor: Laien und Expert:innen zeigten mitunter sogar konträres Antwortverhalten. Aus den Studienergebnissen lassen sich die methodische Eignung der neu konstruierten Vignetten, die generelle Bereitschaft juristischer Stichproben zur Teilnahme an empirischer Forschung sowie Veränderungsimpulse für die juristische Ausbildung und Praxis ableiten.

Inhaltsübersicht

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CRT	Cognitive Reflection Test
NFC	Need for Cognition
PCS	Parallel Constraint Satisfaction
StGB	Strafgesetzbuch
StPO	Strafprozessordnung
WYSIATI	„what you see is all there is”

1 Einleitung

Es ist kein abgeschlossenes rechtswissenschaftliches Studium notwendig, um einen Einblick in das Rechtssystem zu erhalten. Ehrenamtliche Schöffinnen und Schöffen repräsentieren als juristische Laien das Volk in der Rechtsprechung, insbesondere an den Strafgerichten.[1] Sie nehmen mit Berufsrichter:innen an Hauptverhandlungen teil und wirken an der Urteilsfindung mit (Lieber & Sens, 2019a; Machura, 2016). Den Großteil der Hauptverhandlung macht die Beweisaufnahme aus, während der die vorhandenen Beweise eingeführt werden. Nur die Informationen, die gemäß dem Mündlichkeitsprinzip in der Verhandlung besprochen werden, dürfen in die Urteilsfindung einbezogen werden. Sobald die Beweisaufnahme geschlossen und die Schlussvorträge der Staatsanwaltschaft und der Verteidigung gehalten wurden, ziehen sich die Laien und Berufsrichter:innen zur Urteilsberatung zurück. Können Menschen ohne jegliche juristische Ausbildung die Aufgaben der Urteils- und Entscheidungsfindung angemessen erfüllen? Unterscheiden sich Laien von Fachpersonen? Sind Fachpersonen nicht besser für diese Aufgaben geeignet?

Was beide Gruppen, also Laien und Berufsrichter:innen, bei der Erfüllung der Aufgaben *gemeinsam* haben, ist, dass sie ihre Entscheidung nur auf Grundlage der in der Verhandlung eingeführten Beweise treffen und das gefundene Urteil dahingehend begründen müssen. Doch woher kommen diese Beweise, deren Würdigung ausschlaggebend für ein Urteil ist? Die Staatsanwaltschaft leitet das Ermittlungsverfahren, das auf einen Anfangsverdacht gegen eine beschuldigte Person hin eingeleitet wird und den Beginn des Strafprozesses darstellt (Schroeder & Verrel, 2017). Die Polizei ist dabei ein wichtiges Hilfsorgan, da sie Ermittlungen übernimmt und die Staatsanwaltschaft über ihre Erkenntnisse in Aktenform informiert. Mittels der Akte entscheidet die Staatsanwaltschaft, ob ein hinreichender Tatverdacht gegen eine beschuldigte Person vorliegt. Liegt er nicht vor, wird der Fall eingestellt. Liegt er vor, gilt es den nächsten Verfahrensschritt zu bestimmen und beim Gericht zu beantragen. Dazu zählen das Erheben der Anklage, die Einstellung wegen Geringfügigkeit sowie die Einstellung unter Auflagen und Weisungen. Wird

[1] Die Autorin ist seit 2019 als Schöffin tätig.

eine Anklage erhoben, kommen – je nach Prozessvoraussetzung – ein/e oder mehrere Berufsrichter:innen und ehrenamtliche Richter:innen zusammen, um die Hauptverhandlung zu führen. Die Hauptverhandlung ist, vereinfacht gesagt, das Verfahren, in dem ein tatsächliches Urteil gesprochen wird: “In the simplest form, the verdict is either guilty or not guilty” (Hupfeld-Heinemann & Helversen, 2009, S. 275). Deren Inhalt orientiert sich aufgrund des Mündlichkeitsprinzips und der Zweistufigkeit der Beweisaufnahme an den im Ermittlungsverfahren erzielten Erkenntnissen. Die Beweise, mit denen Berufsrichter:innen (und je nach Prozessvoraussetzung auch ehrenamtliche Richter:innen) konfrontiert werden, wurden somit bereits zuvor von der Staatsanwaltschaft begutachtet und gewürdigt. Betrachtet man den langen prozessualen Weg, den Beweismittel im Strafprozess nehmen, entsteht die Sorge, dass es auf diesem Weg zu Fehlern kommt (Sagana, 2018).

Das allgemeine Vorgehen wird durch die Strafprozessordnung (StPO) geregelt, sodass Fehler eigentlich nicht auftreten dürften, es vermutlich aber tun. Diese können einerseits inhaltlicher Natur sein (z. B. fehlerhafte Umsetzung gesetzlicher Vorgaben), andererseits können sie auch von den am Prozess Beteiligten abhängen (z. B. kognitive Denkfehler bei Richter:innen). Schmittat (2017) formuliert es folgendermaßen: „Diesem normativen Programm der StPO stehen jene psychologischen Mechanismen gegenüber, die bei der menschlichen Wahrnehmung, Erinnerung und Entscheidungsfindung in der Regel im Hintergrund ablaufen und im Strafverfahren zu systematischen Verzerrungen führen können.“ (S. 444). Dazu kommt, dass die Würdigung von Beweisen zur Urteilsfindung im deutschen Strafverfahren recht frei von Vorgaben stattfindet (§ 261 StPO).

Die eigentliche Grundlage für die Beweiswürdigung und das daraus resultierende Urteil stammt – wie bereits beschrieben – aus dem Ermittlungsverfahren. Folglich spielt die Staatsanwaltschaft als Herrin dieses Verfahrens, welches letztlich ein Hauptverfahren grundlegend vorbereitet, eine tragende Rolle. Das Hauptverfahren steht aber oft im Fokus der empirischen Urteils- und Entscheidungsforschung (z. B. Guthrie et al., 2001, 2007; Rachlinski & Wistrich, 2017; s. auch Schmittat et al., 2022). Doch aufgrund der Relevanz, die die während der Ermittlungen getroffenen Entscheidungen auf den weiteren Prozess haben, setzt diese Studie noch weit vor dem Hauptverfahren an, nämlich bei der sich am Ende des Ermittlungsverfahrens stellenden Frage nach dem Vorliegen des Tatverdachts. Diese Frage gilt es anhand

der Beweislage zu beantworten. Die Aufgabe der Staatsanwaltschaft, darüber eine Erkenntnis zu erlangen, wird allerdings durch bestimmte Rahmenbedingungen erschwert. Damit ist gemeint, dass das Justizsystem als bösartige Lernumgebung kaum Möglichkeiten bietet, aus gemachten Entscheidungs- und Urteilserfahrungen zu lernen (Rachlinski, 2000). Damit ist auch gemeint, dass die hohe Arbeitsbelastung, die diese Behörde betrifft, dazu führt, dass „die Staatsanwaltschaften zum Nadelöhr bei der Strafverfolgung“ werden (Rebehn, 2021, Abs. 3; s. auch 2.1.5). Auch aufgrund dieser fehlenden personalen und zeitlichen Ressourcen wird vom Gesetz her die Polizei unterstützend tätig (§ 161 StPO). Dies bringt aber auch mit sich, dass die Staatsanwaltschaft in einigen Fällen von der Polizei über Ermittlungsergebnisse nur unterrichtet wird, ohne wirklich federführend beteiligt gewesen zu sein. Somit können auch prozess- und personenbedingte Faktoren auf Seiten der Polizei bereits eine Fehlerquelle für das Strafverfahren darstellen (Sagana, 2018). Die Staatsanwaltschaft hat am Ende des Ermittlungsverfahrens demnach die komplexe Aufgabe, wichtige Entscheidungen auf Grundlage von nicht zwingend selbst ermittelten Beweismitteln zu treffen. „First, the factfinder can never be absolutely certain of the facts in dispute” (Kagehiro & Stanton, 1985, S. 160). Dies ist besonders herausfordernd, wenn die Beweislage die beschuldigte Person nicht eindeutig be- oder entlastet oder gar widersprüchlich ist. Dies ist auch dann herausfordernd, wenn es sich inhaltlich um häufig vorkommende Delikte handelt, die ein routiniertes, ressourcenschonendes Vorgehen „provozieren“.

Von Expert:innen wird erwartet, entsprechend ihrer Expertise richtig zu handeln. Mit Blick auf die Rechtsprechung wäre es angemessen, dass „in einer eher datengeleiteten Vorgehensweise … die strafrelevanten Merkmale zusammengetragen und zu einem Urteil verdichtet [werden]“ (Bieneck, 2006, S. 160). Dies ist anspruchsvolle Denkarbeit. Menschen unter Zeitdruck greifen oft auf schnelle, intuitive Verhaltensstrategien zurück, um den Mangel an Zeit zu kompensieren (Rice & Trafimow, 2012; Rieskamp & Hoffrage, 2008). Dies könnte angesichts der hohen Belastung ebenfalls auf die Mitarbeitenden der Staatsanwaltschaft zutreffen. Auch bereichsspezifische Expertise kann mit intuitivem, heuristischem Verhalten einhergehen (Mishra et al., 2015; J. K. Phillips et al., 2004). Fraglich ist dabei, inwiefern sich in wichtigen juristischen Entscheidungssituationen auf Intuition verlassen werden sollte, zumal die primäre Informationsgrundlage für solche Entscheidungen

(polizeiliche Akte) qualitative Mängel aufweisen kann. Die Forschung zeigt, dass juristisches Verhalten durchaus fehlerhaft sein kann und Urteile durch extralegale Faktoren – wie dem Zeitpunkt einer Verhandlung vor der Mittagspause (Danziger et al., 2011) – beeinflusst werden. Juristische Fachpersonen sind nicht gegen Denkfehler immun (Rassin, 2020). Zudem deuten Disparitäten in den Entscheidungen juristischer Fachpersonen auf unterschiedliche Vorgehensweisen hin (Maguire, 2010). Eine Dynamik aus prozess- und personenbedingten Einflussfaktoren (z. B. Überlastung und kognitive Fehleranfälligkeit) wirkt sich demnach auf den Strafprozess aus, und somit auch auf zu treffende Entscheidungen und Urteile. Doch um welche (extralegalen) Faktoren handelt es sich bei der zentralen Einschätzung über das Vorliegen des Tatverdachtes in einem Delikt mit nicht eindeutiger Beweislage? Inwiefern unterscheiden sich juristische Laien, Noviz:innen und Expert:innen? Was bedeutet dies für die akademische Ausbildung und die berufliche Praxis? Die Studie möchte zur Beantwortung ebenjener Fragen einen Beitrag leisten.

Die Übertragung von allgemeinpsychologischen Mechanismen auf einen bestimmten Anwendungsbereich (hier: strafrechtlicher Kontext) ermöglicht eine größere Alltagsnähe. Aus psychologischer, entscheidungsforschender Sicht sind sowohl die Beweiswürdigung als auch alle anderen Entscheidungen von Interesse, die in den verschiedenen Verfahrensphasen getroffen werden (Schmittat, 2017). Haben kognitive Verzerrungen oder Denkfehler bei der Ermittlung der Beweise bereits eine Rolle gespielt, so werden diese negativen Einflüsse aufgrund der Zweistufigkeit der Beweisaufnahme durch das Verfahren „weitergereicht“. Die Entscheidung, die am Ende des Ermittlungsverfahrens getroffen wird, ist somit wegweisend (Ellison & Brennan, 2016; Schmittat et al., 2022). Aus diesem Grund gebührt diesem Zeitpunkt im Strafprozess eine besondere Bedeutung, weswegen er im Fokus der vorliegenden Studie, die als Quasi-Experiment konzipiert wurde, steht. Der theoretische und empirische Hintergrund (s. 2) befasst sich zunächst losgelöst von psychologischen Inhalten mit dem Ablauf des Strafprozesses, um für die Studie relevante Informationen abzuleiten (s. 2.1). Anschließend beginnt die Darstellung der psychologischen Kerninhalte (s. 2.2), die wiederum im Abschnitt 2.3 mit den juristischen Themen in Verbindung gebracht werden. In den genannten Abschnitten werden die hier dargestellten einleitenden Aspekte weiter ausgeführt und ergänzt.

2 Theoretischer und empirischer Hintergrund

Da in der vorliegenden Studie ein Transfer von empirischen Erkenntnissen auf den Strafprozess stattfindet, erfolgt zum besseren Verständnis zunächst eine zusammenfassende Übersicht über dessen Ablauf, die Verfahrensschritte und die beteiligten Personen (s. 2.1).[2] Dieser Einblick in das juristische Setting steht vorerst möglichst losgelöst vom psychologischen Kernthema, der menschlichen Entscheidungsfindung. Dieses Kernthema ist Gegenstand in Abschnitt 2.2 und wird ebenfalls weitestgehend ohne Bezug zum juristischen Kontext dargestellt. Im Abschnitt 2.3 erfolgt schließlich die Zusammenführung der beiden Themenbereiche. Einen Überblick über die Studie enthält Abschnitt 2.4. Das Kapitel endet mit einer Ausformulierung der Hypothesen und Fragestellungen (s. 2.5).

2.1 Der Strafprozess in Deutschland

Zu Beginn wird anhand eines strafprozessbezogenen Überblicks erläutert, wie der Strafprozess in Deutschland geregelt ist und abläuft, welche Personen beteiligt sind und welche Ziele verfolgt werden (s. 2.1.1). Die einzelnen Verfahrensstadien werden in den Abschnitten 2.1.2 (Ermittlungsverfahren), 2.1.3 (Zwischenverfahren) und 2.1.4 (Hauptverfahren) gesondert betrachtet, bevor in einem Exkurs auf die Anzahlen von Verfahren der Staatsanwaltschaft und der deutschen Strafgerichte eingegangen wird (s. 2.1.5). Die Beweismittel und deren Würdigung sind Gegenstand in Abschnitt 2.1.6. Abschließend steht in Abschnitt 2.1.7 ein Zwischenfazit, um sich nach dieser Zusammenfassung des Strafprozesses auf die für den weiteren Verlauf der Arbeit relevanten Inhalte fokussieren zu können.

2.1.1 Strafprozessbezogener Überblick

Das Strafprozessrecht und die Strafprozessordnung (StPO) regeln den Ablauf des Strafverfahrens und die rechtliche Stellung der Verfahrensbeteiligten. Der erstinstanzliche Strafprozess, in dem ein Sachverhalt zum ersten Mal verhandelt wird und in dem somit bisher noch kein Urteil gefällt wurde, kann in drei Abschnitte unter-

[2] Die Begriffe *Strafprozess* und *Strafverfahren* werden in dieser Arbeit synoym verwendet. Dies gilt auch für die Begriffe *Hauptverfahren* und *Hauptverhandlung*.

gliedert werden (s. Abbildung 2.1). Im Ermittlungsverfahren erforscht die Staatsanwaltschaft den Sachverhalt, wenn konkrete Tatsachen es nach kriminalistischer Erfahrung möglich erscheinen lassen, dass eine verfolgbare Straftat vorliegt. Darauf folgt das Zwischenverfahren, in dem das Gericht entscheidet, ob ein Hauptverfahren eröffnet wird. Dieses Hauptverfahren wird nur eröffnet, wenn im Zwischenverfahren ein hinreichender Tatverdacht bestätigt wird. Dieser liegt vor, wenn bei vorläufiger Tatbewertung auf Grundlage des Ermittlungsergebnisses die Verurteilung in der Hauptverhandlung bei vollgültigen Beweisen wahrscheinlicher ist als ein Freispruch. Anschließend – je nach gerichtlicher Entscheidung im Zwischenverfahren – beginnt das Hauptverfahren vom Aufruf zur Sache bis hin zur Verlesung des Urteils. Das „Verfahren im ersten Rechtszug“ (§§ 151–295 StPO) stellt nahezu chronologisch den Ablauf der prozessualen Schritte dar, wohingegen die „Allgemeinen Vorschriften“ (§§ 1–150 StPO) für das gesamte Verfahren gelten. Straftaten, die im Strafverfahren abgeurteilt werden, und auch ihre Strafbewehrung sind im Strafgesetzbuch (StGB) festgelegt.

Verfahrensbeteiligte sind die beschuldigte Person, Vertreter:innen der Verteidigung und der Staatsanwaltschaft sowie Mitarbeitende der Polizei. Das Gericht, das aus Berufsrichter:innen und gegebenenfalls ehrenamtlichen Richter:innen zusammengesetzt ist, gilt streng genommen nicht als verfahrensbeteiligte Instanz, sondern besitzt eine Art übergeordnete Funktion (Schroeder & Verrel, 2017). Je nach Zeitpunkt im Strafprozess variieren die Bezeichnungen für eine verdächtige Person (Lieber & Sens, 2019a). Eine Person ist Tatverdächtige:r, wenn ein Anfangsverdacht auf eine Straftat besteht. Wird zusätzlich ein Ermittlungsverfahren eingeleitet, gilt die Person als Beschuldigte:r. Im Zeitraum zwischen der Erhebung der öffentlichen Anklage und der Eröffnung der Hauptverhandlung lautet die Bezeichnung Angeschuldigte:r. Als Angeklagte:r gilt jemand nach der Eröffnung der Hauptverhandlung. Jemand wird zur verurteilten Person, wenn eine Strafe rechtskräftig festgesetzt wurde.

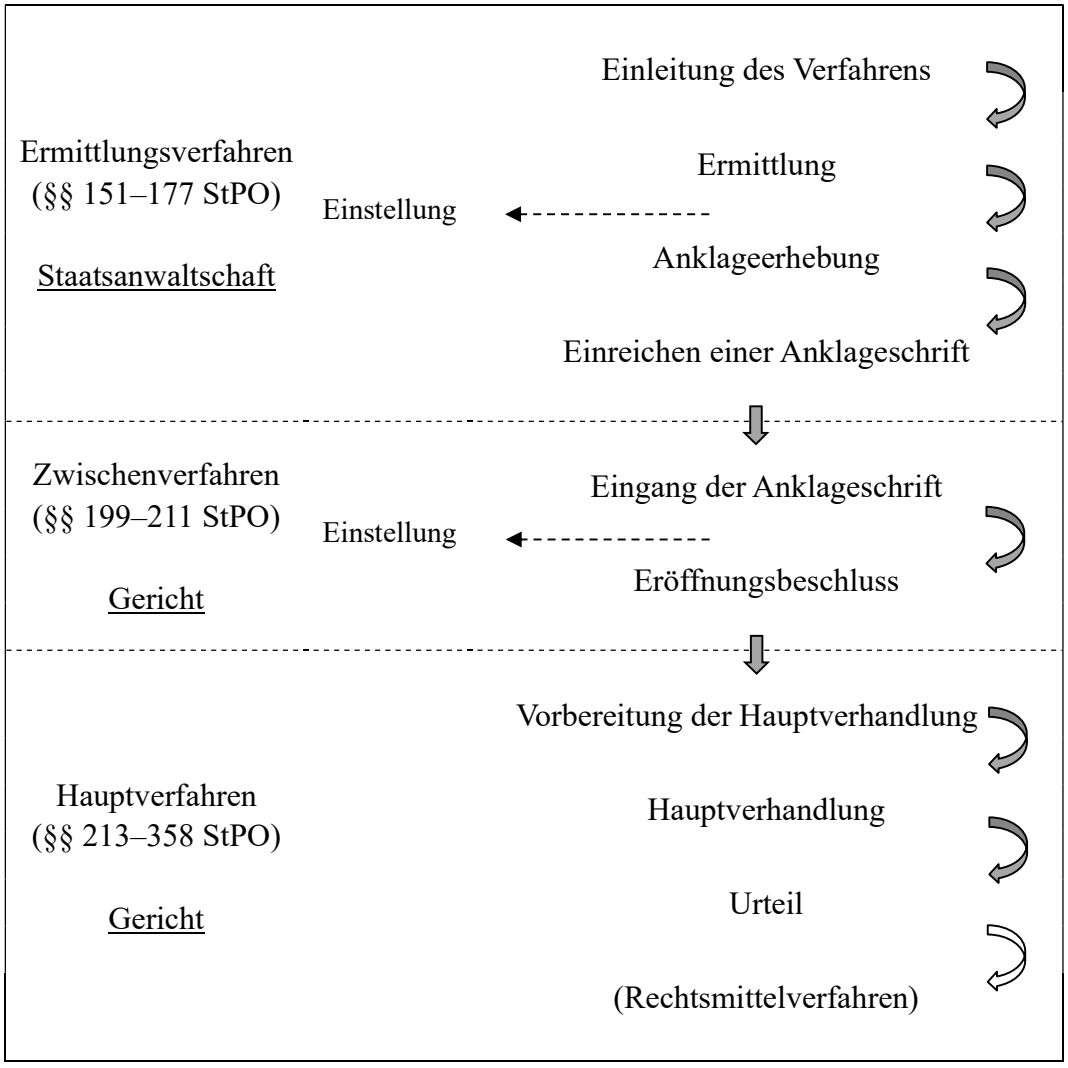

Abbildung 2.1. Verkürzter, schematischer Ablauf des Strafprozesses vom Beginn des Ermittlungsverfahrens bis zum Ende des Hauptverfahrens mit Benennung der jeweils zuständigen Behörde oder Instanz.
Anmerkungen. In Anlehnung an Schroeder und Verrel (2017, S. 239).

Zu den Zielen des Strafprozessrechts und der StPO gehören die Feststellung der Schuld (Verurteilung) oder Unschuld (Freispruch), das Sicherstellen der Rechtsstaatlichkeit des Verfahrens (Schutz der Verfahrensbeteiligten, z. B. vor Willkür) und das Herstellen des Rechtsfriedens (Beilegung von durch die Straftat hervorgerufenen Konflikten). Insgesamt sollen eine gerechte und wahre Entscheidung getroffen sowie Straftaten und ihre Täter:innen zuverlässig festgestellt werden (Lieber & Sens, 2019b; Schroeder & Verrel, 2017). In einem Strafverfahren wird demnach festgestellt, ob „Tatsachen bewiesen werden können, aus denen auf eine strafbare Handlung geschlossen werden kann“ (Lieber & Sens, 2019b, S. 45). Aus diesem Grund sind die verhandelnden Gerichte Tatsachengerichte und die drei prozessua-

len Abschnitte (Ermittlungs-, Zwischen- und Hauptverfahren) können als Erkenntnisverfahren beschrieben werden (Krey & Heinrich, 2018; Schroeder & Verrel, 2017). Für eine Verurteilung sind bestimmte Voraussetzungen nötig. Es gibt drei Elemente, die eine Tat zu einer Straftat machen: Tatbestand (sowohl objektiv als auch subjektiv), Rechtswidrigkeit und Schuld. Lieber und Sens (2019a) ziehen zur Veranschaulichung das Beispiel der Körperverletzung gemäß § 223 StGB heran:

> Wer einen anderen durch einen Faustschlag körperlich verletzt [*objektiver Tatbestand*], die Verletzung will bzw. billigend in Kauf nimmt [*subjektiver Tatbestand*], für die Verletzung keinen Rechtfertigungsgrund (z. B. Notwehr) hat [*Rechtswidrigkeit*] und sich über das Verbotene seines Tuns im Klaren ist [*Schuld*], wird wegen vorsätzlicher Körperverletzung, § 223 StGB [*verwirklichter Straftatbestand*] mit Geld- oder Freiheitsstrafe bestraft [*Rechtsfolge*]. (S. 19)

Ist eines der drei Merkmale nicht zweifelsfrei festzustellen, darf die angeklagte Person gemäß dem Zweifelssatz – lateinisch: in dubio pro reo – nicht verurteilt werden. Es kann zwischen verschiedenen Formen der Täterschaft (z. B. Allein- oder Mittäterschaft) und der Teilnahme (z. B. Anstiftung, Beihilfe) unterschieden werden. Zudem gibt es differenzierbare Phasen einer Straftat (z. B. Planungsphase oder Beendigungsstadium), wobei eine Handlung zu unterschiedlichen Zeitpunkten strafbar werden kann (Lieber & Sens, 2019a).[3]

2.1.2 Das Ermittlungsverfahren

Sobald eine Ermittlungsbehörde (hier: Staatsanwaltschaft) beispielsweise in Form einer Strafanzeige einen Anfangsverdacht erhält, wird ein förmliches Verfahren eingeleitet (§§ 151–177 StPO). Die Staatsanwaltschaft ist zur Strafverfolgung (Legalitätsgrundsatz; § 152 StPO) und zur Objektivität (§ 160 StPO) verpflichtet. Im Ermittlungsverfahren erforscht die Behörde den Sachverhalt, sammelt und sichert Beweise zur Aufklärung einer Straftat (Ackermann et al., 2022). Die Staatsanwaltschaft kann dabei auf Grundlage der StPO über den Einsatz einer Vielzahl von Ermittlungsmethoden entscheiden, um Erkenntnisse zu gewinnen (für eine Übersicht wird insbesondere auf die Abschnitte sieben und acht der StPO verwiesen). Unter wissenschaftliche Methoden fallen beispielsweise Verfahren wie der Vergleich von

[3] An dieser Stelle sind Ausführungen nur begrenzt möglich. Auf weitere relevante Voraussetzungen einer Verurteilung (z. B. Vorsatz, Fahrlässigkeit) oder andere Charakteristiken von Straftaten wird an dieser Stelle nicht eingegangen, sondern auf die Literatur von Lieber und Sens (2019a) und Schroeder und Verrel (2017) sowie auf die StPO verwiesen.

Fingerabdrücken, Untersuchungen von Blutproben, DNA-Analysen, Schriftvergleiche oder Untersuchungen von Stoffen auf deren Zusammensetzung hin. Als rechtsmedizinische Methode gilt die Leichenschau. Auch polizeiliche Observationen und Beobachtungen, Fahndungen, die Überwachung technischer Mittel (z. B. Telefon), Beschlagnahmungen oder Durchsuchungen (Wohnungen, Räume, Personen, Sachen) können Erkenntnisse liefern. Auch wenn es eine Vielzahl von Ermittlungsmitteln und -methoden gibt, so ist deren sinnvoller Einsatz abhängig von den konkreten Umständen eines Einzelfalles und der zu überprüfenden mutmaßlichen Straftat. Ein Unterschied zwischen den Methoden liegt darin, ob die beschuldigte Person davon Kenntnis hat oder ob sie ohne deren Wissen stattfinden. Ein weiterer Unterschied ist, welches Ziel diese Methoden verfolgen, denn sie können der Beweisermittlung oder -sicherung dienen (Schroeder & Verrel, 2017). Da einige dieser Maßnahmen einen schweren Eingriff in die Grundrechte der von ihnen betroffenen Personen bedeuten, ist eine vorherige Anordnung bei der/dem ermittelnden Richter:in einzuholen (Krey & Heinrich, 2018; Schroeder & Verrel, 2017). Die Staatsanwaltschaft entscheidet sich für eine Methode und die/der Ermittlungsrichter:in beurteilt, ob dem stattgegeben wird, sofern die StPO dies voraussetzt. Teilweise darf die Staatsanwaltschaft – und sogar die Polizei – auch ohne richterliche Anordnung tätig werden, z. B. bei Gefahr im Verzug, wenn also ein Nicht-Handeln zu Beweisverlust oder anderweitig zu Schaden führen könnte (Schroeder & Verrel, 2017; s. auch Büchner, 2022). Ermittlungsrichter:innen werden somit auf den Antrag der Staatsanwaltschaft hin tätig und schätzen die Rechtmäßigkeit, aber nicht die Zweckmäßigkeit der beantragten Maßnahme ein. Dadurch bleibt die Herrschaft über das Ermittlungsverfahren bei der Staatsanwaltschaft (Schroeder & Verrel, 2017).

Am Ende des Ermittlungsverfahrens erfolgt durch die Staatsanwaltschaft eine Beweiswürdigung (s. 2.1.7), aus der sich die nächsten prozessualen Schritte für den Strafprozess ergeben. Es gilt, dass die „Beweiswürdigung … nicht gegen Denkgesetze (Regeln der Logik) verstoßen [darf]“ (Schroeder & Verrel, 2017, S. 112). Die Ermittlungen dienen zur Vorbereitung der Entscheidung des Gerichts darüber, ob eine Straftat bewiesen werden kann und ob die/der Tatverdächtige:r auch die/der tatsächliche Täter:in ist (Ackermann et al., 2022). Obwohl für die Einleitung eines Ermittlungsverfahrens zureichende Anhaltspunkte ausreichen (Anfangsverdacht),

so muss an dessen Ende aber ein genügender Anlass (hinreichender Verdacht) vorliegen, damit die Staatsanwaltschaft Anklage erheben kann (Krey & Heinrich, 2018; Schroeder & Verrel, 2017). Für die Staatsanwaltschaft gilt es folgende Frage zu beantworten: „Würde man an Stelle des Gerichts bei dieser Beweislage zu einer Verurteilung kommen?“ (Büchner, 2022, S. 88). Begründet die Beweislage keinen hinreichenden Verdacht, muss das Verfahren eingestellt werden. Eine Einstellung kann auch erfolgen, wenn kein/e Tatverdächtige:r oder die Nichtschuld der/des Beschuldigten ermittelt wurde, wenn sich eine Tat als nicht strafbar herausgestellt hat oder wenn Prozessvoraussetzungen (z. B. Verhandlungsfähigkeit der/des Beschuldigten) nicht erfüllt sind (Schroeder & Verrel, 2017).

Neben den beiden Verfahrensoptionen (Einstellung des Verfahrens oder Erheben einer Anklage) ergeben sich aus der StPO noch weitere Entscheidungsmöglichkeiten für die Staatsanwaltschaft. Gilt für diese Behörde normalerweise das Legalitätsprinzip, also die Verpflichtung zur Strafverfolgung, so gibt es Ausnahmen gemäß dem Opportunitätsprinzip. In bestimmten Fällen kann von einer Strafverfolgung abgesehen werden, wenn zwischen der Straftat und dem Verfahren keine Verhältnismäßigkeit vorliegt. Gründe für eine Einstellung wären die Geringfügigkeit der Schuld und das fehlende öffentliche Interesse an der Strafverfolgung (§ 153 StPO) oder der Verzicht auf Verfolgung unter Auflagen und Weisungen (§ 153a StPO). Bei letzterem wird auf das Erheben einer Anklage verzichtet, aber die beschuldigte Person erhält Auflagen und Weisungen, „wenn diese geeignet sind, das öffentliche Interesse an der Strafverfolgung zu beseitigen, und die Schwere der Schuld nicht entgegensteht“ (§ 153a Abs. 1 StPO). Verfahrenseinstellungen basierend auf dem Opportunitätsprinzip führen zu einer Entlastung der jeweiligen Gerichte und kommen nicht selten vor, sodass eigentlich nicht von Ausnahmen gesprochen werden kann (Schroeder & Verrel, 2017; s. auch 2.1.5). Dies bedeutet aber nicht, dass es nicht erst zu Ermittlungen kommt, sondern dass das Ergebnis der Ermittlungen für eine Einstellung nach dem Opportunitätsprinzip spricht (auch wenn ein Tatverdacht besteht). In Tabelle 2.1 sind die für die vorliegende Studie relevanten Entscheidungsoptionen zusammengefasst. Die Staatsanwaltschaft (nicht die Polizei) beendet mit ihrer Entscheidung in Form der Abschlussverfügung das Ermittlungsverfahren (Krey & Heinrich, 2018).

Tabelle 2.1. Übersicht der Entscheidungsoptionen der Staatsanwaltschaft hinsichtlich nächster Schritte am Ende des Ermittlungsverfahrens

Entscheidungsoption	Rechtliche Begründung und Verortung (StPO)
Anklage	Bieten die Ermittlungsergebnisse genügend Anlass, erhebt die Staatsanwaltschaft öffentliche Klage beim Gericht. (§ 170 Abs. 1)
Einstellung	Bieten die Ermittlungsergebnisse nicht genügend Anlass, wird das Verfahren von der Staatsanwaltschaft eingestellt. (§ 170 Abs. 2)
Einstellung wegen Geringfügigkeit	Wird die Schuld des Täters als gering angesehen und besteht kein öffentliches Interesse an einer Verfolgung, kann von einer Verfolgung wegen Geringfügigkeit abgesehen werden. (§ 153 Abs. 1)
Einstellung unter Weisungen und Auflagen	Von der Erhebung der Anklage wird vorläufig abgesehen und stattdessen werden Auflagen und Weisungen erteilt, z. B. eine Form der Wiedergutmachung, Erbringen von gemeinnützigen Leistungen. (§ 153a Abs. 1)

Auch wenn sie in dieser Arbeit nicht weiter berücksichtigt werden können, so gibt es Sonderformen der Anklageerhebung, die Alternativen zum herkömmlichen Ablauf des Strafverfahrens darstellen (Schroeder & Verrel, 2017). Unter anderem fallen darunter der Antrag auf Aburteilung im beschleunigten Verfahren, „wenn die Sache auf Grund des einfachen Sachverhalts oder der klaren Beweislage zur sofortigen Verhandlung geeignet ist“ (§ 417 StPO) sowie der Antrag auf Erlass eines Strafbefehls. Das Strafbefehlsverfahren (§§ 407–412 StPO) ist ein vereinfachtes Verfahren, das ohne eine Hauptverhandlung auskommt, wenn am Ende der Ermittlungen festgestellt wird, dass eine derartige Verhandlung nicht nötig ist, um die Strafe (Rechtsfolge) einer Tat festzulegen. Das Strafbefehlsverfahren stellt neben den Varianten der Verfahrenseinstellung eine Möglichkeit dar, das Justizsystem zu entlasten, da keine aufwendigen Hauptverhandlungen durchgeführt werden müssen (Schroeder & Verrel, 2017).

2.1.3 Das Zwischenverfahren

Bieten die Ergebnisse am Ende des Ermittlungsverfahrens genügend Anlass, so kann die Staatsanwaltschaft Anklage erheben. Andernfalls kann mit Zustimmung des Gerichts ein Verfahren bereits zu diesem Zeitpunkt des Strafprozesses eingestellt sein. Sobald die Fallakte und die staatsanwaltschaftliche Anklageschrift dem zuständigen Gericht vorliegen, beginnt das Zwischenverfahren (§§ 199–211 StPO). Es gilt zu überprüfen, ob eine überflüssige Hauptverhandlung erspart bleiben kann (Combé, 2007; Schroeder & Verrel, 2017). Das Gericht entscheidet je nach Antrag, Aktenlage und Prozessvoraussetzungen, ob es zur Eröffnung der Hauptverhandlung kommt, ob noch weitere Beweismittel nacherhoben werden müssen oder ob es

Gründe zur Ablehnung des staatsanwaltschaftlichen Antrages gibt. Da das Gericht entscheidet, ob dem Antrag stattgegeben wird (§ 153 Abs. 1, § 153a Abs. 1, § 199 Abs. 1 StPO), hat nunmehr das Gericht und nicht mehr die Staatsanwaltschaft die Herrschaft über das Zwischenverfahren. Verfahrenseinstellungen sind zwar zu diesem Zeitpunkt noch möglich, allerdings ist ein Eröffnungsbeschluss zur Hauptverhandlung wahrscheinlicher (Schroeder & Verrel, 2017). Sollte das Gericht eine Einstellung erwirken wollen, muss die Staatsanwaltschaft allerdings zustimmen, „denn zu beurteilen, ob eine Tat überhaupt verfolgungswürdig ist, also ein öffentliches Interesse an der Strafverfolgung besteht, ist dann doch ureigenste Aufgabe derjenigen, die als Anwältinnen und Anwälte des Staates … auftreten“ (Büchner, 2022, S. 89).

2.1.4 Das Hauptverfahren

Den Schwerpunkt im Strafverfahren bildet die Hauptverhandlung. Sie beginnt mit dem Aufruf zur Sache und endet mit der Urteilsverkündung (§§ 213–295 StPO). Nach dem Aufruf zur Sache und nachdem die angeklagte Person zu ihren persönlichen Angaben vernommen wurde, verliest die Staatsanwaltschaft die Anklage, zu der sich die angeklagte Person äußern oder von ihrem Schweigerecht Gebrauch machen kann. In der darauffolgenden Beweisaufnahme – dem „Kernstück der Hauptverhandlung“ (Lieber & Sens, 2019b, S. 131) – geht es darum, die Rechts- und Sachlage aufzuklären. Gemäß der Unschuldsvermutung ist es die Aufgabe des Gerichts, der angeklagten Person die Straftat und somit ihre Schuld nachzuweisen (Schroeder & Verrel, 2017). Zu den Beweismittel zählen der richterliche Augenschein (z. B. Fotos, Filme, Röntgenaufnahmen; §§ 86–93 StPO), Sachverständige (§§ 72–85 StPO), Zeug:innen (§§ 48–71 StPO) oder das Verlesen von Urkunden und Schriftstücken (§§ 249–256 StPO). Diese Beweismittel dienen zur Klärung der Fragen, ob die Voraussetzungen für einen Strafprozess gegeben sind, ob die zur Last gelegte Tat der angeklagten Person nachgewiesen werden kann (oder nicht) und ob zu einer nachgewiesenen Tat strafmildernde oder -schärfende Umstände identifiziert werden können (Lieber & Sens, 2019a). Die vorgebrachten Beweise werden hinsichtlich ihrer Bedeutung und ihrer Aussagekraft bewertet, wobei diese Bewertungen letztlich eine wichtige Grundlage für die Urteilsfrage darstellen (Lieber & Sens, 2019a). Eine Besonderheit im deutschen Strafprozess ist die Zweistufigkeit der Beweisaufnahme (Schroeder & Verrel, 2017). Die Beweise, die im

Ermittlungsverfahren erhoben werden, müssen in der Hauptverhandlung erneut begutachtet werden, um gewährleisten zu können, dass nur in dieser Verhandlung besprochene Inhalte in die Urteilsfindung und Beweiswürdigung eingehen (Grundsatz der Mündlichkeit). Nach der Beweisaufnahme erfolgen die Schlussvorträge der Staatsanwaltschaft und der Verteidigung. Das letzte Wort erhält zwingend die angeklagte Person. Im Anschluss zieht sich das Gericht zur Beratung zurück, während der das Urteil unter Würdigung der Beweise gefällt wird. Diese Beratung findet auch statt, wenn das Gericht mit einer/m Einzelrichter:in besetzt ist, wobei dieser Beratungsprozess sich dann im Innern der Person selbst abspielt (Schroeder & Verrel, 2017). Das Gericht entscheidet „nach seiner freien, aus dem Inbegriff der Verhandlung geschöpften Überzeugung“ (§ 261 StPO). Zum Schluss werden vor den Verfahrensbeteiligten das Urteil sowie die Gründe dafür verlesen. Bei einem Schuldspruch wird in der Strafzumessung bestimmt, welche Strafart festgelegt wird (Geld- oder Freiheitsstrafe), welcher Strafrahmen ausgehend von der Gesetzesgrundlage (StGB) anwendbar ist und welche strafschärfenden oder -mildernden Umstände zu berücksichtigen sind (Lieber & Sens, 2019a). Ein Urteil ist rechtskräftig, wenn innerhalb einer bestimmten Frist von Seiten der Staatsanwaltschaft oder der verurteilten Person keine Rechtsmittel eingelegt werden (Berufung oder Revision bei Urteilen des Amtsgerichts oder Revision bei Urteilen des Landgerichts oder des Oberlandesgerichts). Ein Rechtsmittel führt dazu, dass ein höheres Gericht einen Fall überprüft. Bei der Berufung kommt es zu einer neuen Beweisaufnahme, da der Sachverhalt erneut überprüft wird. Bei der Revision ist dies allerdings nicht der Fall, da hier nur das Urteil hinsichtlich möglicher Rechtsverletzungen betrachtet wird (Schroeder & Verrel, 2017). Ein rechtskräftiges Urteil wird durch die Staatsanwaltschaft vollstreckt, beispielsweise indem die verurteilte Person zum Haftantritt geladen wird. Somit ist die Staatsanwaltschaft nicht nur Ermittlungs- und Anklage-, sondern auch Vollstreckungsbehörde (Lieber & Sens, 2019a).

2.1.5 Exkurs: Die Arbeit der Staatsanwaltschaft sowie der Strafprozess an deutschen Strafgerichten in Zahlen

Die Staats- beziehungsweise Amtsanwaltschaft erledigte insgesamt 4 938 615 Ermittlungsverfahren im Jahr 2019 in Deutschland (Statistisches Bundesamt, 2020a). Der Großteil der Verfahren wurde durch die Polizei eingeleitet (4 041 350), wohingegen die Ermittlungen in 705 936 Fällen durch die Staats- oder Amtsanwaltschaft

begonnen wurden. Die Vertreter:innen der Staatsanwaltschaft (3 222 960) beendeten mehr Verfahren als die Mitarbeitenden der Amtsanwaltschaft (1 715 691).[4] Unter den insgesamt erledigten Verfahren waren 462 826 Straftaten gegen das Leben und die körperliche Unversehrtheit vertreten, darunter überwiegend die vorsätzliche Körperverletzung (457 885). Weitaus mehr Straftaten bezogen sich auf Eigentums- oder Vermögensdelikte (1 564 287), die sich wiederum in Diebstahl und Unterschlagung (592 451) sowie Betrug und Untreue (971 836) aufgliederten.

In Tabelle 2.2 ist auszugsweise dargestellt, auf welche Art die Ermittlungsverfahren erledigt wurden. Deutschlandweit kam es in 8.48% der Verfahren zu einer Anklage, wohingegen in 28.5% eine Einstellung nach § 170 Abs. 2 StPO erfolgte (s. auch Tabelle 2.1). Betrachtet man die relativen Anzahlen für die Straftat der vorsätzlichen Körperverletzung, so kam es mehrheitlich zur Einstellung (45.24%) und nicht zur Anklage (12.03%). Für das Sachgebiet „Diebstahl und Unterschlagung“ war diese Tendenz in gleicher Richtung, aber weniger stark ausgeprägt (Einstellung: 19.31%, Anklage: 13.92%). Das Einbeziehen der Einstellungen nach § 153a StPO und nach § 153a Abs. 1 StPO würde die relativen Anzahlen jeweils noch erhöhen. Kam es tatübergreifend zu einer Anklage, wurde die Mehrzahl der Verfahren vor dem Amtsgericht (409 326) verhandelt (Landgericht: 9 383). Die am häufigsten bei einer Einstellung verhängte Auflage war das Zahlen eines Geldbetrags an eine gemeinnützige Einrichtung oder an die Staatskasse (141 598 von 167 561).

Von den insgesamt erledigten Ermittlungsverfahren im Sachgebiet „vorsätzliche Körperverletzung“ wurde der Großteil durch die Polizei (424 655) und eine geringere Anzahl durch die Staatsanwaltschaft (32 707) eingeleitet (Statistisches Bundesamt, 2020a). Ein ähnliches Bild zeigt sich beim Sachgebiet „Diebstahl und Unterschlagung“, da die Mehrzahl der Verfahren ebenfalls durch die Polizei (520 216) begonnen wurde (Staatsanwaltschaft: 71 973). Die durchschnittliche Verfahrensdauer vom Tag des Eingangs bei der Staats- oder Amtsanwaltschaft bis zur Erledigung durch ebenjene Behörden betrug 1.7 Monate. Wurde das Ermittlungsverfahren durch eine andere Behörde eingeleitet, vergingen im Mittel 3.7 Monate bis zu

[4] Siehe Fußnote 31.

dessen Erledigung. Dabei dauerte es ab dem Tag der Einleitung des Verfahrens durchschnittlich 2 Monate bis zum Eingang bei der Staats- oder Amtsanwaltschaft.

Tabelle 2.2. Übersicht ausgewählter Statistiken zu Ermittlungsverfahren in Deutschland im Jahr 2019 (Statistisches Bundesamt, 2020a)

Von der Staatsanwaltschaft beim Landgericht und von der Amtsanwaltschaft erledigte Ermittlungsverfahren	
Erledigte Verfahren insgesamt	4 938 651
Anklage	418 709
Antrag auf Entscheidung im beschleunigten Verfahren (§ 417 StPO)	13 785
Antrag auf Erlass eines Strafbefehls	547 665
Einstellung mit Auflage	167 561
Einstellung ohne Auflage	1 214 311
Einstellung wegen Geringfügigkeit (§ 153a Abs. 1 StPO)	454 927
Einstellung nach § 170 Abs. 2 StPO	1 407 425
Erledigte Ermittlungsverfahren in Strafsachen im Sachgebiet „vorsätzliche Körperverletzung“	
Erledigte Verfahren insgesamt	457 885
Anklage	55 099
Antrag auf Entscheidung im beschleunigten Verfahren (§ 417 StPO)	270
Antrag auf Erlass eines Strafbefehls	28 903
Einstellung mit Auflage nach § 153a StPO	11 879
Einstellung wegen Geringfügigkeit (§ 153a Abs. 1 StPO)	20 643
Einstellung nach § 170 Abs. 2 StPO	207 166
Erledigte Ermittlungsverfahren in Strafsachen im Sachgebiet „Diebstahl und Unterschlagung“	
Erledigte Verfahren insgesamt	592 451
Anklage	82 466
Antrag auf Entscheidung im beschleunigten Verfahren (§ 417 StPO)	6 000
Antrag auf Erlass eines Strafbefehls	61 148
Einstellung mit Auflage nach § 153a StPO	16 901
Einstellung wegen Geringfügigkeit (§ 153a Abs. 1 StPO)	74 505
Einstellung nach § 170 Abs. 2 StPO	114 396

Der Vergleich über die Zeit hinweg zeigt eine Verschiebung der Erledigungen durch die Staatsanwaltschaft in der Bundesrepublik (Jehle, 2019). So nahm seit 2003 bis 2017 bei relativ konstant bleibender Gesamtzahl der Fälle die Anzahl der Einstellungen ohne Auflagen zu, wohingegen die Einstellungen unter Auflagen, Anträge auf Strafbefehle sowie Anklagen zahlenmäßig sanken. Jahn (2015) bezeichnet die Staatsanwaltschaft als „Einstellungsbehörde“ (S. 41), da eine Vielzahl an Ermittlungsverfahren eingestellt wird und es oftmals gar nicht erst zu einer

Hauptverhandlung kommt. Seit 2003 zeigt sich – im Vergleich zu *staatsanwaltschaftlichen* Entscheidungen – in den *gerichtlichen* Entscheidungen eine gewisse Konstanz in der Anzahl der Urteile, der Strafbefehle und der Einstellungen mit und ohne Auflagen (Jehle, 2019).

Doch auch an den Strafgerichten kommen Einstellungen häufiger vor als eine Anklageerhebung, wenngleich es regionale Unterschiede in der Handhabung dessen gibt (Schroeder & Verrel, 2017). Die ausgewählten Statistiken bieten einen Überblick über die Verfahren an deutschen Strafgerichten für das Jahr 2019, getrennt nach Amts- und Landgerichten (Statistisches Bundesamt, 2020b). Strafverfahren, bei denen aufgrund der Anklage von einer Freiheitsstrafe bis zu zwei Jahren ausgegangen werden kann, werden am Amtsgericht verhandelt. Fälle, die über die Zuständigkeit des Amtsgerichts hinsichtlich der zu erwartenden Freiheitsstrafe hinausgehen, gehen an das Landgericht. In Tabelle 2.3 sind die Statistiken der Amtsgerichte dargestellt.

Tabelle 2.3. Übersicht ausgewählter Statistiken zu Strafverfahren an deutschen Amtsgerichten im Jahr 2019 (Statistisches Bundesamt, 2020b)

Anzahl der Strafverfahren an deutschen Amtsgerichten	
Erledigte Verfahren insgesamt	660 816
Verfahren vor der/dem Strafrichter:in	438 849
Verfahren vor dem Schöffengericht	39 186
Verfahren vor dem erweiterten Schöffengericht	307
Art der Erledigung der Strafverfahren an deutschen Amtsgerichten	
Erlass eines Strafbefehls (§ 408a StPO)	29 758
Urteil	263 333
Einstellung mit Auflage oder Weisung (§ 153a StPO)	51 815
Einstellung wegen Geringfügigkeit (§ 153 Abs. 2 StPO)	30 254
Einstellung wegen Abwesenheit der beschuldigten Person oder wegen eines anderen in dieser Person liegenden Hindernisses (§ 205 StPO)	26 706
Einstellung wegen Verfahrenshindernisses (§ 206a StPO)	2 999
Ablehnung der Eröffnung des Hauptverfahrens	2 071

Anmerkungen. Es wird das Erwachsenenstrafrecht betrachtet.

Es wird deutlich, dass rund zwei Drittel der erledigten Verfahren vor einzelnen (Straf-)Richter:innen verhandelt wurden. Außerdem erfolgte in Relation zur Gesamtzahl nur in einem Bruchteil der Verfahren die Ablehnung der Eröffnung der Hauptverhandlung (0.31%). Einstellungen kamen dagegen recht häufig vor. Am Amtsgericht machten 2019 die Eigentums- und Vermögensdelikte den Großteil der

Delikte aus (209 162). Darunter fielen Diebstahl, Unterschlagung, Betrug und Untreue. An zweiter Stelle standen Straftaten im Straßenverkehr (113 073), gefolgt von Straftaten gegen das Leben und die körperliche Unversehrtheit (78 840, darunter insbesondere die vorsätzliche Köperverletzung: 78 795) und Straftaten nach dem Betäubungsmittelgesetz (65 364). Die durchschnittliche Dauer der Verfahren betrug 4.3 Monate. Dabei nahm die Hauptverhandlung durchschnittlich 1.2 Tage in Anspruch. Von insgesamt 287 689 Beschuldigten erhielten 252 203 eine Verurteilung und 27 445 einen Freispruch.

Die Tabelle 2.4 beinhaltet die Statistiken der Landgerichte. Einstellungen waren zwar vertreten, kamen aber relativ gesehen seltener vor als bei den Amtsgerichten (Statistisches Bundesamt, 2020b). Vergleicht man die Anzahlen der insgesamt erledigten Verfahren zwischen Amts- und Landgerichten, zeigt sich mengenmäßig die hohe Auslastung ersterer.

Tabelle 2.4. Übersicht ausgewählter Statistiken zu Strafverfahren an deutschen Landgerichten im Jahr 2019 (Statistisches Bundesamt, 2020b)

Anzahl der Strafverfahren an deutschen Landgerichten	
Erledigte Verfahren insgesamt	14 039
Verfahren vor der Großen Strafkammer	9 044
Art der Erledigung der Strafverfahren an deutschen Landgerichten	
Urteil	9 200
Einstellung mit Auflage oder Weisung (§ 153a StPO)	259
Einstellung wegen Geringfügigkeit (§ 153 Abs. 2 StPO)	84
Einstellung wegen Abwesenheit der beschuldigten Person oder wegen eines anderen in dieser Person liegenden Hindernisses (§ 205 StPO)	214
Einstellung wegen Verfahrenshindernisses (§ 206a StPO)	86

Anmerkungen. Es werden das Erwachsenenstrafrecht und erstinstanzliche Verfahren (ohne Berufung) betrachtet.

Straftaten nach dem Betäubungsmittelgesetz wurden an den Landgerichten am häufigsten verhandelt (2 894), gefolgt von Straftaten gegen das Leben und die körperliche Unversehrtheit (2 767), Eigentums- und Vermögensdelikten (1 291) sowie Wirtschafts-, Steuerstrafverfahren und Geldwäschedelikten (1 121). Die durchschnittliche Verfahrensdauer betrug 8 Monate. Dabei nahm die Hauptverhandlung durchschnittlich 4.8 Tage in Anspruch. Von insgesamt 12 758 Beschuldigten wurden 11 787 verurteilt und 957 freigesprochen. An den Landgerichten fanden zwar insgesamt weniger Verfahren statt als an den Amtsgerichten, aber deren zeitlicher

Aufwand war im Durchschnitt um das Vierfache höher. An beiden Gerichten kamen Verurteilungen häufiger vor als Freisprüche, wobei die relativen Anteile der Freisprüche vergleichbar waren (Amtsgericht: 9.6%; Landgericht: 7.5%) .

Während eines Strafverfahrens gilt das Beschleunigungsgebot, um eine unverhältnismäßige Belastung der Verfahrensbeteiligten zu vermeiden. Dies gilt für die beschuldigte Person, für die Geschädigten, aber auch für Mitarbeitende der Staatsanwaltschaft, der Verteidigung und des Gerichts, da jedes Verfahren Personalkapazitäten bindet. Es muss die Balance zwischen Geschwindigkeit und Gründlichkeit der Ermittlungen gefunden werden (Jehle, 2019; s. auch Europäische Menschenrechtskonvention Art. 6 Abs. 1). Im Jahr 2017 dauerten erstinstanzliche Verfahren am Landgericht ab Eingang beim Gericht durchschnittlich 7.7 Monate (Jehle, 2019). Zählt man den Zeitraum der Ermittlungen durch die Staatsanwaltschaft hinzu, ergeben sich 19.1 Monate. Am Amtsgericht beläuft sich die Dauer auf durchschnittlich 4 Monate exklusive beziehungsweise 8 Monate inklusive des Ermittlungsverfahrens.

2.1.6 Beweismittel, Indizien und deren Würdigung: Wahrheitsbegriff und Beweismaß im Strafprozess

Beweismittel werden während des Ermittlungsverfahrens erhoben und können zu verschiedenen Zeitpunkten im Strafverfahren nachermittelt werden. Dies beinhaltet auch das Ermitteln von Hinweisen, die eine beschuldigte Person entlasten (§ 160 Abs. 2 StPO). Der Staatsanwaltschaft stehen verschiedene Ermittlungsmethoden und -formen zur Verfügung (s. 2.1.2). Deren Ergebnisse können bestimmten Kategorien von Beweismitteln zugeordnet werden. Dazu zählen:

- Sachverständige (§§ 72–85 StPO),
- Urkunden und Schriftstücke (§§ 249–256 StPO),
- der richterliche Augenschein (z. B. Fotos, Filme; §§ 86–93 StPO),
- Zeug:innen (§§ 48–71 StPO) sowie
- die Einlassung der beschuldigten Person (§ 157 StPO).

Letzteres wird zwar, im Gegensatz zu den anderen, nicht als förmliches Beweismittel des Strafprozesses angesehen, trägt aber dennoch zur Beweisführung bei (Ackermann et al., 2022).

Die Beweiswürdigung beschreibt den Prozess der Bewertung von Beweisen, während dessen die richterliche Überzeugungsbildung stattfindet (Schweizer, 2015).[5] Der Beweis ist die Voraussetzung dafür, dass sich Richter:innen eine Meinung zu einem Sachverhalt bilden können (Kern, 2019). Im Strafverfahren unterliegt die Urteilsfindung dem Grundsatz der freien Beweiswürdigung (§ 261 StPO), sodass es keine festen Regeln oder Abläufe dahingehend gibt, welche Beweise wie gewertet werden. Dabei wird im Hauptverfahren auch Laien (ehrenamtlichen Richter:innen) zugetraut, eine solche Einschätzung abliefern zu können, denn für die Bewertung der Beweiskraft sollen eher Lebenserfahrung und keine speziellen juristischen Kenntnisse herangezogen werden (Lieber & Sens, 2019a; Machura, 2016). Die freie Beweiswürdigung bringt dabei lediglich eine subjektive und keine objektive Gewissheit mit sich (Ackermann et al., 2022). Beweismittel unterscheiden sich in der Beweiskraft dahingehend, wie sie Richter:innen beeinflussen und von der Wahrheit überzeugen können (Schweizer, 2015). Ackermann et al. (2022) definiert: „Beweisen heißt, dem beurteilenden Gericht einen Sachverhalt durch jedermann überzeugende und beliebig oft reproduzierbare Fakten so darzustellen, dass ein vernünftiger Zweifel an dem … angenommenen Tatgeschehen nicht möglich ist“ (S. 56). Ein Hauptbeweis einer Partei dient dazu, das Gericht von der Wahrheit einer Tatsachenbehauptung zu überzeugen, sodass das nötige Beweismaß überschritten wird. Ein Gegenbeweis der anderen Partei kann das Maß der Überzeugung wiederum schmälern (Schweizer, 2015). Es kann zwischen Haupttatsachen und Indizien (Indizientatsachen) unterschieden werden. Letztere können einen Schluss darüber zulassen, ob die Haupttatsachen wahr sind (Schweizer, 2015). Die Aussagen von Zeug:innen sind Indizientatsachen und können als Indizienbeweis (oder: indirekter Beweis) eine Haupttatsache unterstützen beziehungsweise widerlegen. Mit dem Indizienbeweis ist ein Denkprozess verknüpft, der die Indizien- und die Haupttatsache miteinander in Beziehung setzt (Schweizer, 2015). Ein *direkter* Beweis einer Haupttatsache liegt somit nur dann vor, wenn Richter:innen diesen Beweis selbst in Augenschein nehmen können.[6] Doch wann gilt eine Behauptung als wahr? Wie werden

[5] Da sich die Empirie auf die richterliche Beweiswürdigung im Hauptverfahren fokussiert (s. auch 2.3), wird im Folgenden insbesondere von Richter:innen gesprochen, wenngleich ein Transfer auf die Beweiswürdigung durch andere Fachpersonen im Ermittlungsverfahren möglich ist.

[6] Im weiteren Verlauf der Arbeit wird überwiegend von Beweismitteln gesprochen und nicht näher zwischen Haupt- und Indizientatsachen oder direkten und indirekten Beweisen differenziert.

Beweise gewürdigt? Darauf wird in den Abschnitten 2.1.6.1 und 2.1.6.2 eingegangen.

2.1.6.1 Wahrheitsbegriff und Beweismaß

Das Finden der Wahrheit – und damit einhergehend das Erreichen eines Beweismaßes – steht im direkten Zusammenhang mit dem Prozess der Beweiswürdigung. Eine Straftat wird durch das Strafverfahren erst rekonstruierbar und begreifbar (Schroeder & Verrel, 2017). Im Verfahren geht es nicht darum, die Wahrheit über Tatsachen zu ermitteln, sondern es gilt, die Wahrheit über Tatsachenbehauptungen und Aussagen einzuschätzen. Die Tatsache „Loch im Kleidungsstück" ist nicht wahr oder falsch, aber die Behauptung „Es ist ein Loch im Kleidungsstück" kann wahr oder falsch sein (in Anlehnung an Schweizer, 2015, S. 22). Um „wahr" zu sein, müssen Tatsachen*behauptungen* gemäß der Korrespondenztheorie mit der Wirklichkeit korrespondieren, um als materielle Wahrheit gelten zu können (für eine ausführliche Auseinandersetzung mit dem Wahrheitsbegriff s. Schweizer, 2015). Da aber in einem Fall niemals die absolute Wahrheit gefunden werden kann, darf eine absolute Sicherheit nicht als Standard der Beweiswürdigung gelten (Holländer, 2019; Schweizer, 2015). Insbesondere der Nachweis der Verursachung lässt sich nicht vollends ermitteln, sondern kann nur geschlussfolgert werden (z. B. ob die erhöhte Geschwindigkeit einen Unfall verursacht hat; Althammer & Tolani, 2019). Das Ziel der Beweiswürdigung ist es nicht, Richter:innen von der Wahrheit zu überzeugen, sondern diese Überzeugung ist das Mittel, mit dem das eigentliche Ziel – nämlich die Feststellung der (materiellen) Wahrheit – erreicht werden soll (Schweizer, 2015). Richter:innen erhalten aber nur in den seltensten Fällen Rückmeldung darüber, ob ihre Rekonstruktion des Sachverhalts mit der tatsächlichen Wirklichkeit übereinstimmt (Schweizer, 2019; s. 2.3.3.3).

Zur Überzeugungsbildung ist ein Beweismaß notwendig. Das Beweismaß beschreibt dabei den nötigen Grad der Überzeugung, der erreicht sein muss, um einen Sachverhalt als wahr anzuerkennen (Althammer & Tolani, 2019; Schweizer, 2015, 2019). Es kann zwischen einer überwiegenden Überzeugung und einer vollen Überzeugung unterschieden werden (Schweizer, 2016). Ersteres entspricht einem Wert von über 50%, letzteres einem Wert von über 90% und grenzt somit an absolute Sicherheit. Diese Zahlen sind zwar nicht von den Justizsystemen festgelegt, ermöglichen aber eine Veranschaulichung des Überzeugungsgrades (Schweizer, 2016).

Allerdings stellen die beiden Maße lediglich eine subjektive und keine objektive Einschätzung der Wahrscheinlichkeit oder des Ausmaßes der Überzeugung dar (Kern, 2019; Schweizer, 2019). Sie kommen zudem je nach Land und je nach Rechtsgebiet (Zivil- oder Strafrecht) unterschiedlich zum Einsatz (für eine Übersicht über europäische Länder s. Tichý, 2019). Im deutschen Strafrecht steht die volle Überzeugung (über 90%) im Vordergrund (Althammer & Tolani, 2019; Schweizer, 2016). Eine Erklärung für die Höhe des Überzeugungsgrades sind die mit der Entscheidung verbundenen Kosten: Ein Schuldspruch für eine eigentlich unschuldige Person muss eher vermieden werden als die Freilassung einer tatsächlich schuldigen Person, sodass man mit hoher Sicherheit von der Schuld überzeugt sein muss (Schweizer, 2016, 2019). Der Grad der Überzeugung ist somit nicht dichotomer („überzeugt“ beziehungsweise „nicht überzeugt“), sondern vielmehr gradueller Natur, wobei eine volle Überzeugung immer noch geringer ist als die absolute Sicherheit (100%; Schweizer, 2019). Intuition reicht dabei aber nicht aus, um zur vollen Überzeugung zu gelangen (Schweizer, 2015). Wurden Rechtsexpert:innen in einer Studie von Schweizer (2016) angesichts eines Fallbeispiels direkt befragt, gaben sie ein Maß der vollen Überzeugung als erforderlich an (über 90%). Wurde deren Überzeugungsgrad allerdings statistisch ermittelt, lag dieser unterhalb der genannten prozentualen Grenze. Folglich zeigt sich in der Empirie, dass das tatsächlich gelebte und das theoretisch erforderliche Beweismaß nicht zwingend übereinstimmen.

2.1.6.2 Modelle der Verarbeitung und Würdigung von Beweisen

Aus juristischer Sicht lässt sich Beweiswürdigung definieren „als der *Vorgang der richterlichen Überzeugungsbildung* zur Wahrheit von Tatsachenbehauptungen“ (Schweizer, 2015, S. 13). Kurz gesagt unterscheiden sich Beweiswürdigungsmodelle im Ausmaß der richterlichen Freiheit und des Schutzes vor Willkür. In Deutschland herrscht ein gemischtes subjektiv-objektives Modell vor (Schweizer, 2015). Die damit einhergehende freie Beweiswürdigung und ein eher abstraktes Beweismaß ermöglichen es, die (materielle) Wahrheit darüber zu ermitteln, ob sich eine Straftat tatsächlich zugetragen hat (Ermittlung des Sachverhalts). Durch diese Freiheit lässt sich auf die Gegebenheiten der individuellen Fälle eingehen (Kern, 2019). Auch wenn die Beweiswürdigung zwar frei von gesetzlichen Regeln ist, soll sie sich dennoch an die Grundsätze der Logik und Rationalität halten. Außerdem

dürfen Richter:innen widersprüchliche Beweise und Informationen nicht einfach ignorieren, sondern sie müssen vielmehr das Gesamtbild der Beweise betrachten. Auch eine freie Beweiswürdigung muss vom Gericht nachvollziehbar begründet werden und darf nicht frei von Überprüfbarkeit sein (Schweizer, 2015). Nichtsdestotrotz wird die Beweiswürdigung durch dieses Vorgehen der „personality of the factfinder" (Kern, 2019, S. 53) überlassen.

Wie würdigen Menschen Beweise aus psychologischer Sicht? Diese Fragestellung ist an dieser Stelle von der für die Studie zentralen Frage abzugrenzen, wie sich die durch die Würdigung der Beweise gewonnenen Erkenntnisse anschließend im (Entscheidungs-)Verhalten widerspiegeln (s. 2.4). Neben heuristischen Ansätzen der Entscheidungsfindung gibt es kohärenzbasierte Ansätze (Hupfeld-Heinemann & Helversen, 2009; Schweizer, 2015; Towfigh & Glöckner, 2015). Zwei Modelle werden an dieser Stelle exemplarisch dargestellt: Das *Story Model* (Pennington & Hastie, 1986, 1991, 1992), das als Erklärungsansatz für Entscheidungen in juristischen Kontexten dient, und das *Parallel Constraint Satisfaction Model* (PCS; Glöckner & Betsch, 2008a), dessen Annahmen auch auf den Entscheidungsprozess im Rahmen der Beweiswürdigung übertragbar sind.[7] Zunächst gilt es, die Bedeutung von Kohärenz zu beschreiben. Schweizer (2015) argumentiert, dass die Menge an und die Komplexität von Informationen, die die Beweismittel bereitstellen, nicht ohne Intuition verarbeitet werden können. Intuition beschreibe dabei einen informationsverarbeitenden Prozess, der schnell, ohne Anstrengung und eher unbewusst ablaufe und in einem die Entscheidung beeinflussenden Gefühl ende. Eine der vier Kategorien von Intuition nach Glöckner und Witteman (2010) beschreibt die konstruktive Intuition, nach der mentale Repräsentationen im Sinne der Kohärenzbildung konstruiert werden. Dies bedeutet, dass die Fülle an Informationen aufbereitet wird, sodass als Endergebnis ein stimmiges, kohärentes Ganzes entsteht (s. auch Glöckner & Ebert, 2011).

[7] Die Modelle wären auch in Abschnitt 2.2 (Theorien der Entscheidungsfindung) passend, da ebendiese Modelle ebenfalls als Erklärungsansätze für den Prozess der (juristischen) Entscheidungsfindung dienen. Aufgrund der intendierten Argumentation im Hinblick auf den Vorgang der Beweiswürdigung und aufgrund der Tatsache, dass der Versuchsablauf eine bereits fertige Geschichte beinhaltete (s. 3.4.1), findet eine inhaltliche Abgrenzung statt. Für eine Übersicht über weitere Modelle und Theorien der Beweiswürdigung s. Maegherman (2021).

Kohärenz kann dabei auf verschiedenen Wegen erreicht werden (Schweizer, 2015). Sogenannte Geschichten-Modelle gehen davon aus, dass Beweise gewürdigt werden, indem sie zu einer Geschichte verknüpft werden, um somit narrativ kohärent und stimmig zu sein. Dazu zählt das angesprochene *Story Model* von Pennington und Hastie (1986, 1991, 1992), das auf der Forschung zur Beweiswürdigung von Geschworenen basiert, wobei sich Erkenntnisse auch auf andere Prozessbeteiligte übertragen lassen. Laut diesem Modell spielen allerdings nicht nur die Beweismittel, sondern auch der individuelle Erfahrungsschatz bei dem Versuch eine Rolle, eine möglichst überzeugende Geschichte zu konstruieren. Der Erfahrungsschatz kommt insbesondere dann zum Einsatz, wenn es fehlende Elemente in der Geschichte zu ergänzen gilt. Unterschiede in der resultierenden Geschichte – und somit auch im finalen Urteil – lassen sich laut Pennington und Hastie (1986, 1991) mit diesen individuell unterschiedlichen Erfahrungen erklären, da die eigentliche Beweisgrundlage letztlich für alle Personen gleich ist. Eine Geschichte gilt als kohärent abhängig davon, wie es um die Konsistenz (Freiheit von Widersprüchen), die Vollständigkeit und die Plausibilität der Geschichte bestellt ist (Pennington & Hastie, 1991, 1992).

Neben narrativer Kohärenz gibt es kognitive Kohärenz, die dadurch erreicht werden kann, dass widersprüchliche Beweismittel abgewertet und stimmige Beweismittel iterativ aufgewertet werden (Schweizer, 2015; D. Simon, 1998). Kohärenz kann durch das „Kleinmachen“ von widersprüchlichen Beweisen maximiert werden, sodass die entscheidende Person auch bei gegensätzlichen oder fehlenden Informationen zu einer Überzeugung gelangen kann (Schweizer, 2013). Dies erinnert an das Konzept der kognitiven Dissonanz, demzufolge Dissonanzen beziehungsweise Widersprüchlichkeiten reduziert werden, um dadurch Konsonanz zu erreichen (Festinger, 1957; s. auch Maegherman, 2021). Im *Parallel Constraint Satisfaction Model* (PCS) stellt eine Restriktion die Beziehung zwischen zwei Kognitionen dar, die es zu erfüllen gilt, um Kohärenz zu erreichen (Glöckner & Betsch, 2008a; Schweizer, 2015; D. Simon, 1998). Kognitionen, wie beispielsweise Aussagen von Zeug:innen, können der gleichen Untergruppe (z. B. „wahr“, „falsch“) angehören, sodass sie durch eine positive Verbindung miteinander verknüpft sind. Bei der Zugehörigkeit zu unterschiedlichen Untergruppen liegt eine negative Restriktion vor. Allerdings können eine als „wahr“ eingestufte und eine als „falsch“ eingestufte

Kognition stimmig sein, wenn sie inhaltlich kohärent sind (z. B. wenn beide für oder gegen die Schuld einer angeklagten Person sprechen). In Form von konnektionistischen Netzwerken angeordnet gelten positive Restriktionen zwischen Kognitionen (Knoten) als reizend, wohingegen negative Verbindungen zwischen Knoten hemmend wirken (Glöckner & Betsch, 2008a; Read et al., 1997). Somit erfahren zwei positiv verbundene Knoten stärkere Aktivität als negativ verbundene. Aktiviert ein Reiz (z. B. ein Eindruck durch ein Beweismittel) einen Knoten, so verteilt sich die Aktivität über diese reizenden Verbindungen, um dadurch die Kohärenz der Kognitionen möglichst zu maximieren. Read et al. (1997) veranschaulichen die Mechanismen des PCS-Modells folgendermaßen:

> The activation of each node can be viewed as its degree of acceptability or belief. Thus, belief in a proposition is the result of a set of multiple constraints among the nodes in the belief system. Beliefs that are mostly supported by other beliefs will be positively activated and therefore acceptable, whereas beliefs that are contradicted by many other beliefs will be negatively activated and therefore not believed. (S. 43)

Bei der andauernden und sich wiederholenden Würdigung von (neu dazukommenden) Beweismitteln während eines Verfahrens finden unbewusst Kohärenzverschiebungen in Richtung der favorisierten Entscheidung statt. Dadurch nähert sich die Aktivität im Netzwerk einer gewissen Stabilität an, aus der eine Entscheidung entstehen kann (Glöckner & Betsch, 2008a; Glöckner et al., 2010; D. Simon, 1998). Die Kohärenzverschiebung birgt allerdings die Gefahr der Überzeugung, die richtige Entscheidung getroffen zu haben. Dies kann der Fall sein, obwohl die Beweismittel dies eigentlich nicht zulassen, weil sie objektiv widersprüchlich sind oder gar fehlen, aber dennoch passend gemacht werden und dadurch kohärent wirken (Schweizer, 2013).

Die angeführten narrativen und kognitiven Kohärenztheorien entsprechen dabei einer holistischen Betrachtung von Beweisen “by holistically forming a coherent mental representation of the case“ (Schweizer, 2013, S. 65). Im Gegensatz dazu steht das atomistische (isolierte) Vorgehen „by atomistically assessing the probative value of each item of evidence and integrating the values according to an algorithm” (Schweizer, 2013, S. 65), das sich auf subjektive Wahrscheinlichkeitstheorien bezieht. Schweizer (2013) untersuchte, ob sich die Auffassung von Studierenden über die Schuld einer fiktiven angeklagten Person unterscheidet, je nachdem, ob Beweise eher auf holistische oder atomistische Art betrachtet werden. Es zeigte

sich, dass sich bei einer holistischen Betrachtung der *gleichen* Beweise die Interpretationen derjenigen, die einen Angeklagten verurteilten, stark von den Interpretationen derjenigen unterschieden, die den Angeklagten freisprachen. Bei einer atomistischen Beweiswürdigung zeigten sich diese Unterschiede in den Interpretationen nicht auf diese Weise. Zudem stimmte das Verhalten der Studierenden nicht zwingend mit deren Angaben überein. So verurteilten einige Versuchsteilnehmenden den Angeklagten, obwohl die zuvor persönlich angegebene Prozentgrenze für die Schuldfrage (rund 90%) statistisch gesehen nicht erreicht worden war. Es liegt somit die Vermutung nahe, dass gewisse Faktoren Menschen davon abhalten, einheitlich und normativ zu handeln (s. 2.2; 2.3).

2.1.7 Zwischenfazit: Ableitung von für die Studie relevanten juristischen Inhalten

Die dargestellte Auswahl an Informationen zum Ablauf des Strafprozesses in Deutschland dient zum besseren Verständnis der Herleitung der Forschungsinhalte und der Versuchsplanung.[8] Aus den Abschnitten 2.1.1–2.1.6.2 sind wesentliche Inhalte und Schwerpunkte abzuleiten. Das Ermittlungsverfahren, in dem die Sammlung von Beweismitteln erfolgt, steht zeitlich vor der Hauptverhandlung. Zum Ende dieses ersten Verfahrensschrittes sind die vier ausgewählten Entscheidungsoptionen relevant, die zu ebenjenem Zeitpunkt der Staatsanwaltschaft zur Verfügung stehen: Anklage, Einstellung, Einstellung wegen Geringfügigkeit sowie Einstellung unter Weisungen und Auflagen (s. 2.1.1; 2.1.2). Damit einhergehend ist das Wissen um die vier Kategorien von Beweismitteln nötig (Augenschein, Urkunde, Aussage von Zeug:innen und Einlassung der beschuldigten Person), da auf deren Grundlage eine Entscheidung über den weiteren Verlauf getroffen wird (s. 2.1.6).[9] Dementsprechend spielt auch das Beweismaß (s. 2.1.6.1), das zur Bildung einer Überzeugung notwendig ist, eine wichtige Rolle. Sofern ein Verfahren nicht zuvor eingestellt wird, findet die Beweisaufnahme aufgrund des Mündlichkeitsprinzips letztlich in zwei Stufen statt, nämlich im Ermittlungs- und erneut im Hauptverfahren. Soll im Hauptverfahren ein Urteil gefunden werden, so passiert dies gemäß der

[8] Aus Gründen der Lesbarkeit wird in dieser Zusammenfassung auf Literaturangaben größtenteils verzichtet. Für solche Angaben wird auf den jeweiligen Abschnitt verwiesen.

[9] Das Beweismittel *Sachverständige:r* ist für diese Arbeit nicht weiter von Bedeutung.

StPO im Anschluss an die Beweisaufnahme im Rahmen der freien Beweiswürdigung (s. 2.1.4). Da aber bereits die vorherige Würdigung der Beweismittel durch die Staatsanwaltschaft über den weiteren Verlauf entscheidet, ist diese erste Beweiswürdigung wegweisend für den Prozess (Morgan et al., 2018; D. Simon, 2012). D. Simon (2012) fasst dies folgendermaßen zusammen: „To a large degree, criminal verdicts are determined at the investigative phase, with the trial serving primarily as a ritual that delivers more symbolic than real value“ (S. 204). Im Zusammenhang mit den Beschreibungen eines nicht rein objektiven Beweismaßes sowie mit der freien Beweiswürdigung aus juristischer und psychologischer Sicht wurde bereits angedeutet, dass eben dieser recht freie Umgang mit Beweismitteln – und darauf basierende Entscheidungen – mutmaßlich durch verschiedene Faktoren beeinflusst sein kann (s. 2.1.6.2). Im folgenden Abschnitt zur Entscheidungsfindung wird beleuchtet, welche Modelle der Entscheidungsfindung relevant sind (s. 2.2). Ebenso wird ausgeführt, welche Einflussfaktoren auf den und welche individuellen Unterschiede im Entscheidungsprozess erkennbar werden. Die Übertragung der psychologischen Inhalte auf den strafrechtlichen Kontext erfolgt in Abschnitt 2.3.

2.2 Entscheidungsfindung: Theorien, Einflussfaktoren und individuelle Unterschiede

Es folgt eine Darstellung der für die Studie wesentlichen Grundannahmen von Entscheidungstheorien, in der im Zusammenhang mit Intuition, Rationalität und heuristischen Ansätzen auf solche Mechanismen eingegangen wird, die sich auf die Entscheidungsfindung auswirken können (s. 2.2.1). Der Abschnitt 2.2.2 befasst sich weiterführend mit der zeitlichen Phase nach einer Entscheidung und fokussiert die Bedeutung der Umgebung, der Konsequenzen sowie der Verantwortung auf den post-selektionalen Lernprozess. Darauf folgt eine genaue Betrachtung bestimmter, den Entscheidungsprozess beeinflussender Faktoren (s. 2.2.3), bevor in einem Zwischenfazit die psychologischen Kerninhalte abgeleitet werden (s. 2.2.4).

2.2.1 Von den normativen Grundannahmen zu Dual-Prozess-Annahmen und heuristischen Ansätzen menschlicher Entscheidungsfindung

Urteilen und Entscheiden gelten als alltägliche Denkprozesse.[10] Urteile stellen eine Art Bewertung oder Beurteilung dar und grenzen sich in der Form von Entscheidungen ab, als dass sie sich nicht mit möglichen Handlungskonsequenzen beschäftigen, sondern vielmehr als Ergebnis psychologischer Prozesse gesehen werden (Betsch et al., 2011). Eine auf ein Urteil folgende Entscheidung beruht auf der Wahl zwischen mindestens zwei Optionen, wobei es erwünschte Konsequenzen zu erreichen und unerwünschte Konsequenzen zu vermeiden gilt. Entscheidungssituationen unterscheiden sich darin, ob die Optionen bereits vorgeben oder noch zu generieren sind, und, ob es sich um eine feste oder noch offene Menge an Auswahlmöglichkeiten handelt. Betsch et al. (2011) nutzen eine Dreiteilung zur Darstellung des Prozesscharakters des Entscheidens. Die prä-selektionale Phase betrachte demnach die Generierung und das Aussuchen von Entscheidungsoptionen. Die selektionale Phase befasse sich grundlegend mit der Frage „Wie wird sich entschieden?“ und reiche von der genauen Betrachtung der Optionen bis zur tatsächlichen Entscheidungsfindung und Intentionsbildung. Nachfolgend beziehe sich die post-selektionale Phase auf Prozesse und Erfahrungen, die zeitlich nach der Entscheidung liegen. Das Ergebnis des Prozesses könne eine Feststellung, eine Intention, eine Aktivität oder das Unterlassen einer Handlung sein, wobei sich je nach Modus für die Wahl beziehungsweise die Zurückweisung einer Option entschieden würde.

Wahrscheinlichkeiten spielen in (normativen) Theorien als Determinanten eine grundlegende Rolle. Dabei greift die psychologische Entscheidungsforschung auf Ansätze der Mathematik und Ökonomie zurück. Zu nennen sind das Wert-Erwartungs-Modell (bei rationalen Entscheidungen wird die Option ausgewählt, die im Sinne der Nutzenmaximierung den höchsten erwarteten Wert erreicht) oder die Bedeutung des subjektiven Nutzens von Konsequenzen und dessen Einfluss auf die Entscheidungsfindung (für eine historische Übersicht s. Betsch et al., 2011; Pfister

[10] Sowohl im juristischen als auch im psychologischen Kontext ist vom „Urteil“ die Rede, wobei nicht das gleiche Konstrukt gemeint ist. Im psychologischen Kontext gilt ein Urteil nicht als richterlicher Urteilsspruch, sondern als eine Bewertung oder Beurteilung (Betsch et al., 2011). Was gemeint ist wird aus dem jeweiligen Zusammenhang deutlich.

et al., 2017). Allerdings verhalten sich Menschen nicht unter allen Umständen gemäß den Annahmen der normativen, „idealen“ Ansätze (Stanovich & West, 2000). Während normative Ansätze eine ideale, rationale Entscheidung im Blick haben, befassen sich deskriptive Ansätze mit tatsächlichen Entscheidungen (Stanovich & West, 1998). Beispielsweise kann das sogenannte Framing, also die Art und Weise wie Informationen formuliert werden, die Auswahl der Optionen beeinflussen (Rahmungseffekt; Tversky & Kahneman, 1974; s. 2.2.1.2). Wenngleich sich Forschung zu Erwartungsnutzenmodellen in der Regel auf Lotterien mit Gewinnen und Verlusten – und somit auf konkrete Geldbeträge – bezieht, ist diese Konkretheit bei alltäglichen Entscheidungen selten der Fall, denn Informationen über Optionen und Konsequenzen müssen von der entscheidenden Person erst gesucht und integriert werden (Betsch et al., 2011; Pfister et al., 2017). Außerdem argumentiert H. A. Simon (1955, 1956) mit dem Konzept der begrenzten beziehungsweise eingeschränkten Rationalität, dass eine vollständige und ideale Verarbeitung von Informationen aufgrund der beschränkten kognitiven Kapazitäten des Menschen nicht möglich ist. Auch wären in der Regel nicht alle Informationen bekannt, sodass von einem grundlegenden Informationsmangel ausgegangen werden müsse. Statt der in der Nutzentheorie angewandten Maximierungsregel betrachtet H. A. Simon (1955) vielmehr Entscheidungsstrategien, die sich trotz begrenzter Rationalität einsetzen lassen. Dazu gehört die Regel der Anspruchserfüllung (Satisfizierung), die die erstbeste Option auswählt, deren Konsequenzen am nächsten zu einem festgelegten Kriterium stehen. Bei nicht-kompensatorischen Entscheidungsstrategien (wie der Satisfizierung) muss eine Option eine bestimmte Schwelle erreichen, um als wählbar zu gelten. Kompensatorische Strategien benötigen im Gegensatz dazu eine stärkere kognitive Ressource und können zu einem Konflikt für die entscheidende Person führen, wenn der Ausgleich zwischen positiven und negativen Konsequenzen nur schwer oder nicht gelingt (Pfister et al., 2017).[11] In Situationen, in denen der kognitive Aufwand steigt, ist der Einsatz ressourcenschonender, nicht-kompensatorischer (heuristischer) Strategien sinnvoll (s. 2.2.1). Dazu zählen komplexe Situationen mit zunehmender Anzahl von Entscheidungsoptionen oder solche Momente, in denen unter Zeitdruck gehandelt wird (s. 2.2.3.1). Auch unterscheiden

[11] Für eine zusammenfassende Übersicht über kognitive Entscheidungsstrategien wird auf Rieskamp und Hoffrage (1999, 2008) verwiesen.

sich Situationen nicht nur in der Menge, sondern in der Verfügbarkeit von Informationen, genauer gesagt in deren Vollständigkeit, Übersichtlichkeit und Konkretheit. Dies umschreibt Kahneman (2011) mit "what you see is all there is" (WYSIATI; vgl. S. 113).

Der kognitive Aufwand, der mit der Urteilsbildung und Entscheidungsfindung verbunden ist, stellt einen wesentlichen Faktor für Entscheidungstheorien dar: Dual-Prozess-Theorien differenzieren ebenjenes Ausmaß des Aufwands. Sie postulieren, dass auf verschiedene Prozesse zurückgegriffen werden kann, sodass es zu einem automatischen und heuristischen oder einem kontrollierten und systematischen Vorgehen kommt (für eine Übersicht der Theorien s. Evans, 2008; Stanovich et al., 2014). Der eine Prozess nutzt dabei schnell zu verarbeitende Heuristiken, während sich der andere Prozess der Logik und Wahrscheinlichkeiten bedient. Laut Pfister et al. (2017) sind Art und Umfang des kognitiven Aufwandes auf einem Kontinuum anzusiedeln. Es reiche von als routinisiert zu beschreibenden Entscheidungen, bei denen aufgrund der automatischen Auswahl aus Optionen nur wenig kognitive Ressource nötig ist, bis hin zu konstruktiven Entscheidungen, bei denen die Optionen erst generiert und Präferenzen dafür entwickelt werden müssen. Dazwischen lägen stereotype, auf einem minimalen Bewertungsprozess basierende Entscheidungen sowie reflektierte Entscheidungen, für die es zu den gegebenen Optionen noch keine abrufbaren Werte und Präferenzen gibt.

Eines der bedeutendsten Dual-Prozess-Modelle ist das Modell zu System 1 und System 2 von Kahneman (2011).[12] Der Begriff des Systems ist dabei als Metapher und nicht als physisch trennbare Komponente im Gehirn zu verstehen. Die Systeme unterscheiden sich – gemäß der Grundannahme von Dual-Prozess-Theorien (Evans, 2008) – insbesondere hinsichtlich des kognitiven Aufwandes, den sie erfordern. Das System 1 ist kontinuierlich aktiv und kann als automatisch, schnell, holistisch, unbewusst und anstrengungslos umschrieben werden. Das System 2 ist im Gegensatz dazu kontrolliert, langsam, analytisch, bewusst und kognitiv anstrengend. Laut Kahneman (2011) ist das System 1 faul und folgt dem Prinzip des geringsten (kognitiven) Aufwands. Somit benötigt der anstrengende Einsatz von System 2 Selbstkontrolle und Motivation. Zudem besitzt das System 2 nicht unendliche

[12] Dieses Buch von Kahnemann (2011) ist eine Zusammenfassung der jahrzehntelangen Forschung zum System-1/System-2-Modell.

Ressourcen, sondern es kann zu kognitiver Aus- oder Überlastung kommen. Die beiden Systeme hängen eng miteinander zusammen (Kahneman, 2011). Das System 2 wird mit Informationen aus dem System 1 gespeist, wobei eine eventuelle Korrektur der bereitgestellten Informationen wiederum mit kognitiver Anstrengung einhergeht. Das System 2 schreitet ein, wenn das System 1 in seinen automatischen Prozessen auf Schwierigkeiten stößt oder auf Aufgaben keine schnelle Antwort findet. Aber auch dieses Einschreiten kann die Prozesse des System 1 nicht vollends unterdrücken: Sie arbeiten kontinuierlich weiter, während das System 2 auf niedriger Stufe im Hintergrund läuft (Kahneman, 2011). Auch wenn die Unterteilung in zwei Systeme eher als Metapher zu verstehen ist, beziehen sich Kritikpunkte eben darauf, da dieses Modell als „oversimplified“ (Evans, 2008, S. 270) beschrieben werden kann. Evans (2008) regt an, nicht von Systemen, sondern von Typ-1- und Typ-2-Prozessen zu sprechen. Des Weiteren wird argumentiert, dass intuitive Antworten durchaus korrekt sein können, ohne dass Typ-2-Prozesse intervenieren müssen. Bago und Neys (2017) untersuchten, ob intuitive Antworten (z. B. in Aufgaben zu Basisraten) nach einer Zeit des Überlegens korrigiert werden. Auch wenn in einem Großteil der Fälle (ca. 50%) beide Antworten inkorrekt waren, so waren in rund einem Drittel der Fälle die intuitiven Antworten bereits korrekt und wurden nicht „falsch verbessert“. Dementsprechend wird diskutiert, inwiefern eine dichotome Einteilung der Prozesse sinnvoll ist und – sofern man bei dieser Aufteilung bleibt – inwiefern die beiden Prozesse parallel oder sequentiell ablaufen (Croskerry et al., 2014; Diederich & Trueblood, 2018; Evans, 2008). Pennycook et al. (2015) schlagen ein dreistufiges Modell vor, demzufolge in der ersten Stufe eine erste Antwort generiert wird und in einer zweiten Stufe eine Art Konflikt-Monitoring stattfindet. Werde ein Konflikt entdeckt, erfolge entweder eine Korrektur oder eine Rationalisierung der bereits getätigten Antwort. Auf das Auftreten möglicher Fehler und Verzerrungen in den Abläufen dieser Typ-1- und Typ-2-Prozesse wird in Abschnitt 2.2.1.2 genauer eingegangen. Zunächst erfolgt eine Differenzierung von rationalen und intuitiven Entscheidungen (s. 2.2.1.1).

2.2.1.1 Rationale und intuitive Entscheidungen im Zusammenhang mit den Dual-Prozess-Annahmen

Rationalität und Intuition werden oft mit Prozessen der Entscheidungsfindung in Verbindung gebracht. Mit Blick auf die Rationalität argumentiert Stanovich (2018),

dass es evolutionär gesehen für das menschliche Überleben und die Reproduktion nicht wichtig ist, rational zu handeln. Eine verbesserte Anpassung an die Umwelt gehe nicht zwingend mit einer verstärkten Rationalität einher. Dementsprechend sei das Auftreten von irrationalem Denken und Handeln durchaus möglich. Irrationales Verhalten ist somit solches, das vom Optimum eines normativen Modells abweicht (Stanovich et al., 2014). Rationalität ist von Intelligenz abzugrenzen, sodass auch intelligente Menschen irrationales Handeln zeigen (Stanovich, 2016; Stanovich & West, 2014).

Evans (2010) stellt Intuition und logisches, bedachtes Denken als Kontraste dar. Er vergleicht ersteres mit den Eigenschaften der Typ-1-Prozesse und letzteres mit denen der Typ-2-Prozesse (s. 2.2.1). Während Typ-1-Prozesse demnach mit intuitiven Entscheidungen in Verbindung gebracht werden, so stehen Typ-2-Prozesse im Zusammenhang mit rationalen (reflektierten) Entscheidungen (Evans, 2008, 2010). Intuitive Entscheidungen entstehen somit häufig schnell, ohne Anstrengung und automatisch, wobei nicht immer begründet werden kann, warum man zu einer bestimmten Entscheidung gekommen ist. Die wissensbasierte Intuition beruht auf dem Wiedererkennen (Pfister et al., 2017). Hinweisreize werden in einer Situation wiedererkannt und auf Grundlage der vorhandenen Gedächtnisstrukturen verarbeitet. Insbesondere bei bereichsspezifischen Fachpersonen kann (wissensbasierte) Intuition zu validen Entscheidungen führen, da deren Intuition eben auf Wissen, Erfahrung und Feedback basiert (s. 2.2.3.2). Im Gegensatz dazu ist die Intuition bei Laien eher heuristisch orientiert und es fehlt eine valide Entscheidungsbasis (Kahneman & Klein, 2009; Pfister et al., 2017).[13] Für diese heuristische Intuition ist kein (Vor-)Wissen notwendig (Pfister et al., 2017). WYSIATI spielt demnach bei intuitiven Urteilen und Entscheidungen eine wichtige Rolle, da es die Unempfindlichkeit des System 1 beschreibt, Informationen hinsichtlich ihrer Qualität und Quantität überprüfen zu können (Kahneman, 2011; s. 2.2.1): Vorhandene Informationen werden betrachtet, während Fehlendes ausgeblendet wird. Das System 1 unterdrückt somit Zweifel (im Gegensatz zum System 2). Laut Glöckner und Witteman (2010) ist Intuition im Zusammenhang mit Dual-Prozess-Modellen aber nicht als homogenes Konzept zu verstehen, da vier Kategorien ausgemacht werden

[13] Für weiterführende Ausarbeitungen zu Intuition wird auf die Arbeiten von Evans (2010), Glöckner und Witteman (2010), Hogarth (2010) und Schweizer (2015) verwiesen.

können. Dazu zählen assoziative, abgleichende, akkumulierende und konstruktive Intuition (s. auch Glöckner & Ebert, 2011). Aufgrund der für intuitive Entscheidungen relevanten Charakteristika der Typ-1-Prozesse ist der Verlass auf unzureichende Informationen ein mögliches Problem. Seltene und unwahrscheinliche Ereignisse können in ihrer Häufigkeit überschätzt werden, was wiederum zu extremen Vorhersagen führen kann (Kahneman, 2011; s. 2.2.1.2). Dennoch gilt es zu betonen, dass das System 1 dem System 2 nicht generell unterlegen ist und dass intuitive Entscheidungen nicht unausweichlich schlechter sind als reflektierte Prozesse (Bago & Neys, 2017; Croskerry et al., 2014; Evans, 2008, 2010; Hertwig & Todd, 2003). Laut einer Meta-Analyse von W. J. Phillips et al. (2016), die den Einfluss reflektierter und intuitiver Denkstile auf Entscheidungen betrachtete, war der Zusammenhang zwischen normativer Entscheidungsleistung und Intuition zwar negativ, aber nicht stark ausgeprägt (r = -.09). Ebenso argumentieren Engel und Gigerenzer (2006), dass Heuristiken nicht zwingend als zweitbestes hinter logischem Denken stehen. Doch was genau sind Heuristiken?

2.2.1.2 Heuristiken und Urteilsverzerrungen

Normative Ansätze gehen vorrangig vom Einsatz statistischer Methoden beim Entscheiden aus, wohingegen tatsächlich beobachtbare Entscheidungen von diesen Normen abweichen können (s. 2.2.1). Diese Abweichungen lassen sich mithilfe der Dual-Prozess-Annahmen erklären. Der mit diesen Annahmen im Zusammenhang stehende deskriptive Heuristiken-und-Biases-Ansatz betrachtet, wie verkürzte Strategien (entgegen der normativen Ansätze) durch eine schnelle und weniger kapazitätsfordernde Vorgehensweise zur Urteilsbildung beitragen (Tversky & Kahneman, 1974). Heuristiken dienen als Faustregeln der Urteilsbildung und Entscheidungsfindung, die in der Regel automatisch und intuitiv eingesetzt werden (Kahneman, 2011). Auch die genannte Satisfizierungsregel kann als Heuristik bezeichnet werden (Towfigh & Glöckner, 2015; s. 2.2.1). Der kognitive Aufwand beim Einsatz von Heuristiken ist relativ gering und somit aus ressourcenschonender Sicht von Vorteil. Ein Fokus des genannten Ansatzes liegt auf den systematischen Urteilsverzerrungen (Bias), die aus dem Gebrauch von Heuristiken resultieren. Heuristiken können aufgrund bestimmter Prinzipien, die mit dem System 1 in Verbindung stehen, zu derartigen systematischen Bias führen. Die assoziative Aktivierung bewirkt, dass ein im Gedächtnis aktiviertes Element zur Anregung anderer damit assoziierter

Elemente führt, die aber nicht zwingend sinnvoll oder logisch mit dem ursprünglichen Element verknüpft sind. Aktivierte Informationen erscheinen häufig kohärent, auch wenn diese möglicherweise nicht vollständig sind (WYSIATI; Kahneman, 2011; s. 2.2.1). Gewisse Bias können demnach aus der fehleranfälligen Funktionsweise des System 1 entstehen, wenn das System 2 nicht interveniert und kontrolliert. Urteilsverzerrungen sind sehr hartnäckig und schwierig zu überwinden, zumal sie größtenteils unbewusst ablaufen (Kahneman, 2011). Kahneman (2011) beschreibt, dass eine „Sinngebungsmaschinerie von System 1 … uns die Welt geordneter, einfacher, vorhersagbarer und kohärenter sehen [lässt], als sie es tatsächlich ist“ (S. 254).

Zu den klassischen Heuristiken zählen Repräsentativität, Verankerung sowie Verfügbarkeit (Tversky & Kahneman, 1974).[14] Repräsentativität stellt ein Synonym für Ähnlichkeit dar und beschreibt die Wahrscheinlichkeit, dass eine Kategorie oder ein Ereignis zu einer prototypischen Kategorie oder einem prototypischen Ereignis gehört: Je repräsentativer (ähnlicher) das beobachtbare Ereignis für einen Prototyp ist, desto wahrscheinlicher wird dessen Auftreten eingeschätzt (Tversky & Kahneman, 1974). Die heuristisch ermittelte Wahrscheinlichkeit ist allerdings subjektiv und wird häufig überschätzt, da die Basisrate für die generelle Auftretenswahrscheinlichkeit einer Kategorie oder eines Ereignisses nicht beachtet wird (Basisratenfehler). Dies ist häufig der Fall, wenn die statistische Basisrate für ein Ereignis unbekannt ist. Somit werden insbesondere unwahrscheinliche Ereignisse, die eine niedrige Basisrate haben, in ihrer Wahrscheinlichkeit überschätzt. Ebenso passiert es, dass unwahrscheinliche Ereignisse eine (relativ gesehen) zu große Gewichtung in der Entscheidungsfindung erhalten (Kahneman, 2011). Die Verankerungsheuristik bezieht sich auf das Urteilen hinsichtlich quantitativer Größen. Wird ein solches Urteil verlangt, orientiert man sich – häufig unbewusst – an einem bereits gegebenen (numerischen) Anker und passt den eigenen Wert nach oben oder unten an (Ankereffekt; Tversky & Kahneman, 1974). Dabei können Anker relevant, aber auch irrelevant für eine Aufgabe sein. Auch die Verfügbarkeitsheuristik arbeitet mit Wahrscheinlichkeiten: Je leichter man sich an ähnliche Ereignisse erinnern kann

[14] Da an dieser Stelle nur auf ausgewählte Beispiele eingegangen werden kann, wird für eine detaillierte Auseinandersetzung mit verschiedenen Heuristiken und Bias auf die Literatur verwiesen: Cooke (1991), Englich (2009), Guthrie et al. (2001), Rieskamp und Hoffrage (1999) und Schweizer (2005).

beziehungsweise je größer die Anzahl der vorstellbaren ähnlichen Ereignisse (Salienz), desto wahrscheinlicher wird das Auftreten eines Ereignisses eingeschätzt (Tversky & Kahneman, 1973). Dabei kann die mentale Präsenz von Ereignissen eine Rolle spielen, also wie schnell sich an entsprechende Informationen erinnert werden kann.

Ein ähnliches Konzept stammt von Gigerenzer et al. (1999) und lautet *Fast-and-Frugal-Heuristics*-Ansatz. Dabei werden im Vergleich zum Heuristiken-und-Biases-Ansatz aber weniger die Urteilsfehler betont, sondern vielmehr der Nutzen und die Effizienz von Heuristiken hervorgehoben, da sie schnell und sparsam einsetzbar sind (Gigerenzer et al., 1999; Hertwig & Todd, 2003). Der Ansatz beschreibt laut Gigerenzer et al. (1999) einen adaptiven Werkzeugkasten, der bestimmte Werkzeuge (Heuristiken) zur Urteilsbildung und Entscheidungsfindung beinhaltet. Diese Heuristiken seien schnell und sparsam und können je nach Kontext adaptiv eingesetzt werden. Der Werkzeugkasten beinhalte beispielsweise die Rekognitionsheuristik (Unterscheidung zwischen wiedererkannten und unbekannten Objekten), die *Take-the-Best*-Heuristik (binäre Hinweisreize werden nur so lange verglichen bis der erste Unterschied erkannt wird, auf dessen Grundlage das Urteil getroffen wird) sowie die Folge-der-Mehrheit-Heuristik (Urteil orientiert sich an der Mehrheit der Bezugsgruppe). Diese drei Werkzeuge seien beispielsweise in solchen Kontexten sinnvoll, in denen es viele Hinweisreize gibt, die es im Sinne der begrenzten Verarbeitungskapazitäten zu reduzieren gilt. Zu Fehlern könne es kommen, wenn das für den Kontext inadäquate Werkzeug beziehungsweise das richtige Werkzeug auf fehlerhafte Art und Weise genutzt wird. In neuen Situationen sei es auch möglich, dass noch keine hilfreiche Heuristik im Repertoire vorhanden ist.

Die Theorie der Attributsubstitution von Kahneman und Frederick (2005) dient als ein Erklärungsmodell für die theoretischen Grundlagen der Heuristik-Ansätze. Ausgehend von der Tatsache, dass ein Zielattribut (z. B. die Wahrscheinlichkeiten für Ereignisse) nur schwierig und unter kognitivem Aufwand zu beurteilen ist, bedient man sich gemäß der Theorie eines anderen beobachtbaren Attributs, das bei der Beurteilung helfen soll. Dieses heuristische Attribut ersetzt das nicht beobachtbare Attribut (das aber eigentlich von Interesse ist) und deutet die gesuchte Wahrscheinlichkeit an, indem es Aussagen über die Ähnlichkeit (Repräsentativität) oder

die Verfügbarkeit macht (Kahneman & Frederick, 2005). Kahneman (2011) umschreibt es so, dass sich unbewusst für die Beantwortung einer leichteren (heuristischen) Frage entschieden wird, die wiederum an die ursprüngliche schwierigere Frage angelehnt ist. Da aber das heuristische und das Zielattribut nicht identisch sind, ist dieser Prozess fehleranfällig, weil einem heuristischen Attribut zu viel oder zu wenig Gewicht gegeben wird oder weil aufgrund fehlender Informationen ein heuristisches Attribut gar nicht erst gefunden werden kann (Kahneman & Frederick, 2005).

Ergänzend zu diesem Erklärungsansatz können drei defizitäre Prozesse identifiziert werden, die in Heuristiken-und-Bias-Aufgaben zu Fehlern führen können: inadäquat gelerntes Wissen, das gescheiterte Erkennen einer Korrektur sowie die gescheiterte Durchführung einer Korrektur (Stanovich, 2018; Stanovich, 2018; Stanovich et al., 2014; Stanovich & West, 2008). Dieses Erkennen der Notwendigkeit einer Korrektur sowie die Korrektur an sich ermöglichen es, eine Attributsubstitution zu vermeiden. Ist Wissen aber inadäquat gelernt oder nicht ausreichend vorhanden, kommt es zu Fehlern in den Aufgaben. Doch auch wenn das Wissen vorhanden ist, kann die Notwendigkeit einer Korrektur weiterhin unerkannt bleiben. Somit stehen das Erkennen sowie die Durchführung der Korrektur in einer Abhängigkeit vom vorhandenen Wissen – und die Durchführung der Korrektur hängt wiederum von der erfolgreichen Erkennung ihrer Notwendigkeit ab. Dadurch kann eine schwache Leistung in Heuristiken-und-Bias-Aufgaben in gewissen Defiziten hinsichtlich der oben genannten Abläufe begründet sein, zumal diese Abläufe stark auf Typ-2-Prozessen basieren und gewisse Anstrengungen mit sich bringen (Stanovich et al., 2014). Doch wann interveniert das System 2? Thompson et al. (2011) argumentieren, dass ein Verlass auf Typ-1-Prozesse beziehungsweise ein Wechsel zu Typ-2-Prozessen mit einem Gefühl von Richtigkeit einhergeht: Waren sich Studienteilnehmende hinsichtlich ihrer ersten intuitiven Antwort nicht sicher (Gefühl nach Richtigkeit nur schwach ausgeprägt), so benötigten sie mehr Zeit, um ihre Antwort zu überdenken, und sie änderten diese mit größerer Wahrscheinlichkeit als diejenigen, die von ihrer Initialantwort überzeugter waren. Demnach muss es laut Thompson et al. (2011) einen überprüfenden Mechanismus geben, der das Signal zur weiteren Überprüfung einer Antwort gibt und dieser Mechanismus wird zumindest teilweise durch das Gefühl nach Richtigkeit mediiert. Die Stärke des Gefühls

von Richtigkeit, das auf eine heuristische und intuitive Antwort hin entsteht, hängt nicht nur mit dem Grad der Vertrautheit oder der Bekanntheit der Situation zusammen, sondern auch mit der Geschwindigkeit, mit der assoziierte Erfahrungen abgerufen werden können (Thompson, 2009). Demnach geht ein schwach ausgeprägtes Gefühl von Richtigkeit eher mit einem Einschreiten durch Typ-2-Prozesse einher. Interventionen sind auf verschiedene Weisen möglich. Dies kann ein erster expliziter Versuch sein, die heuristische Reaktion überhaupt erst zu hinterfragen. Weitere Möglichkeiten sind die Rechtfertigung der gegebenen Antwort, das Suchen einer anderen Lösung oder gar das Verwerfen der neu gefundenen Lösung und ein Rückbezug zur heuristischen Antwort (Thompson, 2009). Wurde schließlich eine Entscheidung getroffen, beginnt die post-selektionale Phase.

2.2.2 Die Bedeutung von Lernumgebungen, Konsequenzen und Verantwortung in der post-selektionalen Phase

Die post-selektionale Phase bezieht sich auf Prozesse, die zeitlich nach einer Entscheidung liegen (Betsch et al., 2011; s. 2.2.1). Im Hinblick auf die Konsequenzen lassen sich verschiedene Arten ausmachen. So gibt es Entscheidungen unter Sicherheit (erwartete und tatsächliche Konsequenzen stimmen überein), unter Unsicherheit (Eintreffen der Konsequenzen unterliegt unklaren, unbekannten Wahrscheinlichkeiten) sowie unter Risiko (Eintreffen der Konsequenzen unterliegt bekannten Wahrscheinlichkeiten). Dabei können Varianten von Unsicherheit ausgemacht werden, je nachdem, ob die Ursache der Unsicherheit extern in den Umweltbedingungen verortet wird oder ob sie intern in Abhängigkeit von der entscheidenden Person bleibt (Pfister et al., 2017). Entscheidungssituationen variieren in den Rahmenbedingungen, unter denen sie stattfinden. Umgebungen können dabei gut- oder bösartig sein (Hogarth, 2010). Ersteres beschreibt Situationen, in denen für eine Entscheidung hilfreiche und eindeutige Informationen geboten werden. Letzteres geht von Bedingungen aus, in denen die Informationslage spärlich und uneindeutig ist (Hogarth, 2010; Kahneman & Klein, 2009). Obwohl der Gebrauch von Heuristiken in beiden Umgebungen möglich ist, sollten sie aufgrund ihrer Fehleranfälligkeit nur in gutartigen Situationen eingesetzt werden (Stanovich, 2018; s. 2.2.1.2). Gutartige Umwelten zeichnen sich außerdem durch direktes Feedback und Konsequenzen für die entscheidende Person aus (Hogarth, 2010). So können sich Lernprozesse auf zukünftige Entscheidungen auswirken. Das Effektgesetz als Lerngesetz beschreibt

die erhöhte Auftretenswahrscheinlichkeit eines Verhaltens in einer Situation, wenn dieses Verhalten zuvor in einer ähnlichen Situation zu positiven Konsequenzen geführt hat (Thorndike, 1898, zitiert nach Betsch et al., 2011). Konsequenzen sind das nötige Feedback, das eine sich entscheidende Person benötigt, um daraus lernen zu können.

Eine Entscheidung zu treffen, die sich möglichst leicht vor anderen Personen rechtfertigen lässt, ist ein Motivator in der Entscheidungsfindung (Pfister et al., 2017). Das Gefühl von Verantwortung kann einerseits mit der Tendenz einhergehen, eine Entscheidung zu vermeiden, um dadurch auch einer eventuellen Rechtfertigung zu entgehen. Andererseits kann es die Tendenz des Handelns geben, da ein Nichts-Tun als schlimmer eingeschätzt wird (Pfister et al., 2017). Der Faktor der Verantwortung (oder Rechenschaftspflicht) beinhaltet allerdings mehrere Facetten. So kann es einen Einfluss haben, ob das Publikum, das ein Verhalten beobachtet, bekannt oder unbekannt ist (z. B. Konformität), zu welchem Zeitpunkt dieses Verantwortungsgefühl vorkommt (vor oder nach einer Entscheidung) und vor wem es sich zu rechtfertigen gilt (z. B. vor einer bekannten oder einer fremden Person; Lerner & Tetlock, 1999). Zudem kann es einen Unterschied machen, ob sich die Verantwortung auf den Prozess oder das Ergebnis bezieht, wobei ein Gefühl von Verantwortung für ersteres als vorteilhafter für eine Entscheidung gilt (Hoffmann et al., 2017; Lerner & Tetlock, 1999). Ist davon auszugehen, dass man sich vor Unbekannten für eine Entscheidung rechtfertigen muss, kann dies zu einer aufmerksameren Betrachtung von Hinweisen und zu einem größeren Bewusstsein für die eigenen kognitiven Prozesse führen, um zu vermeiden, sich vor besagtem Publikum zu blamieren: Bias werden reduziert. Allerdings kann das Vermeiden von Blamage solche Bias auch verstärken, indem die Wahl getroffen wird, die am leichtesten zu rechtfertigen ist (Lerner & Tetlock, 1999). Lerner und Tetlock (1999) fassen die Komplexität der Ergebnisse so zusammen, dass „accountability is a logically complex construct that interacts with characteristics of decision makers and properties of the task environment to produce an array of effects“ (S. 270). Hoffmann et al. (2017) argumentieren auf Grundlage ihrer Forschung allerdings, dass ein hohes Gefühl von Verantwortung (für den Prozess) nicht zu einer systematischeren Abwägung oder Integration von Informationen führt. Auch geht Verantwortung nicht zwingend mit akkuraten Antworten – hier: klinische Vorhersage (Ruscio, 2000) – einher. Zudem

bleibt für die entscheidende Person auch im Anschluss an eine vom Umfeld als „schlecht“ eingeschätzte Entscheidung durchaus die Möglichkeit, die eigene Wahl aufzuwerten, um Dissonanz zu reduzieren (Betsch et al., 2011). Laut Bullens et al. (2014) verhalten sich Menschen unterschiedlich, je nachdem, ob eine Entscheidung umkehrbar ist oder nicht. In ihrer Studie waren diejenigen, die eine getroffene Entscheidung nochmals ändern konnten, stärker darauf bedacht, negative Konsequenzen zu umgehen als diejenigen, deren Entscheidung unveränderlich war. Das Eintreten eines negativen Ergebnisses war für die Teilnehmenden ein Hinweis für eine falsche Entscheidung, wohingegen das Ausbleiben eines negativen Ergebnisses eine Veränderung unnötig machte. Das eigentliche Treffen einer Entscheidung beendet nicht unmittelbar den Vorgang des Entscheidens, da gemäß den obigen Ausführungen eben auch die post-selektionale Phase für (zukünftige) Entscheidungsprozesse eine Rolle spielt. Doch welche Faktoren beeinflussen diesen Prozess? Welche individuellen Unterschiede lassen sich ausmachen?

2.2.3 Einflussfaktoren und individuelle Unterschiede im Entscheidungsprozess: Zeitdruck, Expertise, kognitive Reflexion und Need for Cognition

Das Auftreten individueller Unterschiede ist in der Entscheidungsfindung von Bedeutung (Kahneman, 2011; Mishra et al., 2015; Stanovich, 2018). Die Betrachtung individueller Faktoren und Fähigkeiten gilt – neben der Untersuchung von situationalen Faktoren oder von Eigenschaften und Rahmenbedingungen der Entscheidungen – als besonders relevant (Appelt et al., 2011; McElroy et al., 2020). Für diese Studie wurden Einflussfaktoren ausgewählt, die für den Transfer der Entscheidungsforschung auf den Kontext des Strafverfahrens als sinnvoll gelten und die somit als Variablen aufgenommen wurden (s. 2.3; 3.1): Zeitdruck, Expertise, kognitive Reflexion und Need for Cognition. Drei dieser Faktoren (außer Zeitdruck) fallen unter die Beschreibung individueller Unterschiede und Fähigkeiten. In ihrer Datenbank *Decision Making Individual Differences Inventory* (DMIDI) listen und kategorisieren Appelt et al. (2011) Testverfahren, die in diesem Forschungsfeld genutzt werden. Demnach zählt die *Need-for-Cognition*-Skala (Cacioppo & Petty, 1982; s. 2.2.3.4) als epistemische Motivation und der *Cognitive*

Reflection Test (Frederick, 2005; s. 2.2.3.3) als spezifische Fähigkeit und Kompetenz. Zeitdruck ist dagegen ein situationaler Faktor (Appelt et al., 2011). Auf jeden dieser Faktoren wird gesondert eingegangen (s. 2.2.3.1–2.2.3.4).

2.2.3.1 Zeitdruck

Viele alltägliche Entscheidungen finden unter einer Art von zeitlicher Begrenzung statt (Edland & Svenson, 1993; Oh et al., 2016; Ordóñez & Benson, 1997). Zeit gilt als wesentlicher realitätsnaher Faktor in den Prozessen der Informationssuche und der Entscheidungsfindung (Savolainen, 2006). Zeit wird dabei oft als eine Form von Stress oder Druck angesehen, die sich auf die Qualität – oder ganz allgemein auf die Abläufe der genannten Prozesse – auswirken kann (Oh et al., 2016; Zakay, 1993). Zeitdruck dient somit als Stellvertreter für Stress (Liu et al., 2019). Eine weniger stressige Variante des Zeitdruckes ist ein Gefühl der Dringlichkeit (Maule & Hockey, 1993). Zeitdruck erfüllt einerseits die Funktion eines Stressors, spielt andererseits aber auch eine Rolle in der Auswahl oder dem Wechsel der kognitiven Entscheidungsstrategien (Maule & Hockey, 1993; Oh et al., 2016; Ordóñez & Benson, 1997; s. 2.2.1). Unter Zeitdruck kann für die entscheidende Person die Schwierigkeit entstehen, eine gute Entscheidung treffen zu wollen, dies aber in einer begrenzten Zeitspanne schaffen zu müssen: Um beide Ziele zu erreichen, ist unter Umständen ein Wechsel der angewandten Strategien notwendig (E. J. Johnson et al., 1993). Laut Savolainen (2006) lassen sich drei Forschungsansätze ausmachen, wie der Einfluss von Zeit auf die entscheidungsbedingte Informationssuche verstanden werden kann: Zeit kann als fundamentales Attribut einer Situation, als Qualifikation oder Voraussetzung für den Zugang zu Informationen, oder auch als Indikator für den Suchprozess selbst verstanden werden. Methodisch werden in der Regel Stichproben mit und ohne Zeitdruck miteinander verglichen, um Unterschiede in deren Leistungen oder Fähigkeiten festzustellen (Maule & Hockey, 1993).

Menschen unterscheiden sich darin, wie (gut) sie mit Zeitdruck umgehen können (Joslyn & Hunt, 1998; Rastegary & Landy, 1993). Die Beziehung zwischen Zeitdruck und einer gewissen Leistung ist allerdings nicht als linear anzusehen (Rastegary & Landy, 1993). Zeitdruck kann dazu führen, dass sich Personen an situationale Gegebenheiten anpassen und sich die Prozesse der Informationssuche

und Entscheidungsfindung verändern (Rice & Trafimow, 2012; Rieskamp & Hoffrage, 2008). Menschen passen sich an Zeitdruck an, indem sie Informationen selektiv betrachten und filtern oder indem sie Prozesse in ihrem Ablauf beschleunigen (Edland & Svenson, 1993; E. J. Johnson et al., 1993; Stiensmeier-Pelster & Schürmann, 1993). Rieskamp und Hoffrage (2008) schlussfolgern: "People select different strategies depending on the decision situation" (S. 274). Folglich ist eine der Strategien der Wechsel zu einer heuristischen Vorgehensweise, da zeitlicher Stress ansonsten die analytischen Typ-2-Prozesse behindern kann (Rice & Trafimow, 2012). So geht das Erleben von Zeitdruck beispielsweise mit der Tendenz einher, die für eine Entscheidung als schlechteste erachtete Information zu ignorieren, um den Mangel an Zeit zu kompensieren (Oh et al., 2016). Selbst milder Zeitdruck fördert den Verlass auf die Wiedererkennungsheuristik (Hilbig et al., 2012). Ebenso treten Rahmungseffekte unter Zeitdruck verstärkt auf (Guo et al., 2017). In einer Studie von Gonzalez (2004) führte sogar das wiederholte Üben einer Aufgabe (unter starkem Zeitdruck) bei Teilnehmenden nicht zu besseren Leistungen im Vergleich zu denjenigen, die weniger übten, aber mehr Zeit zur Verfügung hatten. Fand das Üben dagegen unter geringem und die eigentliche Aufgabe wiederum unter starkem zeitlichem Stress statt, hatte dies keinen negativen Einfluss auf die Leistungen. Eine heuristische Vorgehensweise war bei kurzem Üben, unter hohem Zeitdruck sowie bei Menschen mit niedrigen kognitiven Fähigkeiten zu erkennen. In einer Meta-Analyse untersuchten W. J. Phillips et al. (2016), ob intuitive oder reflektierte Denkstile die Entscheidungsleistung oder das subjektive Entscheidungserleben vorhersagen: Zeitdruck minderte nur die Effekte von Reflektion auf die Leistung, aber nicht von Intuition.

Eine Anpassung an Zeitdruck ist nicht zwingend maladaptiv (Rastegary & Landy, 1993). Allerdings zeigt sich, dass der Fokus auf die Entscheidungsgeschwindigkeit (anstelle der Genauigkeit) dazu führen kann, dass in der Anzahl, aber auch von der Qualität her weniger Informationen in den Prozess einbezogen werden (Rae et al., 2014). Laut Glöckner und Betsch (2012) kann eine hohe Entscheidungszeit aber auch dahingehend verstanden werden, dass nicht die Menge an Informationen, sondern – im Sinne des PCS-Modells (s. 2.1.6.2) – eher eine abnehmende Kohärenz zu einer längeren Verarbeitung führt. Demnach könne eine steigende Anzahl an kohärenten Informationen durchaus zu einer kürzeren und nicht zu einer verlängerten

Entscheidungszeit führen. Zuvor betonten Glöckner und Betsch (2008b) bereits, dass die Art des Vorhandenseins von Informationen einen wichtigen Einfluss hat: Sind Informationen in Entscheidungsaufgaben noch nicht vorhanden und müssen noch gesucht oder akquiriert werden, so kann ein diese Suche erschwerender Faktor (Zeitdruck) dazu führen, dass Informationen ihrer Wichtigkeit nach auf nicht-kompensatorische Weise betrachtet werden (lexikografische Strategie). Sei ein solch erschwerender Faktor nicht vorhanden, wären Menschen durchaus in der Lage viele Informationen mittels kompensatorischer Strategien zu integrieren (multiattribute-Nutzen-Strategie). Zu einem ähnlichen Schluss kommen Rieskamp und Hoffrage (2008). In deren Studie nutzten Teilnehmende mit Blick auf die für deren Umsetzung benötigte Menge an Informationen nahezu konträre Strategien: Diejenigen ohne Zeitdruck zeigten eine Tendenz zu informationsintensiven, kompensatorischen Prozessen (multiattribute-Nutzen-Strategie), wohingegen diejenigen mit Zeitdruck zum nicht-kompensatorischen Vorgehen tendierten (lexikografische Strategie).

Zeitdruck kann folglich dazu führen, dass vorherrschend nicht-kompensatorische Entscheidungsstrategien eingesetzt werden (Edland & Svenson, 1993). Dabei ist auch relevant, ob alle Informationen bereits vorhanden sind oder sequentiell hinzugefügt werden müssen: Bei der (anstrengenden) sequentiellen Suche werden neue Informationen mitunter ignoriert (Dummel et al., 2016). Die Analyse der gemessenen Antwortzeiten in einer Studie kann Hinweise liefern, welche Entscheidungsstrategien von Teilnehmenden angewandt wurden (Gaissmaier et al., 2011). Doch auch wenn sich eigentlich aufgrund der Rahmenbedingungen einer Aufgabensituation ein Strategiewechsel anbieten würde, fördert das Vorhandensein von Zeitdruck ein Beibehalten vorher gelernter Routinen (Betsch et al., 1999). Auch Ordóñez und Benson (1997) argumentieren, dass unter zeitlichem Stress gewisse Präferenzen für Strategien vorliegen, da nur die Hälfte der Teilnehmenden ihrer Studie einen Strategiewechsel als Reaktion zeigten.[15] Dennoch eignen sich heuristische Entscheidungsstrategien nicht gleichermaßen für den Einsatz unter Zeitdruck. Ein Grund ist die unterschiedliche benötigte Menge an Informationen (Bobadilla-Suarez & Love,

[15] Ordóñez und Benson (1997) vermuten, dass Teilnehmende mit niedriger Ausprägung in Need for Cognition (s. 2.2.3.4) den mit einem Strategiewechsel unter Zeitdruck verbundenen Aufwand gemieden haben. Somit war das Vorhandensein von Zeitdruck nicht die einzige Erklärung des Ergebnisses.

2018). Betrachtet man nicht zeitlichen, aber realen Lebensstress, so geht eine Anhäufung dieses Lebensstresses über 12 Monate hinweg mit verändertem Verhalten einher: Menschen mit geringer kognitiver Geschwindigkeit zeigen weniger zielgerichtetes und eher gewohnheitsmäßiges Verhalten, wohingegen Menschen mit hoher kognitiver Geschwindigkeit davon nicht betroffen sind (Friedel et al., 2017). Stress jeglicher Art hat demnach einen bedeutenden Einfluss auf menschliches Verhalten. Unterschiede lassen sich auch auf neuronaler Ebene ausmachen. Je nach Ausmaß des zeitlichen Stresses ändern sich nicht nur die eingesetzten Strategien, um eine Aufgabe zu lösen, sondern die Anpassung spiegelt sich sogar in der veränderten Aktivierung kortikaler Strukturen wider (Oh-Descher et al., 2017).

Es wurde bereits angedeutet, dass Entscheidungen unter Zeitdruck einen Zielkonflikt auslösen können, da die Güte einer Entscheidung und die dafür benötigte Zeit zueinander in Konkurrenz stehen (E. J. Johnson et al., 1993). Dieser Konflikt wird als *speed-accuracy-trade-off* (SAT) bezeichnet und beschreibt das Abwägen zwischen Schnelligkeit und Genauigkeit (Wickelgren, 1977). SAT kann beispielsweise über Instruktionen, Antwortsignale oder das Setzen von Fristen ermittelt werden (Wickelgren, 1977). Instruktionen, die entweder den Fokus für die Bearbeitung einer Aufgabe auf die Geschwindigkeit *oder* die Genauigkeit legen, können zu Veränderungen im Antwortverhalten führen (Buelow et al., 2019). Grob zusammengefasst ist ein wiederkehrender Befund der, dass ein Augenmerk auf die Geschwindigkeit die Genauigkeit der Antworten reduziert (z. B. Dambacher & Hübner, 2015; Donkin et al., 2014; Hick, 1952; Rae et al., 2014). Laut Donkin et al. (2014) wird dann unter Zeitdruck ressourcenschonend auf Informationen strategisch verzichtet, wenn diese eindeutig redundant zu bereits vorhandenen Stimuli sind. Buelow et al. (2019) nutzten die *Iowa Gambling Task*, um die Einflüsse von Zeitdruck oder Ablenkung auf Teilnehmende zu untersuchen. Es zeigte sich, dass diejenigen unter Zeitdruck zu der gleichen heuristisch geprägten Lösung kamen wie diejenigen, die beim Bearbeiten der Aufgabe abgelenkt wurden. Teilnehmende mit größerem Zeitbudget gaben dagegen eine andere Antwort. SAT muss aber nicht mit Einbußen in der Genauigkeit einhergehen, da motivierende Faktoren (z. B. Belohnung) dem entgegenwirken können (Dambacher & Hübner, 2015).

Ein wesentlicher Faktor für den Umgang mit Zeitstress ist die Expertise (s. 2.2.3.2). Beilock et al. (2004) verglichen geübte und ungeübte Golfer:innen im Sinne des

SAT. Ungeübte Spieler:innen zeigten eine bessere Leistung, wenn die Genauigkeit der Abläufe fokussiert werden sollte. Expert:innen spielten unter Geschwindigkeit besser, vermutlich, weil sie dabei auf ihr automatisiertes Können zurückgreifen konnten. Manipulierte man in einer Folgestudie zusätzlich die Handhabung der Golfschläger, sodass ein Umlernen stattfinden musste, benötigten Expert:innen zunächst auch mehr Zeit, profitierten dann aber erneut von der Instruktion zur Geschwindigkeit (Beilock et al., 2008). Ungeübte Personen zeigten auch mit dem ungewohnten Golfschläger weiterhin die Tendenz zum SAT. Dennoch ist Expertise nicht notwendigerweise ein Schutzfaktor gegen die Einflüsse von Zeitdruck. Ebenso wie Noviz:innen (Studierende) zeigten professionelle Mediziner:innen in einer Studie von Trueblood et al. (2018) die Tendenz zum SAT, wenn es um die Identifikation von Krebszellen auf Bildern ging – wenngleich die Fachpersonen insgesamt besser abschnitten. Auch Kliniker:innen verarbeiten Informationen in Entscheidungssituationen ohne Zeitdruck bewusster als in Situationen mit Zeitdruck (Byrne, 2013). Außerdem ist das Bereitstellen von Hilfsmitteln für die fundierte Entscheidungsfindung unter Zeitdruck mitunter weniger wirksam: Klinische Fachkräfte verbesserten ihre Antwortrate mithilfe einer Literatursuchmaschine unter Zeitdruck nur um rund 6%, wohingegen sich Fachkräfte ohne zeitlichen Stress um etwa 32% verbesserten (van der Vegt et al., 2020).

Joslyn und Hunt (1998) argumentieren, dass die Fähigkeit zum abstrakten Entscheiden ein Hinweis darauf ist, wie gut Menschen Entscheidungen unter Zeitdruck treffen können. Die subjektive Wahrnehmung von Zeit ist eine weitere mögliche Erklärung für den interindividuellen Umgang mit Zeitdruck (Zakay, 1993). Rastegary und Landy (1993) stellen fest: „It is likely that time pressure does not affect everyone in the same way" (S. 217). Zeitdruck wirkt aufgrund der Rahmenbedingungen eher von außen auf die Person ein, während ein Gefühl von Dringlichkeit aus der Person selbst entstammt (Rastegary & Landy, 1993). Menschen mit einem stärkeren inneren Gefühl von Dringlichkeit sind dabei resistenter gegen die Einflüsse von externem Zeitdruck, da für sie die Grenze zum Erleben von Stress höher zu verorten ist, vermutlich aufgrund ihrer ausgeprägten Übung im Umgang mit Dringlichkeit (Rastegary & Landy, 1993). Interindividuelle Unterschiede in dem Gefühl von Dringlichkeit zeigen sich auch auf neuronaler Ebene (Aktivität im Striatum; van Maanen et al., 2016). Bei Zeitdruck, der durch eine Frist ausgelöst wird, spielt die

Präzision in der Wahrnehmung von Zeit eine wichtige Rolle. Menschen mit einer genauen inneren Repräsentation von Zeit sind effizienter in der Verarbeitung relevanter Informationen und zeigen größere Vorsicht beim Geben von Antworten innerhalb der gesetzten Frist (Miletić & van Maanen, 2019). Inwiefern Zeitdruck und der Umgang damit als Belastung gesehen wird, kann mit der individuellen Handlungs- oder Zustandsorientierung zusammenhängen. Eine Person ist handlungsorientiert, wenn sie eigenes Verhalten oder Emotionen beeinflussen und verändern kann. Eine zustandsorientierte Person behält vorhandene behaviorale oder emotionale Zustände bei. Zustandsorientierte Menschen empfinden Zeitdruck eher als Belastung (Stiensmeier-Pelster & Schürmann, 1993).

2.2.3.2 Expertise

Ein Faktor, der Entscheidungsverhalten zu beeinflussen vermag, ist Expertise (Mishra et al., 2015). Die Relevanz der Betrachtung von Expert:innen als Stichprobe begründet sich wie folgt: „We depend on experts in many ways“ (Shanteau & Stewart, 1992, S. 102). Von Expert:innen wird erwartet, dass sie aufgrund ihrer Expertise gute und richtige Entscheidungen treffen (Herbig & Glöckner, 2009). Um Expertise zu untersuchen, bietet sich einerseits die absolute Herangehensweise an, die Menschen mit außerordentlichen Fähigkeiten und Talent betrachtet, oder andererseits die relative Herangehensweise, die einen Vergleich von Expert:innen und Nicht-Expert:innen fokussiert (Chi, 2006). Eine gewisse Regelmäßigkeit über eine langjährige Übungszeit hinweg ist eine wichtige Voraussetzung für den Erwerb von Expertise (Kahneman, 2011). Für die Entwicklung spielt aber weniger die Zeit, sondern vielmehr die Erfahrung eine Rolle (Hutton & Klein, 1999). Shanteau (1988) differenziert hinsichtlich der Ausprägungen zwischen Naiven, Noviz:innen und Expert:innen: Naive besitzen kein oder nur wenig Wissen über einen gegebenen Fachbereich, wohingegen Noviz:innen zwar über vermehrtes Wissen und Erfahrungen verfügen, aber noch nicht das Level der Expert:innen erreicht haben.[16] Fach- oder Sachkunde kann mittels des akademischen Titels oder der Berufsbezeichnung identifiziert werden (Chi, 2006; Shanteau, 1988).

[16] Die Begriffe *Naive* und *Laien* werden in dieser Arbeit synoym verwendet und beschreiben Nicht-Expert:innen.

Expertise ist domänenspezifisch (Hutton & Klein, 1999; J. K. Phillips et al., 2004). Laut Shanteau (1988) gilt es, zwischen Wahrnehmungs- oder kognitiven Expert:innen, Wissens- oder Diagnostik-Expert:innen sowie Beratungs- oder Handlungs-Expert:innen zu differenzieren. Im Vergleich zu Nicht-Expert:innen reagieren Fachpersonen schneller, machen weniger Fehler, erkennen bedeutsame Muster und Informationen, besitzen ein besseres domänenspezifisches Gedächtnis, verbringen mehr Zeit mit dem Verstehen eines Problems, verarbeiten Probleme tiefgehender und erkennen typische Gegebenheiten oder Abweichungen (Hutton & Klein, 1999). Es wird angenommen, dass sich Fachpersonen insbesondere im Ausmaß des domänenbezogenen Wissens von Nicht-Expert:innen abheben (J. K. Phillips et al., 2004). Expert:innen und Nicht-Expert:innen ähneln sich zwar in ihren generellen Fähigkeiten zum Schlussfolgern und Denken, doch die ungleichen domänenspezifischen Leistungen zwischen den Gruppen werden durch die Unterschiede im domänenspezifischen Wissen bestimmt (Chi, 2006). Hutton und Klein (1999) argumentieren, dass es nicht um das reine Wissen geht, sondern darum wie Expert:innen Situationen wahrnehmen und mit diesen Situationen verbundene Handlungen wiedererkennen: Entscheidungen von Fachpersonen haben eher eine wahrnehmende als eine konzeptuelle Komponente. Expert:innen nutzen häufig auf dem Wiedererkennen basierende Strategien, um Entscheidungen zu treffen (J. K. Phillips et al., 2004). Dieser Rückbezug zum Wiedererkennen erinnert an Intuition (s. 2.2.1.1). Dazu passt, dass Expert:innen häufig nicht in der Lage sind, ihr kognitives Entscheidungsverhalten in Worte zu fassen (Shanteau, 1988). Vergleicht man – im Sinne des PCS-Modells (s. 2.1.6.2) – die mentalen Netzwerke, so zeichnen sich die Netzwerke der Expert:innen dadurch aus, dass sie mit einer größeren Anzahl von Knoten aufgebaut sind als die Netzwerke der Noviz:innen, die wiederum größer sind als die der Naiven (Herbig & Glöckner, 2009). Zudem besitzen Noviz:innen zwar theoretisches Wissen, es fehlt ihnen aber (ebenso wie den Naiven) das erfahrungsbasierte Wissen, das Expert:innen aufweisen (Herbig & Glöckner, 2009; Shanteau, 1992). Laut Shanteau (1988) unterscheiden sich Expertise-Gruppen auch dahingehend, dass Fachpersonen ausgeprägte Fähigkeiten der Wahrnehmung und der Aufmerksamkeit besitzen, zwischen relevanten und irrelevanten Informationen unterscheiden, komplexe Probleme vereinfachen und flexibel auf Situationen reagieren. Auch wenn es zunächst naheliegt davon auszugehen, dass fachkundige Menschen mehr Informationen nutzen als Menschen ohne Expertise, ist es vielmehr dieser Gebrauch

und Umgang mit (irrelevanten) Informationen, der die Expertise-Gruppen voneinander abgrenzt (Shanteau, 1992; Shanteau & Stewart, 1992). So sind Fachpersonen eher in der Lage, irrelevante Stimuli oder Informationen zu ignorieren und sich stattdessen auf (wenige) relevante Hinweise zu konzentrieren (Brams et al., 2019; Ettenson et al., 1987; Pachur & Marinello, 2013). Dennoch beeinflussen situationale Faktoren den Umgang von Fachpersonen mit Informationen. In Entscheidungssituationen ohne Zeitdruck werden Informationen von Kliniker:innen bewusster verarbeitet als in Situationen mit Zeitdruck und hoher mentaler Arbeitsbelastung (Byrne, 2013; s. 2.2.3.1). Ähnliche Einflüsse von Zeitdruck zeigen sich bereits bei der Suche nach Informationen durch Fachpersonen (Čavojová & Hanák, 2014). Expert:innen nutzen (bei zunehmender Anzahl an Informationen) eher nichtlineare und nicht-kompensatorische Entscheidungsstrategien als Nicht-Expert:innen (Einhorn, 1971; Pachur & Marinello, 2013; s. 2.2.1).

Fachpersonen sind laut Shanteau (1988) fähig, ihre ursprüngliche Entscheidung abzuändern und an dynamische Entwicklungen anzupassen, in Kooperation mit oder durch Feedback von anderen zu arbeiten und aus vergangenen Entscheidungen zu lernen. Nicht-Expert:innen würden dazu tendieren, ihre Entscheidungen zu rationalisieren und zu verteidigen, anstatt aus ihnen zu lernen. Außerdem würden sich Expert:innen dadurch auszeichnen, dass es ihnen nicht um das Finden der einen richtigen Lösung, sondern um das Vermeiden von großen Fehlern und falschen Entscheidungen geht. Selbstüberschätzung, übermäßiges Vertrauen und der Verlass auf die eigene Kompetenz können aber insbesondere in spezifischen Bereichen erfahrene Menschen zu Fehleinschätzungen verleiten (Kahneman, 2011; Shanteau, 1988; Shanteau & Stewart, 1992). Selbstüberschätzung ergibt sich aus der fehlerhaften Kalibrierung zwischen Sicherheit und Richtigkeit (Chi, 2006; Cooke, 1991; Schweizer, 2005). Ist man in einem Bereich besonders ausgebildet und betreibt man dabei hohe kognitive Leistungen, kann dies zu einer starken Überzeugung vom eigenen Urteilen und Entscheiden führen (Kahneman, 2011). Dies kann zur Folge haben, dass der Einfluss von kognitiven Verzerrungen auf die eigene Entscheidung als weniger stark angesehen oder dass Inkompetenz nicht als solche erachtet wird (Pennycook et al., 2017; Zapf et al., 2018).

Weiss und Shanteau (2012) argumentieren, dass Expertise zwar häufig anhand von Erfahrungswerten oder dem Bildungsgrad abgeleitet wird, dass sich diese Expertise

aber nicht notwendigerweise in den Arbeitsergebnissen widerspiegelt. Dies sei insbesondere in Bereichen zu bedenken, in denen die Arbeit von Expert:innen nicht (objektiv) evaluiert wird. Übermäßiges Vertrauen herrscht, obwohl Expert:innen bei der Einschätzung von Wahrscheinlichkeiten nicht zwangsläufig besser sind als Nicht-Expert:innen (Cooke, 1991): Sie haben ebenso Schwierigkeiten mit der Einschätzung der Wahrscheinlichkeiten von Ereignissen, sodass in beiden Expertise-Gruppen vergleichbare Fehler passieren (z. B. Basisratenfehler; s. 2.2.1.2). Eine Begründung für die teilweise vergleichbare Entscheidungsqualität und die mutmaßlich nicht geringere Fehleranfälligkeit ist, dass sich Fachpersonen ebenso wie Nicht-Expert:innen auf Heuristiken verlassen und daraus Verzerrungen resultieren können (Dror, 2011; Herbig & Glöckner, 2009; Shanteau & Stewart, 1992). In einer qualitativen Untersuchung befragten Mishra et al. (2015) Notfallkräfte der britischen Polizei zu deren Suche nach Informationen in unsicheren Entscheidungssituationen. Die Expert:innen gaben an, in unsicheren, dynamischen und komplexen Situationen nicht nach Informationen zu suchen, sondern sich auf ihre Erfahrungen zu verlassen. Erfahrungen gingen mit einem schnellen und automatischen Entscheidungsverhalten einher, welches mit Typ-1-Prozessen vergleichbar ist (s. 2.2.1). Zudem stand hohe Expertise im Zusammenhang mit hohem Selbstbewusstsein, das wiederum Entscheidungsverhalten nach dem Typ-1-Prozess bekräftigte. Die Suche nach Informationen diente den Expert:innen nicht nur in der Reduktion der Unsicherheit, sondern auch als nachträgliche Rechtfertigung für die getroffene Entscheidung (s. 2.2.2). Laut Dror (2011) können die Mechanismen, die Fachpersonen letztlich ausmachen (z. B. schnelle, automatische Verarbeitung von Informationen), einen Mangel an Flexibilität und Kontrolle mit sich bringen. Anandarajan et al. (2008) diskutieren die Frage, ob sich Expert:innen und Nicht-Expert:innen (Studierende) in der Ausprägung von Heuristiken und Bias unterscheiden. Sie schlussfolgern, dass beide Gruppen Heuristiken nutzen, dass sich aber Nicht-Expert:innen in komplexen Situationen eher auf Heuristiken verlassen als Fachpersonen. Auch wenn Expert:innen nicht weniger Heuristiken in ihren Entscheidungen nutzen, so kann dennoch angenommen werden, dass sie aufgrund ihrer Expertise über für die Aufgabe *sinnvolle* Heuristiken verfügen und von diesen profitieren (J. K. Phillips et al., 2004).

Neben der Anfälligkeit für kognitive Verzerrungen ist der situativ auftretende, fehlende Konsens in Urteilen und Entscheidungen ein „Vorwurf“ an Fachpersonen. Diesem Vorwurf liegt die implizite Annahme zugrunde, dass sich Expert:innen einig sein können beziehungsweise sollen und dass Uneinigkeit demzufolge einen Hinweis auf Inkompetenz oder mangelnde Expertise darstellt (Shanteau, 2000). Laut Shanteau (2000) lässt sich in vielen Fachbereichen allerdings kein goldener Standard und keine Grundwahrheit ausmachen. Dies sei insbesondere in Bereichen der Fall, in denen dynamische Situationen das ständige Anpassen von Expert:innen einfordern. Uneinigkeit könne überdies in der Entwicklung von Expertise eine wichtige Rolle spielen. Fachkundige Menschen sind gerade in solchen Domänen und Settings wichtig, in denen es keine richtigen Lösungen gibt (Shanteau, 1988). Entscheidungen in der realen Welt bringen in der Regel gewisse Eigenschaften und Rahmenbedingungen mit sich, beispielsweise Zeitdruck, unvollständige Informationslagen, mehrere Beteiligte, unklare Ziele und teils massive Konsequenzen (Herbig & Glöckner, 2009; Hutton & Klein, 1999; s. 2.2.2). Expert:innen zeichnen sich dadurch aus, dass sie in diesen dynamischen und komplexen Situationen eben nicht die eine optimale Lösung anstreben, sondern eher die, die in dieser Komplexität am passendsten ist und die zufriedenstellt (Hutton & Klein, 1999). In unklaren, unbekannten Situationen sind sie in der Lage, Handlungsoptionen schnell und sicher zu generieren, deren Einsatz auch unter Zeitdruck zu evaluieren (Hutton & Klein, 1999) und dabei kreativer zu sein als Nicht-Expert:innen (Shanteau, 1988).

2.2.3.3 Kognitive Reflexion

Dual-Prozess-Theorien stehen im Zusammenhang mit dem Einsatz von Heuristiken (s. 2.2.1.2). Als eine Erklärung dafür gilt die Theorie der Attributsubstitution nach Kahneman und Frederick (2005). Das Ersetzen (Substitution) des Zielattributs durch ein heuristisches Attribut läuft im Sinne des Typ 1 schnell und automatisch ab. Dennoch kann dieser Prozess kontrolliert und korrigiert werden. Der *Cognitive Reflection Test* (CRT) von Frederick (2005) dient als Maß zur Vorhersage der Anwendung von Heuristiken (s. 3.4.4). Der CRT gilt als starker Prädiktor für die Leistung in entscheidungsbezogenen, normativen Aufgaben, wenngleich eine Einschränkung folglich darin besteht, dass solche Aufgaben in der Regel in Laborsettings untersucht werden (Juanchich et al., 2016). Der Test beinhaltet drei Items, die intuitiv mit Hilfe des System 1 beantwortet werden und dabei eine falsche Antwort

provozieren. Erst das aufwendige Nachdenken durch Typ-2-Prozesse bringt die richtige Lösung. Dadurch lässt sich die individuelle Neigung zum Verlass auf das System 1 beziehungsweise die Fähigkeit, dem zu widerstehen, ermitteln (kognitive Reflexion). Der Test kann dabei herangezogen werden „as a simple measure of one type of cognitive ability“ (Frederick, 2005, S. 26). Auch wenn sich zwischen dem CRT und Intelligenztests Korrelationen zeigen, scheinen diese beiden Konstrukte durchaus unabhängig zu sein, sodass auch intelligente Menschen irrational handeln können (Stanovich & West, 2008, 2014; s. auch 2.2.1.1). Betrachtet man die Leistungen in Bias-Instrumenten, so ist der CRT ein stärkerer Prädiktor für diese Leistung als Intelligenzmaße (Toplak et al., 2011; West et al., 2012). Eine mögliche Begründung dafür ist, dass Intelligenztests – vereinfacht gesagt – Rationalität nicht in ihrer Erfassung abdecken (Stanovich, 2016; Stanovich et al., 2014; Stanovich & West, 2014). Aufgaben zu Heuristiken und Bias stehen aber durchaus im Zusammenhang mit rationalem Denken: Da der CRT im Gegensatz zu Intelligenzmaßen gewisse Aspekte des rationalen Denkens erfasst, trägt dieser somit substantiell zur Varianzaufklärung bei (Toplak et al., 2011).

Inhaltlich handeln die Items von einem Schläger und einem Ball, von Maschinen und Gegenständen sowie von Seerosenblättern (s. 3.4.4). Niedrige Punktwerte lassen sich mit eher intuitivem Entscheidungsverhalten interpretieren, wohingegen hohe Punktwerte eher bedachtes, reflektiertes Verhalten beschreiben. Der CRT unterscheidet gut zwischen impulsiven und reflektierten Entscheider:innen (Oechssler et al., 2009). Personen mit eher niedrig ausgeprägter kognitiver Reflexion sind anfälliger für Bias, beispielsweise für den Konjunktionsfehler (Oechssler et al., 2009), und sie überschätzen ihre Leistung im CRT mehr als weniger intuitive Personen (Coutinho et al., 2021; Pennycook et al., 2017). Wenngleich der CRT das Ergebnis kognitiver Prozesse in Form der Lösung der jeweiligen Aufgabe abbildet, so beschreibt er dabei nicht *wie* das Ergebnis erzielt wurde (Blacksmith et al., 2019). Szaszi et al. (2017) untersuchten mittels lauten Denkens die Strategien, mit denen sich die Studienteilnehmenden den Aufgaben des CRT näherten und wie sie zu einer Lösung fanden. Rund 77% der Personen hatten direkt zu Beginn die richtige Lösung identifiziert oder erste Gedanken geäußert, die zur korrekten Antwort führten. Laut Szaszi et al. (2017) misst der CRT eher eine allgemeine Präferenz für Geschwindigkeit statt Genauigkeit und nicht nur die reine Fähigkeit zur kognitiven

Reflexion. Betrachtet man nur die Schläger-Ball-Aufgabe, so ist die intuitive Antwort häufig bereits richtig, sodass ein Nachdenken über die Lösung eher dazu genutzt wird, die erste spontane Antwort zu verifizieren, anstatt sie zu korrigieren (Bago & Neys, 2019). Auch Raoelison et al. (2020) argumentieren für die „smart intuitor view“ (S. 30): Personen mit hohen Reflexionsfähigkeiten sind nicht besser darin, falsche intuitive Antworten zu korrigieren, sondern sie wissen von Anfang an intuitiv die Lösung und zeigen demgemäß bessere Leistung. Mittels Messung der Mausbewegung beim Beantworten der CRT-Items zeigten Travers et al. (2016) zudem in ihrer Studie, dass Personen, die sich letztlich für die falsche intuitive Antwort entschieden haben und die somit eine eher niedrig ausgeprägte kognitive Reflexion besaßen, von der zur Auswahl gegebenen korrekten Antwort gar nicht erst angezogen worden waren. Dies lässt die Vermutung zu, dass auch intensives Nachdenken nicht zur richtigen Lösung geführt hätte.

Ein Vorteil des CRT ist, dass die Ergebnisse auf messbarer Performanz und nicht auf einem Selbstbericht zur Reflexionsfähigkeit basieren (Toplak et al., 2011). Auch wenn die Kürze des CRT ökonomische Vorteile hat, so ist einer der Nachteile, dass diese Aufgaben mittlerweile einen großen Bekanntheitsgrad erlangt haben, was zu einer Verzerrung der Ergebnisse führen kann (Haigh, 2016; Toplak et al., 2014). Ein gutes Abschneiden (max. 3 Punkte) im CRT lässt sich somit nicht eindeutig auf hohe Reflexionsfähigkeiten zurückführen: Versuchsteilnehmende, die bereits zuvor in anderen Studien oder durch die Medien in Berührung mit einzelnen Items gekommen waren, erzielten einen höheren Wert (M = 2.36) als diejenigen, für die die Items neu waren (M = 1.48; Haigh, 2016). Allerdings argumentieren Bialek und Pennycook (2018) sowie Šrol (2018), dass vorherige Kenntnis der Items nicht die Vorhersagekraft des CRT mindert, da intuitive Personen auch bei wiederholter Präsentation der Aufgaben eben aufgrund dieser Tendenz erneut Schwierigkeiten haben (können). Im Ausmaß der kognitiven Reflexion gibt es weitere interindividuelle Unterschiede. So weist eine Meta-Analyse auf einen negativen Zusammenhang des weiblichen Geschlechts mit der Anzahl der korrekten Antworten im CRT hin (Brañas-Garza et al., 2015). In einer Studie von (Frederick, 2005) unterschieden sich die beiden Gruppen dahingehend, dass Frauen tendenziell niedrigere Werte erzielten, da sie eher intuitiv die spontane, aber falsche Lösung nannten. Gaben Männer die falsche Antwort, so war diese eher unerwartet, denn sie entsprach

inhaltlich nicht den antizipierten Lösungsmöglichkeiten. Im Zusammenhang mit den Geschlechtern wird diskutiert, inwieweit die Tatsache, dass Männer im CRT besser abschneiden als Frauen, mit mathematischen Fähigkeiten begründet werden kann (Brañas-Garza et al., 2015; Juanchich et al., 2020). Sogar die Angst vor Mathematik kann die Leistung von Frauen im CRT negativ beeinflussen (Morsanyi et al., 2014). Nichtsdestotrotz ist der CRT kein Instrument zur Erfassung rein numerischer oder mathematischer Fähigkeiten (Campitelli & Gerrans, 2014; Liberali et al., 2012). Ausgehend von dieser Debatte gibt es überarbeitete CRT-Versionen, die andere oder zusätzliche Items mit geringerem mathematischem Fokus beinhalten (Sirota et al., 2021; Thomson & Oppenheimer, 2016; Toplak et al., 2014).

Die Leistung im klassischen CRT – also die individuelle Ausprägung der kognitiven Reflexion – korreliert durchaus mit lebensechten Entscheidungen, wenngleich sie nur einen sehr kleinen Anteil der Varianz ausmacht (2%; Juanchich et al., 2016). Juanchich et al. (2016) betonen, dass der CRT in normativen Entscheidungsaufgaben zwar einen prädiktiven Wert besitzt, dass ein Transfer auf lebensechte Entscheidungen aber mit Bedacht geschehen soll. Laut Frederick (2005) treffen Menschen mit unterschiedlichen Reflexionsfähigkeiten dementsprechend unterschiedliche Entscheidungen. Dabei unterstreicht er, dass eine hohe Fähigkeit nicht zwingend zu guten oder richtigen Entscheidungen führt, sondern dass der positive Einfluss dieser Fähigkeit vielmehr von der Art der Entscheidung abhängt.

2.2.3.4 Need for Cognition

Need for Cognition (NFC; auch: Kognitionsbedürfnis) beschreibt die individuelle Neigung, gerne zu denken und sich mit kognitiv anstrengenden Aktivitäten auseinander zu setzen (Cacioppo & Petty, 1982; Cacioppo et al., 1996). Diese Neigung ist relativ stabil und weniger als Fähigkeit, sondern eher als (intrinsische) Motivation oder Persönlichkeitseigenschaft zu verstehen (Cacioppo & Petty, 1982; Petty et al., 2009; s. auch Appelt et al., 2011). Das Kognitionsbedürfnis lässt sich von Intelligenz abgrenzen, da zwar ein positiver Zusammenhang, aber keine Redundanz besteht (Cacioppo & Petty, 1982; Cacioppo et al., 1996; Fleischhauer et al., 2010; Hill et al., 2013). Die *Need-for-Cognition*-Skala erfasst die Tendenz, wie gerne sich Menschen mit kognitiv anstrengenden Aufgaben oder Aktivitäten auseinandersetzen (Cacioppo & Petty, 1982; s. 3.4.3). Die Anzahl der Items, deren Aussagen zugestimmt oder widersprochen werden, variierte im Laufe der Zeit (Beißert et al.,

2014; Bless et al., 1994; Pechtl, 2009). Hohe Werte drücken in der Regel eine hohe Denkmotivation aus. Ein Nachteil einer derartigen Erfassung ist allerdings, dass diese auf Selbstberichten basiert (Toplak et al., 2011).

Laut Cacioppo et al. (1996) sind Menschen mit niedriger NFC-Ausprägung „cognitive misers“ (S. 197), wohingegen diejenigen mit hoher Ausprägung als „cognizers“ (S. 197) bezeichnet werden können. Dementsprechend können die durch die Skala erfassten Ausmaße von NFC als ein Kontinuum verstanden werden. Bleibt man beim Vergleich von Menschen mit hoher versus niedriger NFC-Ausprägung, zeigen sich gewisse Unterschiede in verschiedenen Bereichen, sei es auf kognitiver oder auf behavioraler Ebene (für eine ausführliche Übersicht s. Cacioppo et al., 1996; Petty et al., 2009). Hohe Ausprägungen im Kognitionsbedürfnis stehen beispielsweise im Zusammenhang mit erhöhtem kognitivem Aufwand beziehungsweise erhöhtem Einsatz in der Verarbeitung von Informationen, mit der besseren Leistung in arithmetischen Problemen, Anagrammen oder Glücksspielaufgaben sowie mit größerer Neugier (Cacioppo et al., 1996; Harman, 2011; Verplanken, 1993). Während Personen mit einer starken Neigung zum Denken Informationen über die Umwelt aktiv suchen, darüber nachdenken und reflektieren, verlassen sich diejenigen mit niedrigen Ausprägungen eher auf Mitmenschen, um solche Stimuli und Informationen einordnen zu können (Cacioppo & Petty, 1982; Cacioppo et al., 1996). So kommen Menschen auf verschiedenen Wegen zu Positionen und Meinungen: Personen mit hoher Neigung zum Denken orientieren sich eher an relevanten Aspekten, die sich auf den Inhalt beziehen, wohingegen diejenigen mit niedriger Neigung sich eher nach anderen richten (z. B. Expert:innen oder Gruppenkonsens; Cacioppo et al., 1996). Erstellt man den Bezug zu Dual-Prozess-Theorien (s. 2.2.1), so liegt die Annahme nahe, dass Menschen mit schwach ausgeprägtem NFC zur Nutzung einfacher, heuristischer Stimuli tendieren, wohingegen sich diejenigen mit starker Ausprägung auch mit weiteren Informationen auseinandersetzen (Petty et al., 2009; Verplanken, 1993). Mit Blick auf den kognitiven Aufwand, der zum Bearbeiten einer Aufgabe erforderlich ist (Typ-1/Typ-2-Prozesse), muss laut Cacioppo et al. (1996) die kognitive Repräsentation einer Aufgabe oder eines Stimulus bedacht werden: Handelt es sich um ein komplexes, aber bekanntes Thema, so benötigen Personen mit hohen Ausprägungen weniger kognitiven Aufwand, um sich mit dem Thema zu beschäftigen. Bei unbekannten Themen, zu denen Informationen

oder Meinungen erst generiert werden müssen, sei der eingesetzte Aufwand aufgrund der Neigung zum Denken wiederum größer als bei Personen mit niedrigen Ausprägungen. Allerdings hat eine hohe Ausprägung nur insoweit einen Vorteil für die Bearbeitung einer Aufgabe, als dass diese Aufgabe an sich ausreichend komplex sein muss, denn weiterführendes Denken hat auf simple Entscheidungen kaum Auswirkung: „This type of *overthinking* … normally leads to suboptimal decision choices“ (McElroy et al., 2020, S. 540). Eine hohe Motivation zum Denken führt demnach nicht zwangsläufig zu besseren Entscheidungen (Carnevale et al., 2011; McElroy et al., 2020).

Curşeu (2006) argumentiert, dass denkfreudige Menschen, im Vergleich zu denjenigen mit schwach ausgeprägter Motivation, eher als rational beschrieben werden können. Aber auch wenn die NFC-Skala diese Tendenz zum Denken erfasst, so ist dies nicht mit dem Messen von Rationalität gleichzusetzen (Petty et al., 2009). Eine hohe Motivation führt zwar eher dazu, Denkfehler oder erste Eindrücke angesichts neuer Informationen korrigieren zu wollen, allerdings stellt sie nicht notwendigerweise einen Schutz vor dem Einfluss von Bias dar (Cacioppo et al., 1996; Petty et al., 2009). Personen mit starkem Kognitionsbedürfnis sind beispielsweise eher durch Priming zu beeinflussen als diejenigen mit niedrigen Ausprägungen, weil präsentierte Informationen bei denkfreudigeren Menschen mehr – aber möglicherweise auch falsche – Gedächtnisinhalte aktivieren (Cacioppo et al., 1996). Doch Carnevale et al. (2011) untersuchten eine Stichprobe von Führungspersonen und deren Beeinflussbarkeit durch Bias und stellten im Gegensatz dazu fest, dass eine hohe NFC-Ausprägung mit reduzierten Rahmungseffekten einherging. Einen ähnlichen Befund berichteten bereits Smith und Levin (1996). Menschen mit niedriger Ausprägung sind eher für solche Denkfehler anfällig, die auf heuristischen Prozessen beruhen, wohingegen Menschen mit hoher Ausprägung eher zu auf kognitiver Anstrengung basierenden Bias neigen (Petty et al., 2009). Petty et al. (2009) drücken es folgendermaßen aus: „Although the mechanisms usually differ, individuals high and low in N[F]C can both be susceptible to various biases, regardless of the nature and the source of the biasing factor (e.g., an anchor, a stereotype, or an emotional state)” (S. 325). Laut Frederick (2005) gibt es zwar empirische und konzeptuelle Überschneidungen zwischen kognitiver Reflexion (gemessen mit dem

CRT; s. 2.2.3.3) und NFC, dennoch deutet die schwache Korrelation zwischen den beiden Tests (r = .22) auf eine gewisse Unabhängigkeit hin.

Außerdem spielen situationale Faktoren eine Rolle für die Wirksamkeit von Denkmotivation, denn auch Personen mit hohen Ausprägungen können nicht alle auf sie einwirkenden Informationen auf hohem kognitivem Level bearbeiten (Cacioppo et al., 1996; Ruscio, 2000). Auch wenn bei Personen eine stabile, schwach ausgeprägte Tendenz zum Denken vorliegt, können diese durch mit der Entscheidungssituation verbundenen Anreizen entgegen ihrer Neigung handeln: Die NFC-Effekte treten eher bei Aufgaben auf, bei denen die persönliche Relevanz als niedrig bis mittel eingestuft wird. Liegt eine hohe persönliche Relevanz vor, so sind auch Personen mit niedrigen Ausprägungen zum Nachdenken motiviert (Cacioppo et al., 1996; Petty et al., 2009). Bouckenooghe et al. (2007) untersuchten, welche Strategien beim Erleben innerer Konflikte in wichtigen Entscheidungssituationen angewendet werden. Sie stellten einerseits einen positiven Zusammenhang zwischen NFC und wachsamem Entscheidungsverhalten (r = .17) fest. Andererseits fanden sie negative Zusammenhänge zwischen NFC und der Tendenz, sich panisch um Lösungen zu bemühen (r = -.25), sowie der Neigung, die Verantwortung an andere Personen abzugeben (r = -.31). Zeitdruck als situationaler Faktor (s. 2.2.3.1) beeinflusst das Entscheidungsverhalten derjenigen mit schwachen Ausprägungen von NFC, da unter Zeitdruck vermehrt auf heuristische Strategien zurückgegriffen wird (Verplanken, 1993). Verplanken (1993) stellt die Vermutung auf, dass sich Menschen mit niedrigem Kognitionsbedürfnis aus diesem Grund besser an stressige Umgebungen anpassen können. Weniger denkfreudige Menschen reagieren zudem eher mit einem Strategiewechsel auf zeitliche Begrenzungen als diejenigen mit hoher Ausprägung (Ordóñez & Benson, 1997). Somit gibt es situationale Faktoren, die – je nach Neigung zu NFC – mit bestimmten Reaktionen einhergehen. Des Weiteren werden Menschen je nach Ausprägung unterschiedlich wahrgenommen. In Diskussionsrunden zwischen zwei Personen galten diejenigen mit hoher Denkmotivation als überzeugender, da sie mehr ihre Meinung unterstützende und zugleich valide Argumente lieferten (Shestowsky et al., 1998). Shestowsky et al. (1998) leiten daraus ab, dass Menschen mit diesen gezeigten Verhaltensweisen (die demgemäß vermutlich einen hohen NFC-Wert besitzen) in einer Diskussion eher als Expert:innen wahrgenommen werden.

2.2.4 Zwischenfazit: Ableitung von für die Studie relevanten psychologischen Inhalten

Im Anschluss an die Darstellung juristischer Inhalte wurde bereits ein sich auf ebenjene Inhalte fokussierendes Zwischenfazit gezogen (s. 2.1.7). Ein solches Fazit erfolgt nun für die in den Abschnitten 2.2.1–2.2.3.4 erörterten psychologischen Themenbereiche.[17] Warum wurde auf Entscheidungsforschung eingegangen? In der vorliegenden Studie haben Teilnehmende insbesondere die Aufgabe, sich auf Grundlage eines schriftlich beschriebenen Sachverhaltes ein Urteil darüber zu bilden, ob ein Tatverdacht vorliegt oder nicht, und dementsprechend eine Entscheidung zu treffen („Ja", „Nein"; s. 3.4.2). Da die Entscheidungsoptionen bereits feststehen und sich der Fokus der Studie auf die konkrete Auswahl einer dieser Optionen bezieht, beleuchtet diese Untersuchung die selektionale Phase des Entscheidungsprozesses – wenngleich sich einzelne Items bis zu einem gewissen Grad auch auf post-selektionale Prozesse beziehen (s. 3.4.2). Abschnittübergreifendes Thema war somit die Betrachtung normativer und insbesondere die Ableitung deskriptiver Modelle menschlicher Entscheidungsfindung (s. 2.2.1). Da davon auszugehen ist, dass Menschen nicht (immer) die gleiche Entscheidung wie ihr Gegenüber treffen, müssen mögliche Einflussfaktoren auf den sowie individuelle Unterschiede im Entscheidungsprozess betrachtet werden (s. 2.2.3.1–2.2.3.4).

Ausgehend vom Konzept der begrenzten Rationalität ist eine allumfassende, ideale Integration von für eine Entscheidung notwendigen Informationen nicht möglich. Insbesondere aus dem Grund des (kognitiven) Ressourcenschonens weichen Menschen vom normativen, idealen Entscheidungsverhalten ab. Da Entscheidungssituationen im dafür erforderlichen kognitiven Aufwand variieren können, wurden besonders die deskriptiven Dual-Prozess-Annahmen betrachtet, die verschiedene Prozesstypen differenzieren. So ist zur Entscheidungsfindung einerseits eine schnelle, automatische und heuristische Vorgehensweise (Typ 1) und andererseits ein kontrolliertes, durchdachtes und anstrengendes Verfahren (Typ 2) möglich (s. 2.2.1). Unter dem Blickwinkel dieser Annahmen können intuitive Entscheidungen als Ergebnisse der Typ-1-Prozesse verstanden werden, wohingegen rationale Entscheidungen mit der Natur der Typ-2-Prozesse vergleichbar sind (s. 2.2.1.1). Es gilt fest-

[17] Siehe Fußnote 8.

zuhalten, dass es neben den Dual-Prozess-Theorien auch weitere Entscheidungsmodelle gibt. In Abschnitt 2.1.6.2 wurde bereits beispielhaft auf das PCS-Modell eingegangen. Dennoch wurden duale Prozesse zur Untersuchung der hier angedachten Fragestellungen (s. 2.5) als zugrundeliegendes Konzept gewählt, wenngleich an geeigneten Stellen auch Verweise auf andere Modelle erfolgen. Mit Blick auf die für die Studie ausgewählten Delikte (Diebstahl, Körperverletzung) bietet das duale Konzept eine ausreichende Grundlage: „Especially with regard to professionals' decisions on sentencing in low-level crimes, simple non-compensatory heuristic models seem to provide a more appropriate description of the decision-making process" (Hupfeld-Heinemann & Helversen, 2009, S. 286).

An die Dual-Prozess-Annahmen anknüpfend wurde dargestellt, wie es in der Entscheidungsfindung zu unterschiedlichen Ergebnissen, zu Denkfehlern und Bias kommen kann (s. 2.2.1.2). So können verkürzte Strategien beziehungsweise Heuristiken zwar die kognitiven Ressourcen schonen, aber auch zu systematischen Urteilsverzerrungen führen. Sowohl der Heuristiken-und-Bias- als auch der *Fast-and-Frugal-Heuristics*-Ansatz beschreiben solche verkürzten Entscheidungsstrategien, wobei sich die beiden Ansätze darin unterscheiden, ob solche Strategien eher als nützlich oder als schädlich anzusehen sind. Allerdings ist es Menschen nicht unmöglich, Denkfehler zu erkennen und zu korrigieren. Nach einer intuitiven Antwort entsteht ein Gefühl von Richtigkeit, das insbesondere bei schwacher Ausprägung dazu führen kann, dass eine getroffene Entscheidung oder eine gegebene Antwort hinterfragt und bei Bedarf korrigiert wird. Eine Entscheidungssituation ist nicht losgelöst von der Phase, die sich an eine solche Entscheidung anschließt (s. 2.2.2). Um aus einer Erfahrung lernen zu können, gilt es, mögliche Konsequenzen und erhaltenes Feedback zu betrachten. Situationen und Lernumgebungen unterscheiden sich in der Qualität und der Quantität hilfreicher post-selektionaler Informationen. Auch der Faktor der Verantwortung bestimmt einen Entscheidungsprozess mit, da (je nach Modalität der Verantwortung) die Wahrscheinlichkeit für Bias steigen oder sinken kann.

Somit ist eine Beeinflussbarkeit der Entscheidungsprozesse anzunehmen – auch wenn dies nicht zwingend mit schlechten Entscheidungsergebnissen einhergeht. Towfigh und Glöckner (2015) beschreiben es folgendermaßen: „Häufig müssen Entscheidungen unter Zeitdruck, auf Basis unvollständiger Informationen und unter

einem hohen Grad an Unsicherheit bezüglich der Verlässlichkeit der vorliegenden Informationen getroffen werden“ (S. 270). Daher sind Einflussfaktoren eine Betrachtung wert, die in dieser Studie als unabhängige Variablen oder Prädiktoren berücksichtigt werden (s. 3.1). Zu den personenbezogenen Faktoren zählen Expertise (s. 2.2.3.2), kognitive Reflexion (s. 2.2.3.3) und Need for Cognition (s. 2.2.3.4). Da es sich bei einer der Teilstichproben um Rechtsexpert:innen handelt, die mit weiteren Teilstichproben verglichen werden, muss der Grad der Expertise genauer differenziert werden. Gemäß der Unterscheidung nach Shanteau (1988) stellt die vorliegende Studie einen Vergleich zwischen Naiven, juristischen Noviz:innen (Studierende oder Referendar:innen) und Expert:innen an (s. 3.1). Somit liegt der Fokus auf domänenspezifischer Expertise. Fachkundige und fachfremde Personen unterscheiden sich in ihrem Umgang mit relevanten und irrelevanten Informationen, in ihrem Gebrauch nicht-kompensatorischer und heuristischer Entscheidungsstrategien und in ihrer Fähigkeit, sich an dynamische Entscheidungssituationen anzupassen. Allerdings besteht bei Expert:innen die Gefahr der Denkfehler, der Selbstüberschätzung und des übermäßigen Vertrauens. Fehlender Konsens zwischen Fachpersonen wird zwar tendenziell als Problem wahrgenommen, allerdings ist Expertise insbesondere in den Bereichen von Bedeutung, in denen es keine eindeutige Lösung und keinen goldenen Standard gibt. Dies ist auch für das hier untersuchte juristische Setting zutreffend. Die Fähigkeit zur kognitiven Reflexion ist kein Maß der Intelligenz, sondern beschreibt die Anwendung von Heuristiken. Menschen mit schwach ausgeprägter Fähigkeit gelten als intuitiv (Typ 1), wohingegen eine hohe Ausprägung mit bedachtem (Entscheidungs-)Verhalten (Typ 2) verknüpft wird. Need for Cognition bezieht sich auf die individuelle Motivation zu Denken. Denkfreudige Menschen besitzen eine hohe Ausprägung von NFC – was sich beispielsweise in einem vermehrten kognitiven Aufwand äußert – und gelten als rationaler. Dennoch ist das Kognitionsbedürfnis kein Schutz vor Denkfehlern, zumal bei hoher persönlicher Relevanz auch weniger denkfreudige Menschen zum Nachdenken motiviert werden können. Neben diesen personenbezogenen Faktoren steht Zeitdruck als sehr alltagsnaher situationaler Faktor (s. 2.2.3.1). Dieser wird häufig mit einer Form von Stress gleichgesetzt. Zeitdruck bringt Menschen dazu, sich in begrenzter Zeit um eine möglichst gute Entscheidung zu bemühen. Um dies zu erreichen, müssen entweder Abstriche in der Genauigkeit oder in der Geschwindigkeit gemacht werden (*speed-accuracy-trade-off*). Dabei kommen verschiedene,

eher heuristisch geprägte Strategien (Typ 1) zum Einsatz, denn Informationen können selektiver und schneller betrachtet werden. Menschen sind unterschiedlich gut darin, mit Zeitstress umzugehen. Dafür wichtige Faktoren sind die Expertise, die subjektive Wahrnehmung von Zeit oder auch ein individuelles Gefühl von Dringlichkeit. Im Abschnitt 2.3 werden die dargestellten Erkenntnisse und Befunde der juristischen (s. 2.1) und psychologischen (s. 2.2) Themenbereiche zusammengeführt.

2.3 Die Relevanz von Entscheidungsforschung im Strafprozess: Übertragung empirischer Forschung auf den juristischen Fachbereich

Nach der voneinander nahezu losgelösten Darstellung juristischer sowie psychologischer Inhalte (s. 2.1; 2.2) erfolgt deren Zusammenführung. Zunächst gilt es allgemein zu betrachten, welche Chancen und Risiken oder Grenzen eine solche Verknüpfung mit sich bringt (s. 2.3.1). Anschließend wird in Abschnitt 2.3.2 abgeleitet, welche Erkenntnisse der Entscheidungsforschung sich auf das Strafverfahren übertragen lassen und welche Hinweise für das Vorhandensein von Einflussfaktoren auf den Entscheidungsprozess festzustellen sind. Diese (extra-)legalen Einflussfaktoren ergeben sich aus Besonderheiten des Strafverfahrens. Dabei wird im Rahmen dieser Arbeit zwischen Besonderheiten unterschieden, die entweder vorrangig im Ablauf und den Rahmenbedingungen des Strafverfahrens begründet sind (s. 2.3.3) oder die einen Bezug zu den Verfahrensbeteiligten aufweisen (s. 2.3.4). Aufgrund ihrer Relevanz im Ermittlungsverfahren werden ausgewählte Forschungsbefunde für die staatsanwaltschaftliche Perspektive in Abschnitt 2.3.5 fokussiert.

2.3.1 Chancen und Risiken der Übertragung empirischer Methoden und Erkenntnisse auf den juristischen Fachbereich

Der Strafprozess wird durch Menschen operationalisiert. Folglich sind es deren Wahrnehmungen, Erinnerungen und Entscheidungen, die diesen Prozess steuern (D. Simon, 2011). Forschungsbefunde, die sich mit ebenjenen Konzepten befassen, tragen daher maßgeblich zum Verständnis der strafprozessualen Abläufe bei: „As the process can perform no better than the mental performance of the people involved, it seems sensible to examine its workings from a psychological perspective“ (D. Simon, 2011, S. 145). Die Gegebenheit, dass juristische Entscheidungen nicht

vollends durch Gesetze, Regeln und Richtlinien begrenzt werden, sondern Handlungs- und Verhaltensspielraum besteht, führt zu dem Schluss, dass es dieses Handeln und Verhalten zu untersuchen gilt, um Aussagen über juristische Entscheidungen treffen zu können. Es werden zwar beispielsweise die Auswahl der Entscheidungsoptionen oder der Zeitpunkt für gewisse Entscheidungen festgelegt, nicht aber welche Beweise wie bewertet und interpretiert werden müssen (Ebbesen & Konečni, 1982; s. 2.1.6). Ohne diesen Spielraum wäre eine reine Betrachtung der Gesetze und Regeln ausreichend, um das Justizwesen zu verstehen (Ebbesen & Konečni, 1982). Da psychologische Grundlagen durchaus Gegenstand der Ausbildung sind, lässt sich bereits auf universitärer Ebene zumindest eine gewisse Bedeutsamkeit der Psychologie – im Vergleich zu anderen Fächern – für die juristische Kompetenzentwicklung erkennen (s. Juristenausbildungsgesetz NRW).

Die Tatsache, dass – entgegen der erkennbaren Sinnhaftigkeit des Transfers – eine solche Übertragung wissenschaftlicher Erkenntnisse auf die juristische Praxis wenig stattfindet, war ein Anlass für eine Studie von Redding und Reppucci (1999). Sie untersuchten, wie Richter:innen die Relevanz und Zulässigkeit sozialwissenschaftlicher Befunde einschätzten, die entweder ihrer soziopolitischen Meinung zur Todesstrafe entsprachen oder nicht. Es zeigte sich, dass Richter:innen (ebenso wie die Kontrollgruppe) ihre Meinung unterstützende Befunde höher einstuften als widersprechende Informationen. Diejenigen mit einer eher liberalen Interpretation des Gesetzes werteten den Beitrag sozialwissenschaftlicher Befunde als relevanter für den juristischen Kontext ein. Redding und Reppucci (1999) schlussfolgern, dass eine Nichtübereinstimmung zwischen sozialwissenschaftlichen Befunden und individuellen soziopolitischen Ansichten eine mögliche Begründung für den wenig stattfindenden Transfer darstellt. Auch Sagana und van Toor (2020) setzen sich mit dem Paradox des begrenzten, wenngleich notwendigen Transfers wissenschaftlicher Arbeiten auf die Praxis auseinander. Auf Grundlage empirischer Studien leiten sie ab, dass eine gewisse Skepsis bei den Rechtsexpert:innen herrscht, da diese davon ausgehen, dass solche Untersuchungen aufgrund des fehlenden beziehungsweise begrenzten juristischen Fachwissens auf Seiten der Wissenschaftler:innen nicht die Praxis abbilden können und somit nicht übertragbar sind (s. auch Engel & Gigerenzer, 2006). Studien würden sich beim Zusammenspiel von menschlichen und prozeduralen Faktoren eher auf ersteres beziehen und dabei den Prozessrahmen

nicht ausreichend berücksichtigen. Sagana und van Toor (2020) argumentieren aber auch, dass Studien diese Komplexität der Prozesse nicht immer abbilden können, da es methodische Grenzen gibt: „Legal psychology can be categorized as psychology *and* law or psychology *in* law, but not psychology *of* law“ (S. 227). Dazu passt die Annahme von Engel (2017): „[Methodological] standards do not always match the legal research question” (S. 1). Auch die Tatsache, dass diverse Instanzen und Personen zum Verlauf eines Strafverfahrens beitragen, löst gewisse Dynamiken aus, deren Messung einer Mammutaufgabe gleicht (B. D. Johnson & Stewart, 2016). Schmittat (2017) argumentiert, dass sich Entscheidungsfindung nicht nur auf die Beweiswürdigung reduzieren lässt, sondern dass diese auch andere Entscheidungsmomente innerhalb des Strafverfahrens umfasst. Eine Betrachtung verschiedener Fachbereiche (z. B. Neurowissenschaften, Ökonomie, Psychologie) kann sich eignen, um rechtspolitische Empfehlungen abzuleiten. Dabei muss allerdings berücksichtigt werden, dass sich unterschiedliche Fachbereiche mit verschiedenen Abstraktionsebenen befassen, sodass eine Generalisierung von Befunden einzelner Bereiche zu Fehlschlüssen führen kann (Glöckner, 2008a).

Aufgrund der Relevanz für die vorliegende Studie gilt es, die Vignettenmethode genauer zu betrachten. Vignetten sind Kurzgeschichten, in denen hypothetische Charaktere in Situationen beschrieben werden, zu denen sich Proband:innen je nach Forschungsfrage äußern sollen (Alexander & Becker, 1978; Eifler & Bentrup, 2003; J. Finch, 1987; Schnurr, 2003). Vignetten eignen sich einerseits für professionelle Stichproben, um zu untersuchen, welchen Einfluss diese Professionalität auf das Wahrnehmen, Beurteilen und Entscheiden hat. Andererseits eignen sie sich für die Allgemeinbevölkerung und heterogene Stichproben, in denen Teilnehmende mit sich unterscheidenden Merkmalsausprägungen die gleichen Vignetten erhalten und hinsichtlich ihrer Antworten miteinander verglichen werden (Sauer et al., 2011; Schnurr, 2003). Allgemein eignet sich diese Methodik für die Messung von Einstellungen und Normen, aber auch von Handlungsabsichten und Entscheidungen (Auspurg, Hinz, Liebig & Sauer, 2009; Groß & Börensen, 2009; Schnurr, 2003). Genauer gesagt können Vignetten auch für vielseitige kriminalistische oder juristische Fragestellungen eingesetzt werden, sei es inhaltlich zu Taten wie Vergewaltigung oder Körperverletzung (Alexander & Becker, 1978), zu einstellungsbezogenen Fragestellungen (Angemessenheit von Strafen für Verbrechen: Durham, 1986;

Rossi et al., 1985; Strafeinstellungen: Suhling et al., 2005) oder zu konkreten juristischen Praxisthemen, wie dem Verstoß gegen Bewährungsauflagen (Maguire et al., 2015). Die Vignettenmethode kann auf fachliche Stichproben übertragen werden. Beispiele sind Untersuchungen von Polizist:innen hinsichtlich des Entscheidungsverhaltens bei Festnahmen im Zusammenhang mit häuslicher Gewalt (S. W. Phillips & Sobol, 2010) oder hinsichtlich des Verhaltens bei Verkehrskontrollen (S. W. Phillips, 2009) sowie Studien mit Richter:innen, die aufgefordert waren, sich zu hypothetischen Fällen häuslicher Gewalt zu äußern, in denen die Angaben zum Alkohol- und Drogenkonsum des Beschuldigten variierten (Macdonald et al., 1999). Vignetten können die Komplexität von realen (Entscheidungs-)Situationen aber nur begrenzt abdecken, weil sie in der Regel nur für die Fragestellung relevante Aspekte beinhalten (Bieneck, 2009). Da sie aber mehrere Informationen gleichzeitig enthalten, ermöglichen Vignetten es dennoch, Entscheidungssituationen real zu gestalten (Alexander & Becker, 1978; Auspurg, Hinz, Liebig & Sauer, 2009; J. Finch, 1987). Das Abwägen der Vielzahl an Informationen, welches die Proband:innen leisten müssen, um zu einer Antwort zu kommen, kann Hinweise darauf geben, welche Merkmale relevanter sind als andere, insbesondere wenn diese Merkmale oder Dimensionen systematisch variieren (Auspurg, Hinz, Liebig & Sauer, 2009; Oswald et al., 2009). Zudem wird argumentiert, dass bei der Erfassung von Einstellungen die Verzerrung von Antworten aufgrund von sozialer Erwünschtheit weniger wahrscheinlich ist als bei Methoden, die auf Items basieren (Auspurg, Hinz & Liebig, 2009). Außerdem stellen Vignetten bei komplexen Themen, zu denen die Einstellungen der Teilnehmenden erfragt werden, eine Alternative zu itembasierten Verfahren dar, da sie über allgemeine und möglicherweise oberflächliche Fragen hinausgehen können (Suhling et al., 2005).

> Vignettes are therefore suitable tools for studying decision making comparatively as they provide a common scenario as a starting point to explore decision making in different contexts. But they also allow us to gain insight into the similarities and differences between decision-makers as well as those that are associated with differences in context, culture and the specific idiosyncrasies of different systems. (Maguire et al., 2015, S. 246)

Es muss berücksichtigt werden, dass Vignetten zwar die Absichten für ein Verhalten erfassen, dass dies aber nicht mit der tatsächlichen Umsetzung einer Absicht zusammenhängen muss, insbesondere wenn es sich um prospektives Verhalten handelt (Bieneck, 2009; J. Finch, 1987; Groß & Börensen, 2009). Dies liegt an der

hypothetischen Natur der Fallbeschreibungen, zumal Proband:innen für ihre Entscheidung mit keinerlei Konsequenzen konfrontiert werden und ihre Entscheidung keinen wirklichen Einfluss auf andere Personen hat (Bieneck, 2009). Auch wenn eine Intention als wichtiger Prädiktor für tatsächliches Verhalten gilt, kann dieses nicht mit absoluter Gewissheit vorausgesagt werden. Eifler und Bentrup (2003) zeigten in ihrer Studie, dass Selbstberichte zu intendiertem Verhalten und die tatsächlich beobachteten Häufigkeiten abwichen: Selbstberichte führten zum Überschätzen von hilfreichem Verhalten, wie dem Aufhalten einer Tür für einen Mitmenschen (im Vergleich zu abweichendem Verhalten, wie dem Überqueren einer Straße bei roter Ampel). Ausschnitte und Ergebnisse dieser Studie wurden von Groß und Börensen (2009) repliziert. Sowohl die (mögliche) Diskrepanz zwischen Absicht und Verhalten sowie die nur begrenzt umsetzbare Nähe zur Realität sprechen für ein Fehlen ökologischer Validität in der Vignettenmethode (Maguire et al., 2015).

Die externe Validität psychologischer Studien im juristischen Feld hat ihre Grenzen (Engel, 2017; Konečni & Ebbesen, 1979, 1982; Schweizer, 2005; s. 5.3). Die in Studien präsentierten Fallbeispiele werden inhaltlich so reduziert, dass sie nicht mehr dem realistischen Pensum an Informationen entsprechen. Hypothetische Fallbeispiele bringen im Gegensatz zur Arbeit in der Praxis keine Konsequenzen mit sich (Bieneck, 2009). Des Weiteren wird in Experimenten zur Erhebung von Variablen häufig eine numerische Skalierung genutzt, die in der Praxis nicht zum Einsatz kommt: Oft wird Schuld in Studien numerisch erhoben, wenngleich sie praktisch als dichotome Ja-/Nein-Frage vorkommt (Konečni & Ebbesen, 1979, 1982; Schweizer, 2005). Simulationsstudien haben den Nachteil, dass die Eindrucksbildung und die Entscheidungsfindung mithilfe von ausgewählten, reduzierten Informationen sehr schnell stattfinden, wohingegen Informationen in einem tatsächlichen Strafverfahren von verschiedenen Personen über einen längeren Zeitraum hinweg vermittelt werden. Sowohl die Eindrucksbildung als auch die Entscheidungsfindung entwickeln sich demnach eher graduell (Konečni & Ebbesen, 1979, 1982). Auch der häufige Einsatz leicht verfügbarer Studierendenstichproben wird kritisch angesehen (Konečni & Ebbesen, 1982). Konečni und Ebbesen (1979) argumentieren, dass unterschiedliche Schlussfolgerungen zu vergleichbaren Fragestellungen

(Kautions- oder Urteilsentscheidungen) gezogen werden können, je nachdem, welche Stichproben (Studierende oder Rechtsexpert:innen) und Methoden (simuliert oder naturalistisch) eingesetzt werden. Dies erschwert die Ableitung eindeutiger praktischer Implikationen. Trotz mangelnder externer Validität haben Simulationsstudien gewisse Vorteile: Sie reduzieren Störvariablen, ermöglichen experimentelle Kontrolle, erhöhen die interne Validität und lassen Replizierbarkeit zu (Schweizer, 2005). Somit gilt eine Übertragung empirischer Befunde und Methoden (hier: Vignetten) auf die juristische Praxis trotz gewisser Einschränkungen und Grenzen als gerechtfertigt. Ausgehend von dieser generellen Anwendbarkeit der Methode stellt sich die Frage, welche Erkenntnisse die Übertragung der Entscheidungsforschung auf den Strafprozess überhaupt bietet?

2.3.2 Entscheidungen im Strafprozess aus psychologischer Sicht: Hinweise auf Einflussfaktoren

Empirische Befunde der Entscheidungsforschung lassen sich auf rechtsprechende Personen übertragen (Guthrie et al., 2001). In juristischen Kontexten gibt es in der Regel keine optimalen, normativen Entscheidungen, da diese zu aufwendig und zu ressourcenintensiv wären (Hupfeld-Heinemann & Helversen, 2009; Towfigh & Glöckner, 2015). Laut Hermann (2009) kann das „Strafverfahren … somit als Prozess der stufenweisen Reduzierung von Komplexität gesehen werden, der von normativer Rationalität geprägt wird“ (S. 660). Es liegt der Schluss nahe, dass in solchen Kontexten der Verurteilung auf deskriptive, ressourcenschonende Vorgehensweisen zurückgegriffen wird (insbesondere bei Verbrechen mit niedriger Schwere; Hupfeld-Heinemann & Helversen, 2009; s. 2.2.4). Auch die Dual-Prozess-Annahmen lassen sich auf juristische Kontexte übertragen (s. auch Schweizer, 2009). Sie sind für eine Kontrastierung von intuitivem und reflektiertem Vorgehen – dem sich in dieser Studie durch die Manipulation von Zeitdruck und die Differenzierung von Expertise-Stufen angenähert wird (s. 2.4) – angemessen (Evans, 2008).[18] Guthrie et al. (2007) ordnen richterliche Entscheidungsfindungen so ein, dass *formalistische* Modelle eine logische, überlegte Übertragung der Gesetze und Regeln auf Fälle

[18] Für die Anwendung eines anderen Entscheidungsmodells (PCS; s. 2.1.6.2) auf das juristische Setting s. D. Simon (2004), D. Simon et al. (2001), D. Simon et al. (2004b), Engel et al. (2020) sowie Towfigh und Glöckner (2015).

betonen, wohingegen *realistische* Modelle eher von einem intuitiven Prozess ausgehen, der erst im Nachhinein rational durchdacht wird. Da eine strikte Trennung der Modelltypen in der Praxis nicht haltbar ist und ein Dual-Prozess-Konzept plausibler erscheint, berücksichtigen Guthrie et al. (2007) in ihrem „intuitive-override model“ (S. 6) die Ansätze beider Modelle: Realistisch gesehen ist die juristische Intuition wichtig; formalistisch betrachtet ist überlegtes Vorgehen notwendig. Sie begründen ihre Annahmen mit dem System-1/System-2-Modell von Kahnemann und Frederick (2005, zitiert nach Guthrie et al., 2007). Da empirische Befunde darauf hindeuten, dass Richter:innen in manchen Situationen zwar zu intuitiven Antworten neigen, aber dennoch in der Lage sein können, ihre Angaben zu reflektieren, erachten Guthrie et al. (2007) die Übertragung eines solchen Dual-Prozess-Modells auf juristische Entscheidungsfindung für sinnvoll. Allerdings betonen sie, dass es nicht darum geht, richterliche Intuition vollends zu vermeiden, obwohl diese insbesondere dann gefährlich werden kann, wenn sich in ungeeigneten Situationen darauf verlassen wird (s. auch Rachlinski & Wistrich, 2017). Ähnlich argumentieren Glöckner und Ebert (2011), dass intuitive Prozesse vor allem in komplexen Fällen notwendig sind, in denen es eine große Menge an Informationen zu verarbeiten und zu integrieren gilt. In solchen Fällen nutzen Richter:innen verstärkt einfache, heuristische Strategien (s. auch Dhami, 2003; s. 2.3.4.1).

Die StPO ermöglicht sowohl intuitives als auch reflektiertes Vorgehen. So können intuitive Prozesse im Rahmen der freien Beweiswürdigung zum Tragen kommen, aber aufgrund der Notwendigkeit der Urteilsbegründung stehen diese nicht losgelöst von Überprüfbarkeit (Glöckner, 2008b; Glöckner & Ebert, 2011; s. 2.1.6). Laut Schweizer (2009) sind Entscheidungssituationen mit geringer Begründungsdichte sowie großem Ermessen förderlich für intuitive Reaktionen (Typ 1), wohingegen solche mit hoher Begründungsdichte und eingeschränktem Ermessen reflektierte Entscheidungen begünstigen (Typ 2). Forschung im Hinblick auf den strafrechtlichen Kontext befasst sich mit der Beeinflussbarkeit von Verfahrensbeteiligten oder mit Einflussfaktoren auf die Entscheidungsfindung (z. B. Daftary-Kapur et al., 2010; Niehaus et al., 2009; Oswald & Wyler, 2018). Eine übergreifende Zusammenfassung der Ergebnisse spricht dafür, dass Urteile und Entscheidungen durch diverse Faktoren beeinflusst werden können. Dass eine Einflussnahme stattgefun-

den haben muss, kann durch das Auftreten ungleicher Entscheidungen geschlussfolgert werden. Disparitäten bei juristischen Entscheidungen werden beispielsweise mit der Länge von verhängten Strafen gemessen. Dies ermöglicht Within- oder Between-Vergleiche, sei es mit Blick auf Personen, Jurisdiktionen oder Länder (Sporer & Goodman-Delahunty, 2009). Interindividuelle Unterschiede im gefällten Urteil sind demzufolge ein Hinweis, dass es gewisse Einflussfaktoren geben muss, die letztlich zu dieser Varianz (oder Diskrepanz) beitragen. Bei Kautionsentscheidungen geht es um die Frage, ob eine angeklagte Person bis zum Verhandlungstermin (unter Auflagen) in Freiheit bleibt oder in Gewahrsam kommt. Bei dieser Entscheidung gibt es verschiedene rechtliche Faktoren zu berücksichtigen, wenngleich deren vage Definition einen gewissen Handlungsspielraum ermöglicht. In einer Studie von Dhami und Ayton (2001) zeigte sich, dass sich die untersuchten Gerichte trotz vergleichbarer Fallzahlen in den finalen Kautionsentscheidungen unterschieden und teilweise bei den gleichen Fällen ungleiche Urteile getroffen wurden. Auch Schroeder und Verrel (2017) beschreiben von der Region abhängige Verfahrenstendenzen. Dadurch werden interindividuell unterschiedliche Vorgehensweisen deutlich.

Welche Einflussfaktoren sind nun relevant? Visher (1987) benennt drei Kategorien: Charakteristiken der Richter:innen, Charakteristiken der Opfer und angeklagten Personen sowie die Beweislage und Hintergründe zum Fall. Innerhalb dieser Kategorien lassen sich wiederum Subthemen ausmachen. Mit Blick auf die Charakteristiken der Opfer und der angeklagten Personen sind beispielsweise deren Geschlecht, Herkunft oder Attraktivität zu nennen und mit Blick auf die Richter:innen sind deren Strafeinstellungen oder politische Zugehörigkeit relevant (Niehaus et al., 2009; Rachlinski & Wistrich, 2017; Sporer & Goodman-Delahunty, 2009; Visher, 1987). Zur Beweislage zählen die Qualität und die Quantität der Beweismittel sowie die Art ihrer Präsentation (Niehaus et al., 2009; D. Simon, 2019; Visher, 1987). Die Tatsache, dass juristische Entscheidungen in gewisser Weise von Charakteristiken eines Falles abhängen, ist nicht zwingend negativ zu beurteilen. Allerdings gibt es einerseits legale Faktoren, die begründet auf das zu treffende Urteil einwirken, und andererseits finden sich extralegale Faktoren, die für das Urteil irrelevant sind oder sein sollten. Es ist davon auszugehen, dass letztlich sowohl legale als auch extralegale Aspekte eine juristische Entscheidung ausmachen (s. auch Danziger et

al., 2011). Doch auch Bieneck (2006) argumentiert, dass die Verarbeitung von Informationen bei Entscheidungen im Rahmen der Rechtsprechung datengesteuert sein sollte – zumal Studierende der Rechtswissenschaft in einer analytischen Vorgehensweise der Fallbearbeitung geschult werden. Im Folgenden werden die Einflussfaktoren für die Zwecke dieser Studie dahingehend differenziert, ob sie sich eher auf den Strafprozess (s. 2.3.3) oder auf daran beteiligte Personen (s. 2.3.4) beziehen, wenngleich eine Wechselwirkung zwischen den beiden Bereichen sowie zwischen den verschiedenen Subbereichen besteht. Zur besseren Verständlichkeit wird an geeigneten Stellen mit den Begrifflichkeiten gearbeitet, die Hinweise auf die erfassten Variablen darstellen (s. 3.1). Der Zeitdruck wird, ebenso wie die Beweismittel, bei den prozessbezogenen Faktoren berücksichtigt. Expertise, kognitive Reflexion sowie Need for Cognition werden als personenbezogene Variablen betrachtet.[19] Der Zeitdruck, kognitive Reflexion, Need for Cognition und auch Expertise wurden ausgewählt, weil sie einen Bezug zu den Dual-Prozess-Annahmen aufweisen, die wiederum dieser Arbeit als theoretisches Konstrukt zugrunde liegen. Somit gilt es, deren Einflussgehalt auf juristische Entscheidungen zu untersuchen. Dargestellte Befunde zu Besonderheiten und Einflussfaktoren beziehen sich teilweise auf fachliche und nicht-fachliche Stichproben. Aufgrund ihrer Relevanz für die vorliegende Arbeit wird bestimmte Forschung für Stichproben mit Bezug zur Staatsanwaltschaft an anderer Stelle fokussiert (s. 2.3.5).

2.3.3 Prozessbedingte Einflussfaktoren auf juristische Entscheidungen

Auch wenn die Rahmenbedingungen von gerichtlichen Entscheidungen nicht explizit in der Kategorisierung der Einflussfaktoren nach Visher (1987) genannt werden (s. 2.3.2), können solche Entscheidungen nicht von diesen Bedingungen losgelöst betrachtet werden: „Therefore, judicial decision-making is an interplay between *human* judgment … and *procedural* frameworks" (Sagana & van Toor, 2020, S. 226; s. 2.3.1). Entscheidungen sind in besonderem Ausmaß von der jeweiligen Umgebung abhängig (Hupfeld-Heinemann & Helversen, 2009). Daher wird

[19] Der in der Studie erfasste Delikttyp wird nicht gesondert betrachtet, da dessen Einsatz als Variable vorrangig methodisch statt inhaltlich begründet war, um die neu verfassten Vignetten auf ihre Eignung hin zu überprüfen (s. 3.4.1; 5.2.5).

nachfolgend auf Rahmenbedingungen und Besonderheiten des Strafverfahrens eingegangen, die sich auf den Entscheidungsprozess auswirken können. Dabei kommen legale (z. B. Beweislage; s. 2.3.3.1; 2.3.3.4) und extralegale (z. B. Zeitdruck; s. 2.3.3.2; 2.3.3.3) Wirkfaktoren zur Sprache, wobei inhaltliche Überschneidungen zwischen den Abschnitten möglich sind.

2.3.3.1 Die polizeiliche Akte als Entscheidungsgrundlage im Ermittlungsverfahren

Das Ermittlungsverfahren an sich stellt einen Einflussfaktor dar, der die Entscheidungsumgebung prägt. Die StPO regelt den Ablauf des Strafprozesses und somit auch den Einsatz der beteiligten Akteur:innen wie der Staatsanwaltschaft (s. 2.1). So ist es gemäß dem Legalitätsprinzip die Pflicht der Staatsanwaltschaft, den Sachverhalt bei Verdacht auf eine Straftat zu erforschen, um über die Erhebung einer öffentlichen Anklage entscheiden zu können (§ 160 Abs. 1 StPO). Die Aufgaben der Staatsanwaltschaft sind neben der Erforschung des Sachverhalts auch die Verfahrenssicherung und die Straftatenverhütung (Schroeder & Verrel, 2017). Da diese Aufgabenbereiche die Ressourcen der Staatsanwaltschaft übersteigen (sachliche Behinderung), wird als Ausgleich die Polizei unterstützend tätig (§ 161 StPO). Betrachtet man die Anzahlen der Ermittlungsverfahren sowie der Verfahren, die an deutschen Amts- und Landgerichten verhandelt werden, wird das Ausmaß der zahlenmäßigen Belastung der an diesen Verfahren Beteiligten erkennbar (s. 2.1.5). Dadurch ist nachvollziehbar, dass eine (polizeiliche) Unterstützung vom Gesetz her vorgesehen ist.

Die Aufgaben und Handlungsmöglichkeiten der Polizei werden in § 163 StPO aufgeführt. So gibt es Vorgaben zur Vernehmung einer beschuldigten Person (§ 163a StPO) oder zu Maßnahmen der Identitätsfeststellung (§ 163b StPO). Laut Gesetzesgrundlage ist zwar die Staatsanwaltschaft Herrin des Ermittlungsverfahrens, allerdings zeichnet sich in der Praxis ein verschobenes Bild ab (Schroeder & Verrel, 2017). Satzger (2010) beschreibt dies als eine „*Verpolizeilichung des Ermittlungsverfahrens*“ (S. 98) und Jahn (2015) spricht mit Blick auf die Staatsanwaltschaft von „volljuristischen Hilfsbeamten“ (S. 42) der Polizei. In diesem Zusammenhang wird angemerkt, dass es gewisse Tendenzen seitens Polizei (und Staatsanwaltschaft) gibt, den Richtervorbehalt unter Betonung der Gefahr im Verzug zu umge-

hen und nötige Rücksprachen mit der/dem Ermittlungsrichter:in erst nach dem Ausführen bestimmter Ermittlungen einzuholen (Gusy, 2015; Satzger, 2010). Gründe dafür sind die fehlenden personellen und zeitlichen Ressourcen der Staatsanwaltschaft. Zudem wird der Staatsanwaltschaft häufig erst am Ende der Ermittlungen von der Polizei ein Ermittlungsbericht geschickt, sodass die Staatsanwaltschaft sich, wenn überhaupt, eher mit „Randkorrekturen durch die Anordnung einzelner Nachermittlungen“ (Jahn, 2015, S. 80) beschäftigt. Demnach stellt die Informationssuche und -verarbeitung eine Besonderheit in der strafrechtlichen Entscheidungssituation dar, denn die Ermittlungsergebnisse erhält die Behörde in der Regel in Aktenform: „Die Behörden und Beamten des Polizeidienstes übersenden ihre Verhandlungen ohne Verzug der Staatsanwaltschaft“ (§ 163 Abs. 2 Satz 1 StPO). Die Akten fassen die Polizeiarbeit zusammen, an der die Staatsanwaltschaft unter Umständen nur indirekt aktiv beteiligt war, denn es gibt Fälle, in denen die Polizei zunächst selbstständig ermittelt und die Behörde erst später einbezieht (Büchner, 2022). Somit stellt die Akte eine der wichtigsten Entscheidungsgrundlagen dar und erhält dementsprechend hochrelevante Informationen, wurde aber zuvor federführend von anderen Personen (nämlich den ermittelnden Polizist:innen) zusammengestellt. Die Qualität und die Vollständigkeit der Informationen sind somit nicht direkt erkennbar. Im Sinne des WYSIATI (Kahneman, 2011; s. 2.2.1) ist zu vermuten, dass ein möglicher Informationsmangel nicht erkannt wird, aber die Aktenlage dennoch als umfassendes, kohärentes Material gilt (s. 2.1.6.2).

Dazu passt, dass Kriminaluntersuchungen laut D. Simon (2019) oft unzureichend dokumentiert sind. Dies betrifft beispielsweise die Informationslage zur Befragung tatverdächtiger Personen (Tersago et al., 2020). Dando und Ormerod (2017) untersuchten in ihrer Studie das investigative Vorgehen von Polizist:innen auf Grundlage ihrer Protokolle. Es fanden sich Unterschiede zwischen erfahrenen und weniger erfahrenen Personen in der Vorgehensweise der Beweissammlung. Erfahrene wendeten strategischeres Entscheidungsverhalten an, wohingegen weniger Erfahrene Tendenzen zum Bestätigungsfehler zeigten (s. auch Baber & Butler, 2012). Berufserfahrung führt bei Polizist:innen zum Einsatz nicht-kompensatorischer Entscheidungsstrategien, wohingegen unerfahrene Studierende eher zu kompensatorischen Strategien neigen (Garcia-Retamero & Dhami, 2009). Alison et al. (2013) untersuchten den Einfluss von Zeitdruck auf Polizist:innen im Ermittlungsprozess. Unter

Zeitdruck generierten die Teilnehmenden weniger investigative Hypothesen, wobei allerdings das Zeitgefühl eine moderierende Rolle spielte: Diejenigen, für die die Zeit subjektiv langsamer ablief, entwickelten trotz zeitlichem Stress Hypothesen. Gardner et al. (2019) verglichen Expert:innen diverser forensischer Disziplinen dahingehend, inwiefern sich diese Expertise-Gruppen über die Relevanz oder Irrelevanz bestimmter Informationen einig sind. Zu den Fachbereichen zählten beispielsweise die Spurenanalyse (z. B. Spuren an Waffen, Dokumenten), Chemie (z. B. Toxikologie) und die Ermittlungsarbeit. Generell herrschte in den Fachbereichen jeweils Konsens darüber, welche Informationen wie zu werten sind. So waren Kenntnisse über die beschuldigte Person oder das Opfer (z. B. Name, Geschlecht) für Spurenanalyst:innen oder Chemiker:innen eindeutig irrelevant. Die Gruppe der Ermittler:innen zeigte bei der Einschätzung der Relevanz die größte Variabilität und somit eine gewisse Individualität in den Antworten. Gardner et al. (2019) argumentieren, dass die Ermittlungsarbeit, im Vergleich zu den anderen Disziplinen, weniger spezifische Aufgaben beinhaltet, die wiederum eher auf dem Sammeln als auf dem konkreten Analysieren von Informationen beruhen. Allerdings kann bereits das Sammeln von Informationen und Beweisen am Tatort gewissen Bias unterliegen (Bestätigungsfehler; Ask & Granhag, 2005; s. auch Meterko & Cooper, 2022). Rossmo und Pollock (2018) untersuchten 50 strafrechtliche Fälle, von denen der Großteil unrechtmäßige Strafurteile ausmachte. Von Interesse war dabei, welche Gründe für diese Fehlurteile auszumachen waren. Laut Rossmo und Pollock (2018) traten der Bestätigungsfehler in 37 sowie der Tunnelblick in 24 Fällen auf und beeinflussten die jeweiligen Ermittlungen. Diese Verzerrungen stellten somit die häufigsten festzustellenden Gründe für Fehlurteile dar. Kassin et al. (2013) beschreiben mit dem *Forensischen Bestätigungsfehler* eine Klasse von Effekten, die auf bereits vorhandenen Motiven oder Erwartungen beruhen, aufgrund derer die Sammlung und Interpretation von Beweismitteln beeinflusst wird. Forensische Fachkräfte wüssten demnach um die Vermeidung einer physischen Kontamination der Beweise, wohingegen eine psychologische Kontamination weniger bekannt sei. Hinsichtlich des forensischen Prozesses betonen Morgan et al. (2018) dessen lineare, aber iterative Natur und argumentieren, dass „each subsequent process (and decision) is reliant upon the outcome of a previous process (or decision)" (S. 409). Mit der Beweissammlung am Tatort beginnend gehe es von der Beweisanalyse über deren Interpretation hin zum Gericht, wobei es zu jedem Zeitpunkt des Prozesses

zu Fehlern und Verzerrungen kommen könne (s. auch Dror, 2020; Findley & Scott, 2006; Kang et al., 2012).

Qualitative Unterschiede in der Beweislage, die im Ermittlungsverfahren generiert wird und die letztlich die Entscheidungsgrundlage für die Staatsanwaltschaft darstellt, entstehen bereits zeitlich weit vor der Hauptverhandlung – sofern diese überhaupt stattfindet. In dem Zusammenhang ist die Zweistufigkeit der Beweisaufnahme als prozessbedingte Besonderheit zu nennen (s. 2.1.4), da durch diese die Hauptverhandlung „can be characterized as *pseudo-diagnostic*" (D. Simon, 2011, S. 203; s. 2.1.7). Sagana und Sauerland (2020) argumentieren, dass der Prozess der Informationssammlung während der Ermittlungen einen direkten Einfluss auf die Qualität der Beweismittel hat, die der Staatsanwaltschaft und dem Gericht zur Verfügung gestellt werden (s. auch Kassin et al., 2013; Morgan et al., 2018). In Kautionsentscheidungen verlassen sich Richter:innen in ihren Urteilen auf bereits vorher durch die Polizei oder andere gerichtliche Instanzen getätigte Entscheidungen (Dhami, 2003; s. auch Rachlinski & Wistrich, 2017). Unnötige Hauptverhandlungen finden laut Combé (2007) auch deswegen statt, weil Fehler im Ermittlungsverfahren entstehen, aber bis zur Eröffnung der eigentlichen Hauptverhandlung unkorrigiert bleiben. Beim Bearbeiten der Akte eines Falles ist theoretisch der aufwendige Einsatz von Typ-2-Prozessen erforderlich, um die schnellen und automatischen Denkergebnisse des Typ 1 zu überprüfen, bei Bedarf zu korrigieren und um die Masse an Informationen sachdienlich zu integrieren. Dies erfordert aber (wie jeder Einsatz der Typ-2-Prozesse) einen hohen kognitiven Aufwand (s. 2.2.1). Juristische Entscheidungsfindung beinhaltet eine Art „Rückwärtsdenken", da es ausgehend vom Effekt (Beweislage in der Fallakte) auf die Ursache (Täterschaft oder Unschuld) zu schließen gilt (Rassin, 2020). Zudem gibt es verschiedene Arten von Urteilen: evaluativ, prädiktiv, sozial sowie die Wahrheit einschätzend (Betsch et al., 2011). Die von Verfahrensbeteiligten zu treffenden Urteile besitzen eine hohe Komplexität, da je nach Fall und Fragestellung eine oder mehrere dieser vier Klassen von Inhalten bedient werden müssen, um sodann zu einem geeigneten richterlichen Urteilsspruch zu kommen. Dies trägt mitunter zu hohen kognitiven Anstrengungen bei.

2.3.3.2 Zeit als begrenzte Ressource

Die komplexen Aufgaben der Staatsanwaltschaft – oder allgemein der am Strafverfahren Beteiligten – benötigen nicht nur kognitiven Aufwand, sondern auch Zeit. Letzteres ist im beruflichen Alltag je nach Arbeitsbelastung eine knappe Ressource, sodass manche Entscheidungen unter Zeitdruck gefällt werden müssen. Insbesondere das Beschleunigungsgebot erschwert es vermutlich den juristischen Fachpersonen, das Gleichgewicht zwischen Geschwindigkeit und Gründlichkeit zu erreichen (s. 2.1.5). Diese Form des zeitlichen Stresses kann einen Einfluss auf das Entscheidungsverhalten haben (s. 2.2.3.1). Unter Zeitdruck fehlen der entscheidenden Person die Ressourcen zum Überlegen und Nachdenken (Kang et al., 2012). Liu et al. (2019) betrachteten in ihrer Studie das Strafverhalten unter Zeitdruck. Im Rahmen der Aufgabe konnten Teilnehmende eine andere fiktive Person für ungerechtes Verteilen finanzieller Ressourcen bestrafen oder den Schaden der ungerecht behandelten fiktiven Person kompensieren. Unter Zeitdruck zeigten die Proband:innen verstärkt die Tendenz zur Strafe und weniger altruistisches Verhalten. Richter:innen arbeiten oft unter Zeitdruck und sind mit kognitiver Überlastung konfrontiert, sodass ein Verlass auf intuitive Reaktion wahrscheinlicher wird (Guthrie et al., 2007). Dies ist insbesondere aus dem Grund problematisch, als dass Zeitdruck zu einem verstärkten Einsatz von Heuristiken oder Routinen und dadurch wiederum zu einer größeren Fehleranfälligkeit führen kann (Sagana, 2018; Schweizer, 2009; s. auch Rachlinski & Wistrich, 2017).

2.3.3.3 Routinen und bösartige Lernumgebungen

Doch nicht nur das Erleben von Zeitdruck, sondern auch das Wiederkehren von Entscheidungssituationen führt zu Routinen (Betsch et al., 2011). Mit erneutem Blick auf die Zahlen der (Ermittlungs-)Verfahren (s. 2.1.5) kann abgeleitet werden, dass sich insbesondere in den Amtsgerichten die Fallthemen, sprich die Arten der Vergehen und Verbrechen regelmäßig wiederholen – wenngleich sich die individuellen Begebenheiten eines Falles unterscheiden. Auch wenn die Wiederholungen von Situationen als Lerngelegenheiten gesehen werden könnte, fehlt in der Praxis in der Regel das lehrreiche Feedback (Rachlinski & Wistrich, 2017; Spellman, 2007). Da Rechtsexpert:innen wenig Feedback zu ihren Entscheidungen erhalten, fallen Lerneffekte notgedrungen klein aus (Glöckner & Ebert, 2011; Sagana, 2018; Schweizer, 2015). So kann auch keine Rückmeldung über eventuell aufgetretene

Denkfehler und Bias erfolgen (Pantazi et al., 2020). Dementsprechend können reflektierte Überlegungen nach dem Typ 2 nur eine begrenzte Wirkung haben, wenn diese Überlegungen nicht von gemachten Lernerfahrungen profitieren (Guthrie et al., 2007). „Das ist … besonders gefährlich, weil wiederholtes Entscheiden, selbst ohne relevantes Feedback, dazu führt, dass die Überzeugung, richtig entschieden zu haben, steigt, obwohl es dafür keine rationale Basis gibt…“ (Schweizer, 2015, S. 269). Spellman (2007) fasst es folgendermaßen zusammen: „And trial judges can sit through hundreds of cases and never do the focused study or have the fast reliable feedback necessary for developing expertise” (S. 7). Eventuelle Fehler haben außerdem selten Konsequenzen für Richter:innen (Guthrie et al., 2007), wodurch ein Lerneffekt gering ausfällt.

Laut Schroeder und Verrel (2017) sind fehlerfreie Verfahren nicht möglich, sondern Fehler sprechen eher für die Menschlichkeit der Justiz. Nichtsdestotrotz müsse eine Korrektur von fehlerhaften Entscheidungen in Verfahren möglich sein. Dass „fast alle Versäumnisse … im Laufe des Verfahrens noch *nachgeholt*, Fehler *geheilt* werden“ (S. 14) können, sei typisch für den Prozesscharakter des Strafverfahrens. Dennoch ist dieser Prozesscharakter aufgrund seiner Dauer – zum Beispiel wegen der langwierigen Abläufe von Berufung oder Revision – nicht förderlich für das Lernen der entscheidenden Personen (Schweizer, 2009, 2015). Getroffene juristische Entscheidungen sind nur bedingt reversibel, doch laut Bullens et al. (2014) kann die Möglichkeit zur Änderung einer Entscheidung das damit verbundene Verhalten beeinflussen. Zwar gibt es die Möglichkeit der Beschwerde, diese stellt aber zumeist eine (zeit-)aufwendige Korrektur eines Urteils dar. Nichtsdestotrotz kann die Aussicht auf eine Beschwerde auch ein Motivator sein, sich beim Urteilen Mühe zu geben, um eine gute argumentative Grundlage zu haben, falls ein eigenes getroffenes Urteil von anderen Beteiligten beanstandet und überprüft wird. Für Richter:innen ist es zudem bereits bei der Übernahme eines Falles bekannt, dass das getroffene Urteil letzten Endes begründet werden muss (Maegherman, 2021). Dies kann ebenso wie die Tatsache, dass eine solche Begründung in der Regel vor einem Publikum erfolgt, ein Gefühl von Verantwortung fördern (s. 2.1.4; 2.2.2). Allerdings steht zu Beginn eines Hauptverfahrens nicht zwingend fest, vor wem es sich im Rahmen der Urteilsverkündung – neben den ständig anwesenden Verfahrensbeteiligten – zu erklären gilt, z. B. Zuschauende, Presse (Maegherman, 2021).

In einer Studie von Huber und Seiser (2001) wurde der Unterschied von Rechtfertigung und Überzeugung untersucht. Proband:innen wurden aufgefordert, sich für die Unterbringung einzelner krimineller Jugendlicher zu entscheiden. Während ein Teil der Gruppe die Aufgabe bekam, die Entscheidung im Nachgang zu rechtfertigen, erhielt der andere Teil die Aufgabe, eine Wahl lediglich zu empfehlen. Beide Experimentalgruppen nutzten die vorhandenen Informationen stärker als die Kontrollgruppe, wenngleich sich die Informations*menge* in den Experimentalgruppen nicht unterschied. Diejenigen, deren Aufgabe die Überzeugung war, benötigten aber mehr Zeit und lieferten elaboriertere Argumentationen als die Gruppe *Rechtfertigung*. Die Aussicht darauf, sich im Nachgang noch weiter mit einer Entscheidung (vor anderen Personen) auseinander setzen zu müssen, kann demnach den Prozess beeinflussen (s. auch Bieneck, 2006). Fraglich ist, inwiefern der zeitlich verzögerte Moment der Überprüfung juristischer Urteile – sofern er im Rahmen einer Beschwerde überhaupt stattfindet – zu vergleichbaren Effekten führt. Maegherman (2021) verglich niederländische und deutsche Richter:innen, deren Aufgabe sich dahingehend grundlegend unterscheidet, dass erstere aufgrund der Gesetzesgrundlage eher eine Rechtfertigung darlegen, wohingegen letztere eine Erklärung für ihr Urteil abgeben. Eine derartige Erklärung erfordert eine tiefergehende Auseinandersetzung mit einem Fall als eine rechtfertigende Begründung. Die Ergebnisse von Maegherman (2021) zeigten, dass die verschiedenen Anforderungen Auswirkungen auf den Umgang mit Beweismitteln in einer Mord-Vignette hatten: Die Stichproben unterschieden sich zwar nicht in der Wahrnehmung der Schuld, allerdings ging die Notwendigkeit einer Erklärung mit einer stärkeren Betrachtung von entlastenden Beweisen einher, die wiederum ein positiver Prädiktor für einen Freispruch waren.

Im Sinne der Einteilung von Lernumgebungen nach Hogarth (2010) lassen sich juristische Entscheidungssituationen als bösartig definieren (s. auch Schweizer, 2009, 2015; s. 2.2.2). Lehrreiches Feedback ist nur begrenzt möglich (Rachlinski, 2000). D. Simon (2019) betont, dass die Kritik nicht an die Rechtsexpert:innen gerichtet ist, sondern sich vielmehr auf die Umstände bezieht, unter denen von diesen Fachpersonen erwartet wird, ein Urteil zu treffen. Mit Blick auf die Konsequenzen gibt es die Besonderheit, dass die Beteiligten im Hauptverfahren (Mitarbeitende der

Verteidigung und der Staatsanwaltschaft, Richter:innen) nur indirekt von den Entscheidungen betroffen sind. Urteile hinsichtlich der Schuld oder des Strafmaßes betreffen diese Beteiligten nicht, da sie sich direkt auf den angeklagten Menschen beziehen („Ultimately lawyers decide on people's lives", Engel, 2017, S. 14). Dies schmälert nicht zwingend den Aufwand (des System 2), der zur Findung eines gerechten Urteils benötigt wird, aber die Verschiebung der direkten Betroffenheit auf andere Personen führt dementsprechend auch zu einer Verschiebung der erlebten Konsequenzen. Das Konsequenzerleben als wichtige Feedbackkomponente in der Entscheidungsfindung findet im Strafverfahren eher indirekt statt, zumindest was die konkreten Inhalte der Schuld und des Strafmaßes angeht. Im Ermittlungsverfahren ist das Konsequenzerleben unklar. Die in diesem Verfahrensabschnitt gesammelten Beweise und getroffenen Entscheidungen dienen letztlich zur inhaltlichen Vorbereitung des Zwischen- und insbesondere des Hauptverfahrens (Zweistufigkeit der Beweisaufnahme, Prinzip der Mündlichkeit; s. 2.1.1). Welche Informationen und Entscheidungen letztlich überhaupt ausgewertet werden und zum Urteil beitragen, zeigt sich somit in der Regel erst zeitlich verzögert in der Hauptverhandlung. Ob diese überhaupt stattfindet, beruht aber wiederum mit großer Schnittmenge auf den Handlungen und Entscheidungen, die im Ermittlungsverfahren stattgefunden haben (s. 2.3.3.1).

2.3.3.4 Überzeugungskraft von Beweismitteln: Objektive Beweislast, Unschuldsvermutung, Beweismaß und Verwertungsverbote

In vielen Experimentalstudien werden bestimmte Variablen (z. B. die Beweise betreffend) innerhalb eines Strafprozesses manipuliert, um Auswirkungen auf Urteile zu erfassen (Hupfeld-Heinemann & Helversen, 2009). Beweismittel werden in ihrer Würdigung unterschiedlich gewichtet, beispielsweise in Abhängigkeit bereits erfolgter Interpretationen (Sagana & Sauerland, 2020). Zu den hier relevanten Beweismitteln zählen der Augenschein, Urkunden, Zeug:innen und die beschuldigte beziehungsweise die angeklagte Person (s. 2.1.6).

Beweismittel der Kategorie *Augenschein* lassen sich mit den Sinnen wahrnehmen. In Form von beispielsweise Fotos oder Skizzen stellen sie eine Art physischen Beweis dar (wie *Urkunden* auch). Physische Beweismittel, die eine Verbindung zwischen dem Tatort und der beschuldigten beziehungsweise der angeklagten Person herstellen, haben einen großen Einfluss auf die Einschätzung der Schuldfrage

(Pearson et al., 2018; s. auch Daftary-Kapur et al., 2010). Dies liegt vermutlich daran, dass eine solche eindeutige Verbindung bei der Bewertung von Aussagen helfen kann.[20]

Ein Merkmal, das mit Blick auf die beschuldigte oder die angeklagte Person genannt werden muss, ist die Vorstrafe. In Deutschland gibt das Bundeszentralregister darüber Auskunft (§ 3 Bundeszentralregistergesetz). Da es sich um ein schriftliches Dokument handelt, das in Verhandlungen in der Regel verlesen wird, gilt es im Sinne der Beweiskategorien für die Zwecke dieser Studie als Beweismittel *Urkunde* (s. 3.4.2) – auch wenn sich der Inhalt direkt auf die beschuldigte oder die angeklagte Person bezieht. Im anglo-amerikanischen Raum dürfen Vorstrafen nicht für die Einschätzung der Wahrscheinlichkeit genutzt werden, dass eine Person ein Verbrechen tatsächlich begangen hat, da sie nur zur Einschätzung der Glaubwürdigkeit der Person eingesetzt werden sollen (Allison & Brimacombe, 2010; Oswald, 2009). In einer Studie von Greene und Dodge (1995) stieg die Wahrscheinlichkeit für eine Verurteilung, wenn Proband:innen Auskunft über die Vorstrafen der angeklagten Person erhielten, im Vergleich zu Situationen, in denen sie Auskünfte über vorherige Freisprüche oder gar keine derartigen Informationen erhalten haben. Auch in einer Studie von Pearson et al. (2018) zeigte sich, dass eine einschlägige Vorstrafe die wahrgenommene Überzeugung der Schuld um 10 Punkte (100-Punkte-Skala) anhob, wohingegen eine Vorstrafe für ein anderes Verbrechen zu einer Erhöhung um 5 Punkte führte. Insbesondere in Fällen mit nicht eindeutiger Beweislage können Vorstrafen laut Pearson et al. (2018) für eine Verurteilung entscheidend sein (s. auch Allison & Brimacombe, 2010). In einer Literaturübersicht zum Einfluss von Vorstrafen auf die Schuldfrage kommt Schmittat (2022) allerdings zu dem Schluss, dass dieser Effekt eher als klein einzuschätzen ist und in Abhängigkeit zu Moderatorvariablen steht (z. B. Einschlägigkeit und Zeitpunkt der Vorstrafe). Einschlägige Vorstrafen werden auch für die Bewertung der Stärke eines Alibis berücksichtigt und können zu einer höheren Einschätzung der Schuld führen (Allison & Brimacombe, 2010).

[20] Der CSI-Effekt beschreibt die hohen Erwartungen von Geschworenen (also Laien) an forensische Beweisstandards, auch wenn diese Standards nicht als fehlerfrei gelten und selbst Richter:innen nicht eigens dafür ausgebildet sind, solche Beweise zu bewerten und einzuschätzen. Für weiterführende Informationen s. Daftary-Kapur et al. (2010), Kassin et al. (2013) und Morgan et al. (2018).

Ein Alibi ergibt sich oft aus den Angaben der beschuldigten oder der angeklagten Person, welche wiederum als Beweismittel der *Einlassung* gelten.[21] Allison und Brimacombe (2010) argumentieren, dass schwache Alibis nicht zwingend Schuld implizieren, wenngleich in ihrer Studie starke Alibis durchaus mit als niedriger eingeschätzter Schuld und größerer Glaubhaftigkeit einhergingen. Die Einlassung kann aber auch die Form eines Geständnisses annehmen. Nach dem Geständnis einer verdächtigen Person besteht die Gefahr, dass keine umfassenden Ermittlungen in andere Richtungen mehr durchgeführt werden (Schmittat, 2017). Laut Tersago et al. (2020) bemühen sich Richter:innen mehr darum, eine Leugnung zu falsifizieren als ein Geständnis: „The relationship between convictions and the consistency of confessions with other evidence is stronger than the relationship between acquittals and the consistency of denials with other evidence“ (S. 184–185). Geständnisse haben eine enorme Kraft – insbesondere wenn sie mit objektiven Beweismitteln übereinstimmen (Tersago et al., 2020) – obwohl es durchaus Gründe für falsche Geständnisse gibt (z. B. eine Bedrohungssituation durch externe Personen). Ein falsch abgelegtes (und wieder zurückgezogenes) Geständnis birgt für die angeklagte Person ein Risiko, weil es für Verfahrensbeteiligte kaum möglich ist, ein solches Geständnis zu ignorieren (Schmittat, 2017). Ein Grund, warum es bei Einlassungen schwerfällt, zwischen Wahrheit und Täuschung zu unterscheiden, ist der Mangel an validen Hinweisen (Wyler, 2021). Zusätzliche Beweismittel können die Richtigkeit einer Aussage untermauern oder widerlegen. Mit Blick auf die beschuldigte oder die angeklagte Person gibt es noch weitere (extralegale) Faktoren, die eine Entscheidung beeinflussen können. Beispielsweise stehen Kautionsentscheidungen zwar durchaus im Zusammenhang mit Art und Schwere der Straftat, aber auch mit Geschlecht, Alter und ethnischer Herkunft der angeklagten Person (Dhami & Ayton, 2001; s. auch Forsterlee et al., 2006; Niehaus et al., 2009). Auch die Attraktivität dieser Person kann eine Wirkung auf die Einschätzung der Schuld haben (Scurich et al., 2016; Sporer & Goodman-Delahunty, 2009). Ebenso lassen

[21] Es ist auch möglich, dass sich ein Alibi aus anderen Beweismitteln ableiten lässt, wenn keine Einlassung der Person erfolgt (Schweigerecht). Für eine Übersicht zu Befragungsmethoden in der Vernehmung von Tatverdächtigen s. Wyler (2021).

sich Faktoren hinsichtlich der geschädigten Person ausmachen (z. B. Geschlecht, Herkunft, Ausmaß des Schadens; Sporer & Goodman-Delahunty, 2009).[22]

Eine weitere Beweiskategorie lautet *Zeug:in*. Aussagen können sich auf die Zeit unmittelbar vor, während oder nach der Tat sowie auf Beschreibungen von Handlungen, Umständen und Personen beziehen. Die Fähigkeiten der Zeug:innen zum Wahrnehmen und Erinnern werden allerdings oft falsch eingeschätzt, obwohl sie in ihrer (selektiven) Wahrnehmung einer Tat auf vielseitige Weise beeinflusst werden können (z. B. Stereotype; Schmittat, 2017). Ebenso wie bei der Einlassung ist eine Einschätzung der Glaubwürdigkeit bei Aussagen von Zeug:innen wichtig (Niehaus et al., 2009; s. auch Effer-Uhe & Mohnert, 2019). Auch hier kann ein Mangel an Informationen die Korrektheit dieser Einschätzung schmälern. D. Simon (2019) fasst Gründe zusammen, warum Aussagen problembehaftet sein können. Dazu zählen die fehlerhafte Wahrnehmung der Tat (z. B. aufgrund begrenzter Aufmerksamkeit oder schlechter Sichtverhältnisse), polizeiliche Ermittlungen (z. B. Suggestivfragen) sowie eine über die Dauer eines Verfahrens stattfindende Veränderung der erinnerten Inhalte (z. B. Integration zusätzlicher Informationen). Wixted et al. (2018) argumentieren, dass solche Aussagen aber durchaus zuverlässig sein können, wenn sie zuvor nicht kontaminiert wurden. Aussagen von Zeug:innen haben unter Umständen bei der Klärung der Schuldfrage aber eine geringere Kraft als physische Beweismittel, wie beispielsweise Abgleiche von Erbinformationen (Pearson et al., 2018).

Die objektive Beweislast liegt im Strafverfahren nicht bei der angeklagten Person, denn das Gericht muss deren Schuld beweisen (Schweizer, 2019; D. Simon, 2012). Ein Freispruch ist allerdings nicht zwingend mit Unschuld gleichzusetzen (Tersago et al., 2020). Die Unschuldsvermutung erschwert intuitives Handeln aus gutem Grund: Rechtsexpert:innen müssen – auch wenn es schwerfällt – entgegen der intuitiven Vermutung denken und handeln, dass sich eine beschuldigte oder eine angeklagte Person nicht ohne Grund in der Situation befindet (D. Simon, 2019). Durch die Unschuldsvermutung lässt sich aber das Risiko zumindest reduzieren, dass eine eigentlich unschuldige Person auf intuitiver Entscheidungsgrundlage zu Unrecht

[22] Da diese Variablen in dieser Studie nicht manipuliert wurden, wird an dieser Stelle nicht näher auf einzelne Befunde eingegangen.

verurteilt wird. Scurich und John (2017) konnten zeigen, dass als Geschworene eingesetzte Teilnehmende, die nicht auf die Unschuldsvermutung hingewiesen worden waren, einen gewissen Argwohn angesichts der Tatsache, dass es überhaupt zu einer Beschuldigung gekommen war, als einen Hinweis für Schuld nutzten (s. auch D. Simon, 2012). Schuld wird in der Regel als Dichotomie verstanden. Während normalerweise nur zwischen „schuldig" und „nicht schuldig" unterschieden wird, kann eine zusätzliche dritte Option („nicht bewiesen") ausdrücken, dass die Schuld zwar nicht über das nötige Maß hinaus bewiesen ist, aber es dennoch gewisse Zweifel an der Unschuld der angeklagten Person gibt – auch wenn das rechtliche Ergebnis letztlich mit der Unschuld gleichzusetzen ist (Freispruch). Hope et al. (2008) untersuchten, inwiefern die Ergänzung dieser Option zu veränderten Reaktionen führt. Bei schwacher oder starker Beweislage wurde sich nicht vermehrt dafür entschieden. Wenn die Beweislage aber als moderat beschrieben werden konnte, wählten Teilnehmende vermehrt diese dritte Möglichkeit anstatt der „schuldig"-Option. Laut Hope et al. (2008) schätzten die Teilnehmenden dies aber als eine Art Freispruch zweiter Klasse ein, auch wenn sich dadurch die Gelegenheit ergab, die eigenen Ansichten zur (Un-)Schuld der angeklagten Person zu differenzieren (s. auch Curley et al., 2022; Koch & Devine, 1999; Ormston et al., 2019). Ergänzend dazu zeigte sich in einer Studie von Smithson et al. (2007) allerdings nicht, dass Entscheider:innen die „nicht bewiesen"-Option als ausweichende Alternative nutzten, um ein extremes Schuld-/Unschuld-Votum zu vermeiden.

Im strafrechtlichen Kontext gilt eine volle Überzeugung als notwendiges Beweismaß, was einem Überzeugungsgrad von etwa 90% oder mehr entspricht (s. 2.1.6.1). Für die überwiegende Überzeugung reichen mehr als 50% aus. Zumindest mit Blick auf Laienrichter:innen ist allerdings zu vermuten, dass diese ein unterschiedliches Verständnis des Beweismaßes haben, je nachdem, wie es erklärt, instruiert und in welchen Situationen es angewendet wird (Baucum et al., 2018; Daftary-Kapur et al., 2010; Dhami et al., 2015; Mueller-Johnson et al., 2018; D. Simon, 2012; Wright & Hall, 2007). Kagehiro und Stanton (1985) verglichen ein quantifizierbares und ein juristisches Beweismaß. Sie zeigten, dass Naive von einem als Zahl oder als Wahrscheinlichkeit ausgedrückten Überzeugungsgrad profitierten (z. B. „51%"), im Gegensatz zum Einsatz der juristischen Begriffe (z. B. „überwiegende Überzeu-

gung“). Die Kombination beider Formen führte aber nicht dazu, dass die juristischen Begrifflichkeiten besser verstanden wurden.[23] Laut Lundberg (2016) kann die Schwere des Verbrechens dazu führen, dass das Beweismaß angepasst wird. Bei schweren Verbrechen sei die Wahrscheinlichkeit für einen Schuldspruch niedriger, sodass das für einen Schuldspruch nötige Beweismaß mit steigender Schwere angehoben werde. Unter Umständen haben auch Expert:innen ein ungleiches Gefühl darüber, wann welches Maß erreicht wurde (s. auch Schweizer, 2016). Auch wenn in einer Studie von Lavie et al. (2020) die juristische Expertise mit einer besseren Passung zwischen der individuellen Interpretation und der gesetzlich angedachten Definition des Grades einherging, so stimmten diese Größen dennoch nicht überein, da die individuelle Zahl höher angesetzt wurde. Zu ähnlichen Ergebnissen kam Schweizer (2016) dahingehend, als dass Richter:innen einen anderen Standard anlegten als es die gesetzliche Grundlage vorsah. Schweizer (2016) äußert zudem die Vermutung, dass es, unabhängig von der gesetzlichen Vorgabe, eine Art intuitive oder natürliche Entscheidungsgrenze gibt. In einer Literaturübersicht kommt Conklin (2020) zu dem Schluss, dass die politische Zugehörigkeit von Richter:innen oder die Wahrnehmung über die angeklagte Person zu einer Anpassung des Beweismaßes führen kann. Legale Standards können individuellen Einflüssen unterliegen, müssen aber nicht: „This interplay is not universal among jurors“ (Conklin, 2020, S. 290). Trotz dieser empirischen Ausgangslage argumentieren Glöckner und Engel (2013), dass Menschen durchaus in der Lage sind, legale Standards anzuwenden, ohne dass numerische Wahrscheinlichkeiten zur Verurteilung genutzt werden müssen. D. Simon (2012) fasst darüber hinaus gewisse Chancen und Risiken, die mit dem Beweismaß verbunden sind, folgendermaßen zusammen:

> Yet, even if the standard of proof did reduce conviction rates, it would not necessarily increase the accuracy of verdicts. The standard is merely a sorting mechanism, and is devoid of any diagnostic properties of its own. It relies on the fact finder’s ability to correctly assess the accuracy of evidence in the particular case. (S. 193)

Eine weitere Besonderheit des Strafverfahrens, die ähnlich wie die Unschuldsvermutung gewissermaßen wider Intuition arbeitet, sind Beweisverwertungsverbote. Derartige Verbote beziehen sich beispielsweise auf solche Beweismittel, die un-

[23] Kagehiro und Stanton (1985) stellen die Vermutung auf, dass dies daran gelegen haben könnte, dass die juristische Beschreibung *vor* der Quantifizierung des Beweismaßes präsentiert wurde.

rechtmäßig erhoben worden sind (z. B. Misshandlung, § 136 StPO). „Die höchstrichterliche Rechtsprechung setzt – wenig überraschend – erhebliches Vertrauen in die Fähigkeit der (Berufs-)Richter zur Nichtberücksichtigung bereits erhobener, sodann jedoch für unverwertbar erachteter Beweismittel“ (Lindemann, 2015, S. 131). Empirische Befunde weisen allerdings darauf hin, dass Fachpersonen ebenso wie Laien unter Umständen nicht zuverlässig in der Lage sind, solche Informationen unberücksichtigt zu lassen (z. B. Daftary-Kapur et al., 2010; Landsman & Rakos, 1994; Lieberman & Arndt, 2000; Lindemann, 2015; Pickel et al., 2009; Rachlinski & Wistrich, 2017; D. Simon, 2012). Dies kam in einer Studie von Wistrich et al. (2005) gar bei einigen Richter:innen vor, die zuvor selbst diejenigen gewesen waren, die die Unzulässigkeit der Beweismittel bestimmt hatten. Somit kann es in Einzelfällen zur Integration sachlich ungeeigneter Informationen kommen – auch wenn durchaus der Versuch unternommen wird, unzulässige Beweismittel zu ignorieren (Daftary-Kapur et al., 2010). Zudem gibt es Rahmenbedingungen, unter denen dies besser gelingen kann. In einer Studie von Fleming et al. (1999) mit Schein-Geschworenen spielte die Frage eine Rolle, ob letztlich unzulässige Beweise in einem Vergewaltigungsfall mit einer milden oder einer starken Verletzung der Prozessordnung erhoben wurden. Als milde Verletzung galt beispielsweise ein seit wenigen Minuten abgelaufener Durchsuchungsbefehl, als starke Verletzung galt dagegen das Fehlen eines solchen Befehls. Schein-Geschworene waren besser dazu in der Lage, die Beweise zu ignorieren, wenn diese nicht mit einer starken, sondern nur mit einer milden Prozessverletzung einhergingen.

Weiterhin bleibt offen, was bei strafrechtlichen Urteilen – im Sinne der normativen Nutzentheorie (s. 2.2.1) – als Gewinn oder Verlust interpretiert werden kann, sprich, welche Situationen erstrebenswert sind und welche es zu vermeiden gilt. Für die Wissenschaften sind falsch-positive Ergebnisse zu vermeiden (Alpha-Fehler). In vielen juristischen Kontexten ist dies ebenfalls der Fall, da keine eigentlich unschuldige Person unrechtmäßig verurteilt werden darf (Engel, 2017; s. 2.1.6.1). Das Entscheidungserleben kann ein Hinweis darauf sein, ob ein Entscheidungsprozess eher als positiv oder als negativ angesehen wird. Zu Eigenschaften positiver Erfahrungen zählen ein schneller Prozess und Zufriedenheit mit der Entscheidung, wohingegen Stress, wahrgenommene Schwierigkeit und Prokrastinieren eher mit negativen Erfahrungen in Verbindung gebracht werden (W. J. Phillips et al., 2016).

Das Gefühl von Reue kann ein Indikator dafür sein, dass sich eine Person im Nachgang eine andere Entscheidung gewünscht hätte (Engel, 2017): Reue kann einerseits den Umstand betreffen, dass eine Person handeln wollte, dies aber nicht tat, und andererseits ist es möglich, dass jemand handelte, dies aber nicht tun wollte. Das Gefühl von Sicherheit kann laut Engel (2017) dabei helfen, beide Situationen zu vermeiden. Wie bereits angedeutet gibt es in juristischen Kontexten aber häufig nicht die eine richtige Entscheidung (Gigerenzer, 2006; s. 2.2.3.2) – zumal diese für Rechtsexpert:innen aufgrund der unrealistischen objektiven Wahrheit unerreichbar ist (Maegherman, 2021; s. 2.1.6.1). Eine Möglichkeit, sich der Ermittlung der Güte einer Entscheidung zu nähern, ist die Untersuchung der schriftlichen Urteilsbegründung, die Hinweise auf (fehlende) Gerechtigkeit oder auf das Vorkommen von Bias liefern kann (Maegherman, 2021). Doch auch wenn sich Entscheidungen von Rechtsexpert:innen nicht auf simple, dichotome Weise in „richtig" oder „falsch" einordnen lassen (Rachlinski & Wistrich, 2017), so wurde in den Abschnitten 2.3.3–2.3.3.4 doch dargestellt, welche prozessbedingten Wirkfaktoren und Besonderheiten sich auf juristische Entscheidungsprozesse auswirken können (Rahmenbedingungen des „bösartigen" Ermittlungsverfahren: z. B. defizitäre Informationsgrundlage, Zeitdruck, reduzierte Lerneffekte; Beweislage: z. B. Arten der Beweismittel, Unschuldsvermutung, Beweismaß, Beweisverwertungsverbote).[24]

2.3.4 Personenbedingte Einflussfaktoren auf juristische Entscheidungen

Gigerenzer (2006) argumentiert, dass die beste Lösung für juristische Frage- und Problemstellungen in der Regel nicht erreicht werden kann. Weiter führt er aus, dass sich Menschen angesichts dieser Unmöglichkeit nicht in eine Schockstarre begeben, sondern trotzdem zu Lösungsideen kommen. Es kann geschlussfolgert werden, dass Menschen nicht zwingend die gleichen Lösungsideen entwickeln, sondern dass interindividuelle Unterschiede zum Tragen kommen. In Abschnitt 2.3.2 wurde bereits angedeutet, dass Fachpersonen ungleich entscheiden und handeln, was als Varianz (oder Diskrepanz) gedeutet wurde. Maguire (2010) differenziert

[24] Auch wenn juristische Entscheidungen und Urteile das Ergebnis eines *legalen* Prozesses sind, evaluieren Laien solche Urteile oft dahingehend, ob diese mit ihren persönlichen Einschätzungen übereinstimmen, unabhängig davon, ob das Strafverfahren rechtmäßig durchgeführt wurde oder nicht (D. Simon & Scurich, 2011, 2013).

zwischen Inkonsistenz und Ungleichheit beim Urteilen: Inkonsistenz zeigt sich, wenn ähnliche Fälle *begründet* unterschiedlich verhandelt werden; Ungleichheit beschreibt den Umstand, dass ähnliche Fälle *unbegründet* unterschiedlich verhandelt werden. Dabei ist relevant, welche (extra-)legalen Faktoren zur Inkonsistenz oder gar zur Ungleichheit beitragen.

Hinweise darauf, dass Fachpersonen nicht zwingend einer Meinung sind, gibt es über verschiedene Länder hinweg (z. B. Finnland, England und Wales: Davies, 2004; Irland: Maguire, 2010; USA: Austin & Williams, 1977; Kramer & Ulmer, 1996). Die Gründe für diese Diskrepanz sind vielseitig. Unterschiede in den Entscheidungen können – wie bereits angedeutet – auf extralegale Faktoren (z. B. Einstellungen der Jurist:innen, Charakteristiken der beschuldigten oder angeklagten Person) oder auf legale Faktoren (z. B. Stärke der Beweislage, Schwere des Verbrechens) zurückzuführen sein (Austin & Williams, 1977; s. auch Ellison & Brennan, 2016; Kramer & Ulmer, 1996; Maguire, 2010). Weitere mögliche Erklärungen für die Diskrepanz sind Zeitdruck, fehlende prozessuale Regeln und fehlende Informationen (Dhami & Ayton, 2001; s. 2.3.3.1; 2.3.3.2). Eine Möglichkeit der Reduktion von Ungleichheit sind fachliche Richtlinien. Anderson et al. (1999) verglichen die durchschnittliche Länge von Gefängnisurteilen vor (1986–1987) und nach (1988–1993) der Einführung von föderalen Urteilsrichtlinien in den USA. Diese Richtlinien beinhalten beispielsweise eine Tabelle, anhand derer die Schwere eines Verbrechens an die Vorstrafen der beschuldigten Person angepasst werden kann. Der zu erwartende durchschnittliche Unterschied zwischen zwei Richter:innen reduzierte sich im Schnitt von 4.9 auf 3.9 Monate (s. auch Ellison & Brennan, 2016; Kautt, 2009). Nichtsdestotrotz ist Inkonsistenz beim Urteilen laut Maguire (2010) durchaus zu erwarten:

> Firstly, individualised sentencing systems by their very nature always give rise to a certain degree of inconsistency. Differences in case factors means that no two cases are ever precisely the same and this results in small but often noticeable differences in sentencing outcomes between two seemingly similar cases. (S. 18)

Es gilt zu wiederholen und weiter auszuführen, dass juristische Entscheidungen auf Grundlage legaler Faktoren stattfinden sollten, es aber prozess- und personenbezogene Einflüsse gibt, die zu Inkonsistenzen – oder gar Ungleichheiten – beim Urteilen führen können (s. 2.3.2). Rachlinski und Wistrich (2017) schlussfolgern: „Judges do not easily set such extralegal matters aside“ (S. 31). In Abschnitt 2.3.3

wurde bereits auf prozessuale Wirkfaktoren eingegangen, was in diesem Abschnitt durch die Betrachtung von personenbezogenen Einflüssen ergänzt wird. Gemäß Vishers (1987) dreigeteilter Kategorisierung von Einflussfaktoren entspricht dies am ehesten den Charakteristiken der Richter:innen (s. 2.3.2). Allerdings werden an dieser Stelle nicht nur die rechtsprechende Person betrachtet, sondern auch Laienrichter:innen und andere am Strafverfahren Beteiligte. Auch wenn einzelne Einflussfaktoren im Folgenden eher getrennt betrachtet werden, so sind dennoch Interaktionen möglich: „Some of these factors may not act in isolation, but may interact with each other in complex ways“ (Sporer & Goodman-Delahunty, 2009, S. 380). Obwohl Visher (1987) eine Kategorisierung vornimmt, gibt es laut Sporer und Goodman-Delahunty (2009) keine einheitliche Definition (extralegaler) Wirkfaktoren. Für die Zwecke dieser Studie erfolgt die inhaltliche Gruppierung von personenbezogenen Variablen mit Blick auf die Beteiligten (Expertise, s. 2.3.4.2; kognitive Reflexion, s. 2.3.4.3; Need for Cognition, s. 2.3.4.4).[25] Zunächst erfolgt ein Exkurs zum Auftreten von Heuristiken und Bias im juristischen Kontext (s. 2.3.4.2)

2.3.4.1 Exkurs: Heuristiken und Bias im juristischen Kontext

In Abschnitt 2.3.2 wurde die Sinnhaftigkeit der Übertragung von Dual-Prozess-Annahmen auf die juristische Entscheidungsfindung begründet. Der Einsatz von Heuristiken und das Auftreten von Bias kann mit diesen Annahmen in Verbindung gebracht werden (s. 2.2.1; 2.2.1.2). Demzufolge kann gewissermaßen im Umkehrschluss beim Auftreten solcher Heuristiken und Verzerrungen auf das Wirken von Typ-1- und Typ-2-Prozessen, also auch von Intuition und Reflektion, geschlossen werden, was wiederum den Prozess der juristischen Entscheidungsfindung mutmaßlich beeinflusst. Laut Rachlinski und Wistrich (2017) verlassen sich Richter:innen sehr stark auf intuitives Denken, um rechtliche Fragen und Probleme anzugehen. Dabei würden einfache, mentale Abkürzungen zum Einsatz kommen. Doch welche Hinweise finden sich im juristischen Kontext für dieses interindividuelle, extralegale Wirken?

[25] Für weiterführende Informationen zu Eigenschaften juristischer Fachpersonen (z. B. Geschlecht, politische oder straftheoretische Einstellungen) wird auf Collins et al. (2010), Ellison und Brennan (2016), Niehaus et al. (2009), Rachlinski und Wistrich (2017) sowie Sporer und Goodman-Delahunty (2009) verwiesen. Für Informationen über den Einfluss von Emotionen auf juristische Entscheidungen s. Gabriel (2009).

Heuristiken und Bias treten in den verschiedensten Entscheidungskontexten auf. Gemäß der Ansätze von Tversky und Kahneman (1974) sowie von Gigerenzer et al. (1999) stehen diese aber nicht notwendigerweise im Zusammenhang mit schlechten Entscheidungen, sondern erfüllen durchaus einen positiven Nutzen. Es liegt die Annahme nahe, von Fachpersonen aufgrund ihrer Expertise zu erwarten, dass sie ihre fachliche Aufgabe von irrelevanten Faktoren und Prozessen unbeeinflusst erfüllen können. Doch Dror (2020) nennt es einen Irrtum, dass Expert:innen gegen Einflüsse und Bias immun sind. Rassin (2020) bringt es folgendermaßen auf den Punkt: „Professional judges are susceptible to bias" (S. 85–86). Deren Beeinflussbarkeit ist dabei mannigfaltig, sodass lediglich beispielhafte Befunde berichtet werden können. In alltäglichen Situationen ist es eine effiziente Tendenz, vorhandene Informationen von Mitmenschen zunächst als wahr anzuerkennen (Pantazi et al., 2020). Dies stellt aber in gerichtlichen Situationen eine Quelle des Wahrheits-Bias dar, denn eine Besonderheit in Gerichtsverhandlungen ist die Tatsache, dass es aus verschiedenen Gründen sehr wahrscheinlich sein kann, dass präsentierte Informationen unwahr sind (z. B. weil sich die angeklagte Person schützen möchte; s. auch Engel, 2020). Der Wahrheits-Bias beschreibt die Tendenz, Informationen zu glauben, obwohl es Hinweise auf deren Inkorrektheit gibt. Pantazi et al. (2020) untersuchten mittels Kriminalberichten, inwiefern Richter:innen und Geschworene vom Wahrheits-Bias beeinflusst wurden. Sie zeigten, dass die Teilnehmenden von solchen Informationen beeinflusst wurden, von denen man wusste, dass sie inkorrekt sind. Auch die Tatsache, dass die Beteiligten für ihre Entscheidungen zur Verantwortung gezogen werden könnten, führte nicht zu weniger starken Verzerrungen.[26] In der Studie von Rassin (2020) zeigten sich Richter:innen in einer Vignette zu einem Fall von Körperverletzung von irrelevanten Kontexteffekten beeinflusst: Informationen, wie beispielsweise das Aussehen der tatverdächtigen Person, wirkten sich auf die Verurteilungsrate aus. Die Präferenz der Teilnehmenden für bestimmte Nachermittlungen deutete dahingehend auf den Bestätigungsfehler hin, als dass zusätzliche belastende Hinweise gesucht werden wollten, die die bereits belastende Beweislage weiter untermauerten. Auch wenn dies nicht mit klassischem heu-

[26] Gleichwohl zeigen sich empirische Hinweise darauf, dass Personen durchaus aufmerksamer werden, wenn sie es erwarten getäuscht zu werden (Chambers & Zaragoza, 2001; s. auch Levine et al., 2004).

ristischem oder urteilsverzerrtem Entscheidungsverhalten gleichzusetzen ist, befassten sich Danziger et al. (2011) in ihrer Studie mit einem besonderen extralegalen Einflussfaktor, nämlich der (Essens-)Pause von Richter:innen. Die Rate von für die Kläger:innen günstigen Bewährungsurteilen sank von rund 65% vor einer Pause auf nahezu 0% und stieg nach einer Pause auf den vorherigen Wert an. Die Wahrscheinlichkeit für eine Bewährung war somit am Arbeitsbeginn und nach einer Pause höher als am Ende einer Arbeitseinheit. Guthrie et al. (2001) untersuchten die Beeinflussbarkeit von Richter:innen im Vergleich zu Naiven hinsichtlich ausgewählter Bias. Die Leistung juristischer Fachpersonen war bei drei von fünf Urteilsfehlern vergleichbar mit der Leistung von Nicht-Fachpersonen (z. B. Rückschaufehler, Ankereffekt).[27] Bei Rahmeneffekten und der Repräsentativitätsheuristik schnitten sie hingegen besser ab. Hinsichtlich des Ankereffekts zeigte sich in einer Studie von Bieneck (2006), dass Personen im Rechtsreferendariat weniger beeinflusst waren als Studierende der Rechtswissenschaft. Miller (2019) erforschte die Wirkung eines auf das Geschlecht bezogenen Bias in einem Fall von Sorgerecht und in einem Fall von Diskriminierung im beruflichen Kontext. Richter:innen wurden durch ihre Geschlechtsideologie nachweislich beeinflusst, weil ihre Entscheidung anhand ihrer Einstellung zu Geschlechterrollen vorhergesagt werden konnte. Da Naive ähnlich entschieden, konnte juristische Expertise diesen Bias folglich nicht reduzieren. Doch auch andere am Justizwesen beteiligte Personengruppen neigen zu Heuristiken oder Urteilsverzerrungen (z. B. Bewährungshelfer:innen; Kautt, 2009). Sagana (2018) argumentiert, dass Richter:innen häufig Heuristiken nutzen, um die Komplexität der Fälle, mit denen sie sich befassen, zu reduzieren. Insbesondere Zeitdruck, Unsicherheit und Ambiguität sind dabei treibende Kräfte. Somit spielen Urteilsfehler im Entscheidungskontext des Strafrechts eine Rolle und können – bereits im Ermittlungsverfahren (s. 2.3.3) – eine gewisse Abwärtsspirale auslösen (Sagana, 2018; s. auch Maegherman, 2021).

Warum können Heuristiken und Bias im juristischen Kontext auftreten? Da die Art und Weise der Feststellung von Tatsachen und die Würdigung der Beweise aus der Hauptverhandlung nicht vorgegeben werden (s. 2.1.6), ist dieser (intuitive) Prozess

[27] Allerdings argumentieren Sagana und van Toor (2020), dass es aufgrund der strafprozessualen Vorgaben nicht ohne die Vorgabe eines Ankers geht, da in den Plädoyers Vorschläge für das Strafmaß zu nennen sind (s. 2.1.4).

fehleranfällig und kann Verzerrungen und Fehlurteile begünstigen (Lieber & Sens, 2019a; Rachlinski & Wistrich, 2017). Pantazi et al. (2020) argumentieren, dass Heuristiken und Bias, die auf der Grundlage von präsentierten Informationen entstehen (z. B. Bestätigungsfehler), bedeutsame Wirkfaktoren in diesem Kontext darstellen, denn in ebendiesen gerichtlichen Verhandlungen sollen Urteile und Entscheidungen eigentlich gemäß der gesetzlichen Theorie auf reiner Informationsgrundlage getroffen werden (s. auch Bieneck, 2006). Sie umschreiben dieses Phänomen als meta-kognitive Kurzsichtigkeit. Damit ist laut Pantazi et al. (2020) die Unfähigkeit verbunden, vorliegende Informationen hinsichtlich ihrer Qualität und ihrer Herkunft beurteilen zu können. Dies erinnert an Kahnemans (2011) Konzept des WYSIATI und an die mit der polizeilichen Akte als Entscheidungsgrundlage verbundenen Risiken (s. 2.3.3.1). Die Tatsache, dass auch juristische Fachpersonen Tendenzen zu gewissen Urteilsfehlern zeigen, wird an dieser Stelle als Hinweis auf das Auftreten von Typ-1- und Typ-2-Prozessen gedeutet, was einen Transfer dieser Dual-Prozess-Annahmen auf das juristische Setting untermauert.[28] Auf weitere Hinweise intuitiven Verhaltens von Richter:innen wird an späterer Stelle im Zusammenhang mit kognitiver Reflexion eingegangen (s. 2.3.4.3).

Auch wenn Unterschiede zwischen den Entscheidungen von Rechtsexpert:innen laut Maguire (2010) zu erwarten sind, so birgt die Tatsache, dass man in einem Bereich besonders ausgebildet ist, die Gefahr, zu Fehleinschätzungen verleitet zu werden und die eigenen Urteile und Entscheidungen zu überschätzen (Kahneman, 2011; s. 2.2.3.2). Erfahrene Personen sehen den Einfluss von Urteilsfehlern auf die eigenen Entscheidungen – der gemäß der obigen Ausführungen aber vorhanden zu sein scheint – als weniger stark an (Zapf et al., 2018). Obwohl sich Rechtsexpert:innen in einer Studie von Dhami und Ayton (2001) angesichts der gleichen Fallbeispiele in ihren Kautionsentscheidungen unterschieden, äußerten sie ein starkes Vertrauen in die Angemessenheit ihrer jeweiligen Entscheidung. Eine juristische Ausbildung und ein großer Erfahrungsschatz gehen mit größerem Vertrauen in die eigene Entscheidung einher (Dickert et al., 2012). Eine derartige Selbstüberschätzung

[28] Für weiterführende Ausführungen zu spezifischen Bias – Rückschaufehler, Rahmungseffekt, Ankereffekt, Bestätigungsfehler, Tunnelblick – im juristischen Kontext wird auf Effer-Uhe und Mohnert (2019), Englich et al. (2006), Findley und Scott (2006), Guthrie et al. (2001), Maegherman (2021), Nickolaus (2018), Orr und Guthrie (2006), Rachlinski (2000), Rachlinski und Wistrich (2017), Oswald und Wyler (2018) und Schweizer (2005, 2007, 2009) verwiesen.

kann Rechtsexpert:innen aber daran hindern, an eigentlich geeigneten Stellen zu zweifeln (Schweizer, 2005), besitzen sie doch das Vertrauen in sich und die Überzeugung von Bias unbeeinflusst zu sein (Landsman & Rakos, 1994).[29]

Allerdings stellt juristische Expertise bedingt einen Schutzfaktor dar: Mit Blick auf kulturell polarisierende Streitfragen (z. B. Legalisierung von Marihuana) wurden Richter:innen und Mitglieder der Rechtsanwaltschaft in ihren Antworten weniger von kulturellen Werten beeinflusst als Studierende und Laien (Kahan et al., 2016). Insbesondere bei berufserfahrenen Jurist:innen zeigen sich wiederkehrende Entscheidungssituationen in ihrer langjährigen Übungszeit. Solche Situationen fördern die Nutzung von Routinen, um ein sich wiederholendes Auseinandersetzen mit ähnlichen Inhalten und Abläufen zu vermeiden. Routinen sind ebenso ressourcenschonend wie intuitives Verhalten, bergen aber das Risiko für kognitive Verzerrungen. Verantwortung stellt einen wichtigen regulativen Faktor dar, der einen Effekt auf die Entscheidungsfindung haben kann: Richter:innen bemühen sich darum, den Gesetzen und Vorgaben zu entsprechen (Wistrich et al., 2015). Somit sind Expert:innen eine wesentliche Quelle für die Gesellschaft, besonders bei wichtigen Entscheidungen (Klein et al., 2017). Laut Schmittat und Burgmer (2020) erachten Laien die Richter:innen als moralische Expert:innen in einem Arbeitsfeld, in dem es keine eindeutigen Entscheidungsstandards gibt. Schmittat und Englich (2016) untersuchten, inwiefern Rechtsexpert:innen vom Entscheidungsbias betroffen sind. Auch wenn Expert:innen ihre Entscheidung unterstützende Informationen positiver bewerteten als widersprechende Angaben, zeigte sich ein Gruppenunterschied: Domänenspezifische Expert:innen zeigten diese Tendenz weniger als allgemeine Rechtsexpert:innen, die sich wiederum nicht von Naiven unterschieden. Wurde ein Gefühl von Verantwortung induziert, näherte sich diese Tendenz der allgemeinen Fachpersonen zur bestätigenden Informationssuche der Tendenz der domänenspezifischen Personen an. Verantwortung und Domänenspezifität können demnach protektive Faktoren gegen mögliche Bias darstellen.

[29] Der *bias blind spot* kann im Deutschen als Verzerrungsblindheit bezeichnet werden. Der Begriff umschreibt die menschliche Annahme, man selbst werde beim Urteilen und Entscheiden weniger durch kognitive Verzerrungen und Bias beeinflusst als andere Personen. Für weiterführende Literatur s. Ehrlinger et al. (2005), Pronin et al. (2002) und Pronin et al. (2004).

Maßnahmen zur Reduktion von Bias setzen bei der Falsifikation an. Die Aufgabe ist es, sich mit möglichen Gegenbeweisen und Alternativhypothesen für vorliegende Sachverhalte zu beschäftigen. Dies kann beispielsweise dem Bestätigungsfehler entgegenwirken, wird aber in der juristischen Praxis von Richter:innen wenig angewendet: „However, the open questions revealed that judges may be excessively focused on excluding alternative scenarios" (Maegherman, 2021, S. 85). Dieser Ansatz der Falsifikation kann nicht nur auf juristische Fachpersonen angewendet werden, sondern auch auf andere am Ermittlungsprozess Beteiligte.[30] Eine wichtige Forderung ist in dem Zusammenhang, dass die Fachpersonen der Strafverfolgung verstärkt Einsicht in die Ermittlungsarbeit nehmen sollen, um die dort generierte Informations- und Beweislage besser evaluieren zu können (Findley & Scott, 2006). Zumal es im Ermittlungsverfahren deren Aufgabe ist, nicht nur be-, sondern auch entlastende Beweise zu sichern (§ 160 Abs. 2 StPO; s. 2.1.6).

2.3.4.2 Expertise im juristischen Kontext

Jurist:innen haben eine langjährige akademische Ausbildung hinter sich, was ohne gewisse kognitive Stärken kaum möglich wäre (für eine Übersicht der Ausbildung s. Glöckner et al., 2013). Dennoch schützt kognitive Leistungsfähigkeit nicht uneingeschränkt vor Bias, da sie nur im begrenzten Zusammenhang mit der kognitiven Reflexionsfähigkeit steht (Stanovich & West, 2008, 2014; s. 2.2.3.3). Die langjährige Ausbildung spricht aber dafür, dass diese Personen als besondere Expert:innen für ihr Arbeitsfeld bezeichnet werden können. Dickert et al. (2012) argumentieren, dass die Kongruenz zwischen Urteilen steigt, wenn sich die Ausbildungen der urteilenden Menschen ähneln und dadurch die Überschneidungen der Wissensstrukturen größer werden. In ihrer Studie zeigte sich, dass die Urteile von Fachpersonen eher mit tatsächlichen Gerichtsentscheidungen übereinstimmten als die von Studierenden oder Laienrichter:innen. Wichtig für die Entwicklung von Expertise ist neben einer Regelmäßigkeit auch eine langjährige Übungszeit (Kahneman, 2011; s. 2.2.3.2).

[30] Da eine Ausführung solcher Maßnahmen für die vorliegende Studie nicht weiter von Relevanz ist, wird auf die Literatur von Findley und Scott (2006), Kassin et al. (2013), Lidén et al. (2019), Maegherman (2021) und Schmittat (2022) verwiesen.

Doch wie sieht es auf der Ebene der juristischen Noviz:innen aus? Mitchell (1989) argumentiert, dass sich Rechtsexpert:innen und -noviz:innen dahingehend unterscheiden, dass Expert:innen durch ihr Wissen und ihre Erfahrungen bereits kognitive Schemata entwickelt haben. Dadurch „sehen" Fachpersonen Verknüpfungen und Verbindungen, die nicht explizit in juristischen Sachverhalten dargestellt sind. Zudem nutzen Rechtsexpert:innen ein eigenes Fachvokabular, wodurch deren Verständigung und Denken anders erfolgen kann als bei Noviz:innen (Mitchell, 1989). Unterschiede zwischen gelernten und ungelernten Personen in der Entscheidungsfindung basieren größtenteils auf den Wissensstrukturen, die durch die juristische Ausbildung angeeignet werden (Dickert et al., 2012). Nievelstein et al. (2010) verglichen angehende Jurist:innen (neue vs. fortgeschrittene Studierende) und Fachpersonen in der Bearbeitung komplexer Fälle. Die Besonderheit war, dass Teilnehmende auf ihr konzeptuelles Wissen zurückgreifen mussten, da sie keine unterstützenden Quellen erhielten (Gesetzestext). Fachpersonen schnitten dabei besser ab als Studierende. Studierende im ersten Jahr, die noch kein konzeptuelles Wissen aufbauen konnten, profitierten im Vergleich zu den Fortgeschrittenen im dritten Jahr nicht von Fachbüchern. Allerdings war die Leistung der Expert:innen ebenfalls schwach: Auch Fachpersonen scheinen fachliche Hilfsmittel in der Ausübung ihrer Arbeit zu benötigen. Somit ist laut Nievelstein et al. (2010) das Lernen, mit solchen Mitteln umzugehen, als Bestandteil juristischer Ausbildung nicht zu unterschätzen. Glöckner et al. (2013) untersuchten auf Grundlage einer umfassenden Stichprobe von Jurastudierenden deren Leistungsentwicklung beim Lösen komplexer Fälle über ein Jahr hinweg. Studierende wurden anhand ihrer Leistung unterteilt, um den Einfluss von Übungsklausuren zu ermitteln. Diejenigen mit hoher Anfangsleistung zeigten eine lineare Lernkurve und profitierten mehr von auf bestimmte Themenbereiche bezogenen Klausuren, wohingegen diejenigen mit niedriger Leistung von eher allgemein gehaltenen Klausuren profitierten (konkave Lernkurve). Glöckner et al. (2013) schlussfolgern, dass das Schreiben von Übungsklausuren innerhalb der juristischen Ausbildung ratsam ist. Angehende Jurist:innen unterscheiden sich also darin, wie sie ihre Leistung steigern können und wie sie die Entwicklung von Expertise – zumindest gemessen an Klausurnoten – fördern können. Bieneck (2006) verglich Studierende der Rechtswissenschaft sowie Rechtsreferendar:innen hinsichtlich ihrer Beeinflussbarkeit durch den Ankereffekt. Es zeigte sich ein Einfluss von Expertise dahingehend, als dass sich Studierende in der Bestimmung eines

Strafmaßes mehr durch diesen Effekt leiten ließen. Im Sinne der relativen Herangehensweise zur Untersuchung von Expertise lohnt demnach nicht nur ein Vergleich von juristischen Expert:innen und Nicht-Expert:innen, sondern auch von Noviz:innen, weil sich auch innerhalb dieser „mittleren" Gruppe Unterschiede zeigen (Chi, 2006; s. 2.2.3.2).

2.3.4.3 Kognitive Reflexion im juristischen Kontext

Neben Expertise werden zwei weitere personenbezogene, extralegale Einflussfaktoren berücksichtigt: kognitive Reflexion und Need for Cognition (s. 2.3.4.4). Niedrige Reflexionsausprägungen stehen im Zusammenhang mit intuitivem Verhalten (Typ-1-Prozesse), wohingegen hohe Ausprägungen auf reflektiertes Vorgehen schließen lassen (Typ-2-Prozesse; s. 2.2.3.3). Die Tatsache, dass Rechtsexpert:innen – wie in Abschnitt 2.3.4.1 beschrieben – zumindest in einigen Situationen durchaus auf heuristische, intuitive Prozesse zurückgreifen, lässt die Vermutung zu, dass diese Stichprobe auch unterschiedlich schwach oder stark ausgeprägte Fähigkeiten der kognitiven Reflexion aufweist. Zumal diese Fähigkeit zur Reflexion nicht mit Intelligenz gleichzusetzen ist, welche wiederum für die akademische Ausbildung eine wichtige Voraussetzung ist. Guthrie et al. (2007) präsentierten 252 Richter:innen die Items des CRT (s. 3.4.4). Bei einem Maximalwert von 3 Punkten zeigte der Wert von $M = 1.23$ an, dass diese Fachpersonen in ihrer Leistung mit anderen gebildeten Erwachsenen vergleichbar waren, wenngleich rund ein Drittel keine der drei Aufgaben richtig beantworten konnte. Falsche Antworten entsprachen in der Regel den intuitiven, aber falschen Optionen. Die Anzahl derer, die die Fragen korrekt beantworteten, stieg im Laufe des Tests von rund 28% (Item 1) auf etwa 50% (Item 3) an (Item 2: 44%), was auf den Lerneffekt schließen ließ, dass sich sukzessiv weniger auf intuitive Reaktionen verlassen wurde. Demnach war zwar die Tendenz zu intuitivem Verhalten erkennbar, ließ sich aber reduzieren. Laut Guthrie et al. (2007) bedeutet dies aber nicht, dass Richter:innen auch in ihrer beruflichen Funktion zu intuitivem Verhalten neigen, auch wenn der CRT dies zumindest für solch einfache, gewöhnliche Aufgaben andeutet. Dazu passt, dass die Verzerrungsblindheit – die Annahme man selbst neige weniger zu Urteilsfehlern als Mitmenschen – negativ mit kognitiver Reflexion korreliert: Eine subjektiv als gering eingeschätzte Beeinflussbarkeit durch Bias geht mit einem größeren Verlass auf Intuition einher (Scopelliti et al., 2015). Dass es aber durchaus zu Urteilsfehlern

kommen kann, die auf Intuition und Typ-1-Prozessen basieren, wurde im Abschnitt 2.3.4.1 zu Heuristiken und Bias bereits ausgeführt und an dieser Stelle durch die Darlegung interindividueller Reflexionsfähigkeiten untermauert (s. auch Rachlinski & Wistrich, 2017).

2.3.4.4 Need for Cognition im juristischen Kontext

Auch wenn es inhaltliche Überschneidungen gibt, so stellen die kognitive Reflexion und Need for Cognition zwei verschiedene Konstrukte dar (s. 2.2.3.4). Die Übertragung der Annahmen von NFC (auf Entscheidungen von Geschworenen) lässt erwarten, dass „jurors who are high in NFC are more likely to reason through complex legal tests rather than engage in reasoning based on arguably less cognitively demanding, commonsense principles" (McKay et al., 2014, S. 497; s. auch Leippe et al., 2004). Unterschiede in den Ausprägungen von NFC können mit Bezug zur Beweiswürdigung und zu den eigentlichen Urteilen betrachtet werden (s. auch Wood et al., 2019). Shestowsky und Horowitz (2004) untersuchten den Einfluss des Kognitionsbedürfnisses auf die Beratungen von Schein-Geschworenen, an deren Ende das Urteil entsteht. Diejenigen mit hoher Denkfreude sprachen signifikant länger und wurden als überzeugender und aktiver wahrgenommen, auch wenn sie nicht zwingend validere oder logischere Argumente darboten. Diejenigen mit niedriger Ausprägung konnten dagegen besser zwischen starken und schwachen Argumenten differenzieren. Des Weiteren fokussierte sich diese Gruppe eher auf das Erreichen eines Konsenses, wohingegen Menschen mit hoher Ausprägung dazu tendierten, überzeugen zu wollen. Henderson und Levett (2020) gingen in ihrer Studie der Frage nach, wie Menschen mit hoher Denkfreude den Zusammenhang zwischen der Aussage einer Person und (widersprüchlichen) Beweisen wahrnehmen. Generell stieg die Wahrscheinlichkeit für eine Verurteilung, wenn die Aussage mit der Beweislage übereinstimmte. Hohe Ausprägungen gingen mit einer weniger wahrscheinlichen Verurteilung einher, auch wenn dabei nicht besser zwischen den unterschiedlichen Beweislagen differenziert werden konnte. Leippe et al. (2004) beschreiben dagegen ein kurvilineares Modell: Schein-Geschworene mit mittlerer Ausprägung neigten in dieser Studie bei einer starken Beweislage eher zur Verurteilung als diejenigen mit hoher oder niedriger Motivation. Bei besonders stark ausgeprägter Denkfreude war die Wahrscheinlichkeit, die Person schuldig zu sprechen, am geringsten – vermutlich, weil auch bei einer starken Beweislage noch Zweifel

an der Schuld bestanden. Betrachtet man eine fachliche Stichprobe von Studierenden der Rechtswissenschaft sowie Rechtsreferendar:innen, waren diejenigen mit niedrigem Kognitionsbedürfnis anfälliger für den Ankereffekt (Bieneck, 2006). Allerdings werden Menschen mit hoher Ausprägung mitunter stärker von der Glaubwürdigkeit der Verteidigung beeinflusst als Menschen mit niedrigen NFC-Werten (Wood et al., 2019). Wood et al. (2019) deuten das Ergebnis ihrer Studie derart, dass die Teilnehmenden die besagte Glaubwürdigkeit als (starkes) Beweismittel nutzten und in ihre Beweiswürdigung integrierten. Auch wenn sich diese Befunde insbesondere auf Schein-Geschworene beziehen, lässt sich dennoch zusammenfassen: „Research in psychology and law has shown that differences in the amount and depth of thinking between individuals high and low in N[F]C can influence legal judgments“ (Petty et al., 2009, S. 325). Bieneck (2006) agumentiert, dass sich ein hohes Kognitionsbedürfnis im juristischen Kontext darin äußert, dass das Bearbeiten von Fällen Freude bereitet. Damit würde einhergehen, dass sich Jurist:innen mit hohen NFC-Ausprägungen – trotz nicht eindeutiger Beweislage – mehr auf die vorliegenden (Fall-)Daten konzentrieren als diejenigen, denen diese Aufgabe weniger Freude bereitet.

2.3.5 Die Verbindung von Prozess und Person im strafrechtlichen Ermittlungsverfahrens

Es wurden juristische und psychologische Inhalte dargestellt und mit Blick auf den Strafprozess und auf daran Beteiligte miteinander verknüpft (s. 2.1.7; 2.2.4; 2.3.3; 2.3.4). An dieser Stelle erfolgt eine für die Studie wesentliche Fokussierung, denn größtenteils betrafen die betrachteten prozess- und personenbedingten Einflussfaktoren keine Entscheidungssituationen im Ermittlungs-, sondern im Hauptverfahren (Schuldfrage, Strafmaß). An anderer Stelle wurde bereits betont, dass das Ermittlungsverfahren aufgrund seiner zeitlichen Position sowie seiner inhaltlichen Funktion hochrelevant für das gesamte Strafverfahren ist (s. 2.1.7): "A decision made early in the process … has the potential to impact the judicial outcome of the case” (Morgan et al., 2018, S. 409). Erfolgt eine Anklage am Ende des Ermittlungsverfahrens, ist dies ein starker Prädiktor für die spätere Verurteilung (Schmittat et al., 2022). Mit Blick auf gerichtliche Diskrepanzen führen Ellison und Brennan (2016) weiter aus: “Likewise, more would be learned about sentencing disparity if researchers paid more attention to upstream outcomes, especially the pre-conviction

decisions made by prosecutors" (S. 340). Jurisdiktionen unterscheiden sich darin, zu welchem Zeitpunkt eine Art Staatsanwaltschaft die Strafverfolgung übernimmt und welche Aufgaben es zu erfüllen gilt (Lidén et al., 2019). Im Gegensatz zum deutschen Strafverfolgungszwang (s. 2.1.2) besitzt die Strafverfolgung in den USA mehr Freiheitsgrade. Entscheidungen über das Veranlassen von Strafverfolgungen in Fällen sexueller Gewalt werden von extralegalen Faktoren beeinflusst (z. B. Charakteristiken der beschuldigten oder der geschädigten Person; Alderden & Ullman, 2012), wobei sich in der Handhabung dessen regionale Unterschiede zwischen Städten eines Landes ergeben können (Holleran et al., 2010).

Lidén et al. (2019) verglichen schwedische Jura- und Psychologiestudierende sowie Mitarbeitende der Strafverfolgung in acht Fallszenarien, in denen sie entweder selbst über die Verhaftung der tatverdächtigen Person entschieden oder über die Entscheidung anderer Mitarbeitenden informiert wurden. Die Expert:innen zeigten abwägendes Verhalten in der Betrachtung der Beweise. Sobald sich aber – insbesondere in Fällen, in denen eine Verhaftung stattgefunden hat – für eine Anklage entschieden wurde, veränderte sich die Denkweise in eine „schuldbestätigende" Richtung (s. auch Engel & Glöckner, 2013; Findley & Scott, 2006). Studierende beider Fachrichtungen zeigten bereits früher als Fachpersonen Tendenzen zum Bestätigungsfehler, wonach das juristische Studium nicht übermäßig vor einem solchen Bias schützte. Sommers et al. (2014) untersuchten den Einfluss extralegaler Faktoren auf die Entscheidung der Strafverfolgung im Rahmen von Gewaltverbrechen. Sie konnten zeigen, dass relevante legale Faktoren richtungsweisend für solche Entscheidungen sind (z. B. Delikttyp). Dennoch wurden (geringe) Einflüsse extralegaler Variablen gemessen (z. B. Geschlecht und Herkunft der beschuldigten Person oder das Alter der geschädigten Person). Egli Anthonioz et al. (2019) argumentieren, dass bereits die Zuordnung der Rolle als Richter:in, Verteidiger:in oder Vertreter:in der Staatsanwaltschaft Auswirkungen auf die Informationssuche sowie die Beweiswürdigung hat: „Since the assignment of the role of … prosecutor induces a motivation to prefer the hypothesis that the defendant is guilty, the assessment of the evidence and the search for new evidence will likely be biased, unbeknownst to the subject" (S. 454). Dies zeigte sich auch in der Studie von Egli Anthonioz et al. (2019) mit Kriminologie- und Strafrecht-Studierenden: Sobald sich für eine Untersuchung des Falles entschieden wurde, wirkte sich die eingenommene

Rolle erwartungsgemäß auf die Verurteilungen aus, da sich diejenigen in der Rolle der Verteidigung am wenigsten und diejenigen mit der Aufgabe der Strafverfolgung am meisten für Schuldsprüche aussprachen. Auf deskriptiver Ebene zeigten sich die Differenzen auch bei der Handhabung be- und entlastender Beweise, da Verteidiger:innen sich weniger auf belastende Hinweise stützten als diejenigen, die sich in der Funktion der Strafverfolgung wiederfanden. Engel und Glöckner (2013) beschreiben es als schwierig, eine (durch die Rolle induzierte) Interpretation zu verlassen, sobald man in ihr gefangen ist – auch wenn, wie in ihrer Studie, die Rollenzuteilungen aufgehoben und monetäre Anreize für das Treffen einer unbeeinflussten Entscheidung in Aussicht gestellt wurden. In besagter Studie mit einer studentischen Stichprobe machte die zugeteilte Rolle in der Vorhersage des Urteils rund 14% aus. Allerdings konnte Zeit zum Nachdenken diesen Bias reduzieren. Schmittat et al. (2022) untersuchten die Wirkung einer Alternativgeschichte der Verteidigung auf die Schuldfrage. Erhielten Jurastudierende zusammen mit den Ermittlungsergebnissen eine alternative Geschichte, reduzierte sich die Wahrscheinlichkeit einer Anklage vermutlich, weil die Plausibilität der polizeilichen Geschichte reduziert wurde.

Empirische Studien zeigen somit, dass die Rolle als Vertreter:in der Staatsanwaltschaft das Vorgehen im Strafverfahren beeinflussen kann (z. B. die Handhabung der Beweise). Die Frage, ob eine Anklage zu erheben ist, wird maßgeblich durch die polizeilichen Ausführungen bestimmt (Schmittat, 2022). In Abschnitt 2.3.3.1 wurde bereits die Relevanz der polizeilichen Akte als Entscheidungsgrundlage im Ermittlungsverfahren abgeleitet. Schmittat et al. (2022) fassen zusammen:

> The decision to charge is based on potentially incomplete or biased files and shapes the course of the following trial, as this decision – the first notification the court receives – may cause the judge(s)/ jury to cling to the hypotheses of the investigative authorities (S. 24)

Somit stellt nicht nur die Aktenlage, sondern auch die als Mitglied der Staatsanwaltschaft eingenommene Rolle und die damit einhergehenden Aufgaben einen Wirkfaktor auf die Entscheidungsfindung dar. Findley und Scott (2006) argumentieren weiterführend: „But prosecutors‘ assessments of guilt can be flawed both by the information provided to them and the feedback they receive. Prosecutors are particularly vulnerable to distortions based on the types of information to which they have access" (S. 329). Demzufolge gilt es, die Entscheidungsprozesse am Ende

des Ermittlungsverfahrens zu untersuchen (s. auch Lidén et al., 2019; Schmittat et al., 2022). Da die Unschuldsvermutung wider Intuition arbeitet (s. 2.3.3.4), liegt die Vermutung nahe, dass am Ende des Ermittlungsverfahrens nach der Vorlage einer Fallakte durch die Polizei zunächst ein intuitives, den Bestätigungsfehler begünstigendes „Schuldig"-Denken stattfindet. Dies könnte dazu führen, dass insbesondere belastende Beweismittel genauer betrachtet und zu einer kohärenten Geschichte konstruiert werden (*Story Model*; s. 2.1.6.2) – möglicherweise befeuert durch prozess- und personenbedingte Wirkfaktoren wie Zeitdruck, kognitive Reflexion oder Need for Cognition. Andererseits stellt dieser Zeitpunkt im Ermittlungsverfahren womöglich auch den Höhepunkt der Neutralität oder Objektivität dar (Lidén et al., 2019; s. auch Büchner, 2022), sodass sich ein die beschuldigte Person belastendes Denken erst im Anschluss einstellt. Wie handhaben Personen in komplexen Situationen eine nicht eindeutige Beweislage, auf Grundlage derer die Frage nach dem Tatverdacht (und nach dem nächsten Verfahrensschritt) beantwortet werden muss? Was beeinflusst diese Handhabung mit Blick auf den Prozess und die Person? Die Beantwortung dieser Fragen ist das Ziel der Studie.

2.4 Überblick über die Studie

Urteile können dahingehend untersucht werden, ob eine Person wiederholt die gleiche Entscheidung trifft oder ob verschiedene Personen im gleichen Fall zum gleichen Ergebnis kommen (Dhami & Ayton, 2001). In der vorliegenden Studie wurde letzteres betrachtet. Es galt zu untersuchen, welche Faktoren sich auf die Entscheidung auswirken (prozess- oder personenbedingt). Die vorrangig interessierende Entscheidung ist zeitlich am Ende des Ermittlungsverfahrens zu verorten, wenn sich die Staatsanwaltschaft anhand der (polizeilichen) Aktenlage für oder gegen das Vorliegen eines Tatverdachtes ausspricht. Aus dieser Entscheidung ergeben sich wiederum prozessweisende Handlungsschritte (s. 2.1.2). In Abbildung 2.2 sind die Variablen und deren inhaltliche Einordnung veranschaulicht.

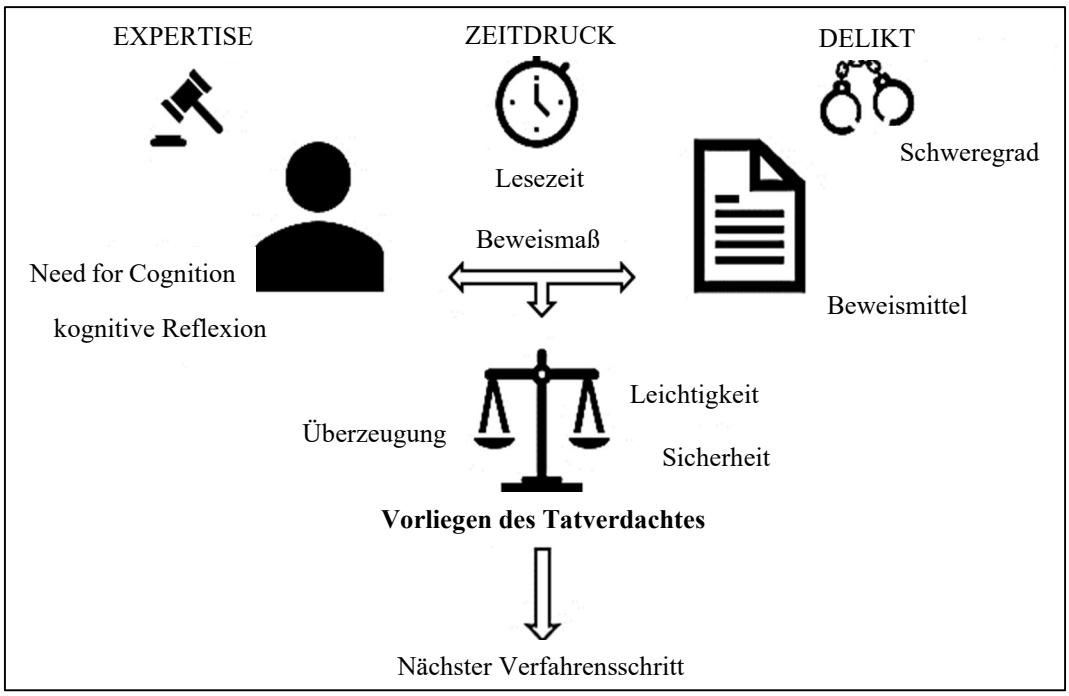

Abbildung 2.2. Veranschaulichung der untersuchten Variablen mit Microsoft-365-Piktogrammen. *Anmerkungen.* Die menschliche Figur stellt die an der Studie teilnehmende Person dar. Das stilisierte Schreiben repräsentiert die polizeiliche Akte als Versuchsmaterial (Vignette).

Da die Frage nach dem Vorliegen des Tatverdachtes aus der Perspektive der Staatsanwaltschaft chronologisch *vor* der definierten Entscheidung über den konkreten nächsten Verfahrensschritt stattfindet (im Sinne einer „wenn, dann"-Situation), steht in dieser Studie ebenjene Einschätzung des Tatverdachtes im Fokus. Diese Einschätzung lässt sich anhand einer Dichotomie ermitteln („ja, liegt vor", „nein, liegt nicht vor"; abhängige Variable). Liegt er nicht vor, ist eine Einstellung nötig. Liegt der Tatverdacht vor, ermöglichen sich die Verfahrensschritte der Anklage, der Einstellung wegen Geringfügigkeit und der Einstellung unter Weisungen und Auflagen (abhängige Variable: nächster Verfahrensschritt). Das Beweismaß (abhängige Variable) stellt in der Praxis keinen eindeutigen Punktwert dar, ab dessen Erreichen eine Anklage erhoben wird oder nicht, wenngleich es aber die Aufgabe der Staatsanwaltschaft ist, die Beweislage in eine solche dichotome Entscheidungen zu übersetzen (Lidén et al., 2019). Es wurden nicht nur juristische Fachpersonen einbezogen, sondern auch Noviz:innen (Studierende und Referendar:innen) und Laien (unabhängige Variable: Expertise; Chi, 2006; Shanteau, 1988). Dies ermöglicht direkte Vergleiche zwischen den Gruppen, z. B. hinsichtlich der Interpretation darüber, wie einig oder uneinig beziehungsweise ähnlich oder unähnlich sich die Teilstichproben in gewissen Aspekten sind (s. 2.2.3.2). Auch wenn sich die Studie

inhaltlich um die Perspektive der Staatsanwaltschaft dreht, so wurde eine gemischte juristische Stichprobe untersucht. Befunde zu anderen juristischen Berufsgruppen (z. B. Richter:innen, Vertreter:innen der Rechtsanwaltschaft) lassen sich auf diese spezielle Teilstichprobe übertragen, da die Fachpersonen zumindest von der Grundausbildung her vergleichbar sind (s. auch Glöckner et al., 2013). In diesem Quasi-Experiment erhielten Studienteilnehmende eine Vignette zu einem Fall von Diebstahl oder einem Fall von Körperverletzung (randomisierte unabhängige Variable: Delikt). Diese Vignetten enthielten jeweils die Zusammenfassung der polizeilichen Ermittlungen, in der vier Kategorien von Beweismitteln abgedeckt waren (s. auch Lidén et al., 2019). Die gesamte Beweislage sprach nicht eindeutig für oder gegen die Täterschaft der beschuldigten Person. Um zu vermeiden, dass Unterschiede in der Schwere des Delikts zu ungewollten Veränderungen im Strafmaß führen (Lundberg, 2016), wurde sich für diese beiden Delikte entschieden, da sie zumindest gemäß des StGB vergleichbar sind (abhängige Variable: Schweregrad des Delikts). Für Einblicke in den Entscheidungsprozess wurden Teilnehmende zu ihrem Entscheidungserleben (abhängige Variablen: Leichtigkeit, Sicherheit, Überzeugung), zum angewandten Beweismaß sowie zum Umgang mit den Beweismitteln befragt (z. B. genutzte Anzahl oder Relevanz einzelner Beweismittel; jeweils abhängige Variable). Auch qualitative Angaben zu gewünschten Nachermittlungen wurden erhoben.

Da die Staatsanwaltschaft – nicht die Polizei – in der Abschlussverfügung mit ihrer Entscheidung das Ermittlungsverfahren beendet, ist laut prozessualen Vorgaben deren Perspektive an dieser Stelle richtungsweisend. Empirische Erkenntnisse zeigen aber, dass sich das prozessuale Vorgehen an sich sowie personenbezogene Einflussfaktoren darauf auswirken können. Die Annahmen von Dual-Prozess-Theorien lassen sich auf das juristische Setting übertragen (s. 2.3.2). Daher wurden im Zusammenhang mit diesen Annahmen Wirkfaktoren ausgewählt, die bisher unzureichend als solche im juristischen Kontext untersucht worden sind. Als personenbezogene abhängige Variablen und Prädiktoren werden die kognitive Reflexion und Need for Cognition berücksichtigt. Zur Berücksichtigung einer gewissen Alltagsnähe findet für die Beweiswürdigung sowie für die sich daran anschließende Tatverdacht-Frage eine Zeitdruckmanipulation statt, um dadurch eine heuristische oder reflektierte

Vorgehensweise induzieren und diese Experimentalgruppen dahingehend vergleichen zu können (randomisierte unabhängige Variable: Zeitdruck; Maule & Hockey, 1993; Rice & Trafimow, 2012).

2.5 Hypothesen und Fragestellungen

In dieser Studie werden nicht nur konfirmatorische, sondern auch explorative Analysen durchgeführt. Erstere werden in Form von Hypothesen dargestellt, letztere werden als Fragestellungen formuliert.

Hypothese H1:

Es zeigt sich ein Haupteffekt von Expertise auf die Entscheidung hinsichtlich des Tatverdachts in Interaktion mit Zeitdruck, unabhängig vom Delikttyp: Laien bejahen den Tatverdacht eher als Noviz:innen und Expert:innen, insbesondere unter Zeitdruck.

Hypothese H2:

Es zeigt sich ein Haupteffekt von Expertise auf das Beweismaß, unabhängig von Zeitdruck und Delikttyp: Die Beweislage überzeugt Laien mehr als Noviz:innen und Expert:innen.

Hypothese H3:

Wird der Tatverdacht bejaht, zeigt sich ein Haupteffekt von Expertise auf die Entscheidung hinsichtlich des nächsten Verfahrensschritts in Interaktion mit Zeitdruck, unabhängig vom Delikttyp: Laien entscheiden sich eher für eine Anklage als Noviz:innen und Expert:innen, insbesondere unter Zeitdruck.

Hypothese H4:

Es zeigt sich ein Haupteffekt von Expertise auf den eingeschätzten Schweregrad der Delikte, unabhängig von Zeitdruck: Laien stufen den Schweregrad für das Delikt „Körperverletzung“ höher ein als den Schweregrad für das Delikt „Diebstahl“, wohingegen Noviz:innen und Expert:innen diese Tendenz nicht zeigen.

Fragestellung F1:

Hinsichtlich des wahrgenommenen Entscheidungsprozesses: Unterscheiden sich Laien, Noviz:innen und Expert:innen in den Prozessmerkmalen Leichtigkeit, Sicherheit und Überzeugung?

Fragestellung F2:

Unterscheiden sich Laien, Noviz:innen und Expert:innen in ihrer Fähigkeit zur kognitiven Reflexion und in ihrem Need for Cognition?

Fragestellung F3:

Sagen die kognitive Reflexion, Need for Cognition, die Expertise, der Zeitdruck, der Delikttyp oder die genutzte Lesezeit die Entscheidung hinsichtlich des Tatverdachtes vorher?

Fragestellung F4:

Sagen die kognitive Reflexion, Need for Cognition, die Expertise, der Zeitdruck, der Delikttyp oder die genutzte Lesezeit die Anzahl der Beweismittel vorher, die zur Entscheidung hinsichtlich des Tatverdachtes herangezogen wird?

Fragestellungen F5:

Unterscheiden sich Laien, Noviz:innen und Expert:innen darin, wie sie die Relevanz einzelner Beweismittel einschätzen?

Fragestellung F6:

Sagen die Beweismittel mit der höchsten und niedrigsten Relevanz die Entscheidung von Laien, Noviz:innen und Expert:innen hinsichtlich des Tatverdachts vorher?

3 Methode

Die Beschreibung des methodischen Vorgehens beginnt mit der Darstellung der Planung (s. 3.1) und des Ablaufs des Versuchs (s. 3.2). Informationen über eine durchgeführte Poweranalyse sowie über die Rekrutierung und Zusammensetzung der Stichprobe folgen im Anschluss (s. 3.3). Die eingesetzten Materialien (s. 3.4), deren Gütekriterien (s. 3.5) sowie die Ergebnisse einer zuvor durchlaufenen Pilotstudie zur Überprüfung der selbst erstellten Vignetten (s. 3.6) werden ebenfalls vorgestellt. Das Kapitel endet mit Übersichten der Maßnahmen zur Dateninspektion beziehungsweise zur Datenbereinigung (s. 3.7) sowie der durchgeführten statistischen Verfahren und den dafür notwendigen Voraussetzungen (s. 3.8).

3.1 Versuchsplanung

Das Projekt wurde am 05. Januar 2022 im Open Science Framework präregistriert (Ruppenthal, 2022). Diese querschnittliche Studie vereinte einen experimentellen und einen korrelativen Ansatz (Peters & Dörfler, 2019a), da sowohl experimentelle Gruppenzuordnungen als auch die Erfassung von Variablen mittels Fragebogen zum Einsatz kamen. Sie enthielt prüfende und explorative Elemente (s. 2.5).

Der Studie lag ein faktorielles 2 (Zeitdruck: ohne vs. mit) x 2 (Delikt: Diebstahl vs. Körperverletzung) x 3 (Expertise: Expert:in vs. Noviz:in vs. Laie)-Design zugrunde. Sämtliche Faktoren wurden zwischen den Versuchsteilnehmenden variiert (Between-subjects-Design). Die Zuweisung der Teilnehmenden zu den jeweiligen Untergruppen der Faktoren *Zeitdruck* und *Delikt* erfolgte randomisiert unter der Restriktion gleicher Gruppengrößen. Derartige Randomisierungsprozesse in der Zuordnung zu Experimentalgruppen verhindern das Auftreten von Verzerrungen und ermöglichen die Kontrolle personenbezogener Störvariablen (McCready, 2006; Peters & Dörfler, 2019a; Trimmel, 2009). Der Faktor *Expertise* ließ keine Randomisierung zu und wurde durch Selbstselektion realisiert, weswegen es sich um ein Quasi-Experiment handelt (Peters & Dörfler, 2019a; Sonnentag, 2006; Trimmel, 2009). Es stehen mehrere abhängige (AV) und unabhängige (UV) Variablen bezie-

hungsweise Prädiktoren im Fokus der Untersuchung (mehrfaktorielles Quasi-Experiment; Peters & Dörfler, 2019a). Die Variablen *Expertise*, *Zeitdruck* und *Delikt* besitzen die größte Relevanz und wurden daher in nahezu jeder Hypothese oder Fragestellung berücksichtigt (s. 2.5; s. auch Abbildung 2.2). Zudem wurden die UVs beziehungsweise Prädiktoren *kognitive Reflektion*, *Need for Cognition*, *Lesezeit*, *Beweismittel mit niedrigster* und *höchster Relevanz* erhoben. Neben den entscheidungsbezogenen AVs *Tatverdacht* und *nächster Verfahrensschritt* wurden die AVs *Beweismaß*, *Schweregrad des Delikts*, *Leichtigkeit* bzw. *Sicherheit* bzw. *Überzeugung* hinsichtlich der Tatverdachtsentscheidung sowie die *Anzahl der Beweismittel*, die für die Entscheidungsfindung benötigt wurden, erfasst. Die Variablen *kognitive Reflexion*, *Need for Cognition* und die *Beweismittel mit niedrigster* und *höchster Relevanz* wurden je nach Testverfahren auch als AVs eingesetzt.

3.2 Versuchsablauf

Die fragebogenbasierte Onlinestudie wurde mittels der Software *Qualtrics* realisiert. Zu Beginn erhielten die potenziellen Versuchsteilnehmenden auf der Startseite der Umfrage Informationen zum Ablauf der Studie. Dazu gehörten neben der thematischen Einordnung der Studie, deren Ablauf sowie einer erste Beschreibung der Aufgaben auch Angaben zu den Themen Anonymität, Freiwilligkeit und Datenschutz. Die Anonymität der Versuchsteilnehmenden konnte gewahrt werden, da die Antworten keine Rückschlüsse auf einzelne Personen zulassen. Es wurden keinerlei direkte personenbezogenen Daten erhoben (z. B. IP-Adresse des zur Teilnahme genutzten Endgeräts). Dies hat den Vorteil, dass von ehrlichen, da anonymen Antworten ausgegangen werden kann (Goddard & Villanova, 2006). Die Proband:innen wurden vorsorglich darauf hingewiesen, dass bestimmte Delikte in den Vignetten behandelt würden und bei Bedenken hinsichtlich dieser sensiblen Inhalte auf eine Teilnahme verzichtet werden solle. Es erfolgte zudem ein Hinweis auf externe Unterstützungsmöglichkeiten. Die Abbildung 3.1 stellt schematisch den Ablauf der Onlinebefragung dar, der in vier Blöcke unterteilt werden kann. Der erste Block repräsentiert das Vorgehen bei den Gruppenzuordnungen. Der zweite Block beinhaltet die Darbietung der Vignette sowie die Fragen, die sich auf das Gelesene beziehen. Im dritten Block kommen die Fragebögen für NFC und kognitive Reflexion zum Einsatz. Der vierte Block beinhaltet auf die jeweilige Expertise-Gruppe

abgestimmte demografische Fragen. Der zweite Block wurde bewusst vor den dritten positioniert, und nicht umgekehrt: Da sich die Haupthypothese H1 auf die Vignette bezieht (s. 2.5), sollte durch das frühe Präsentieren der dafür relevanten Variablen eine möglichst große Stichprobe sichergestellt werden (für nähere Angaben zu Dropout-Raten s. 3.3).

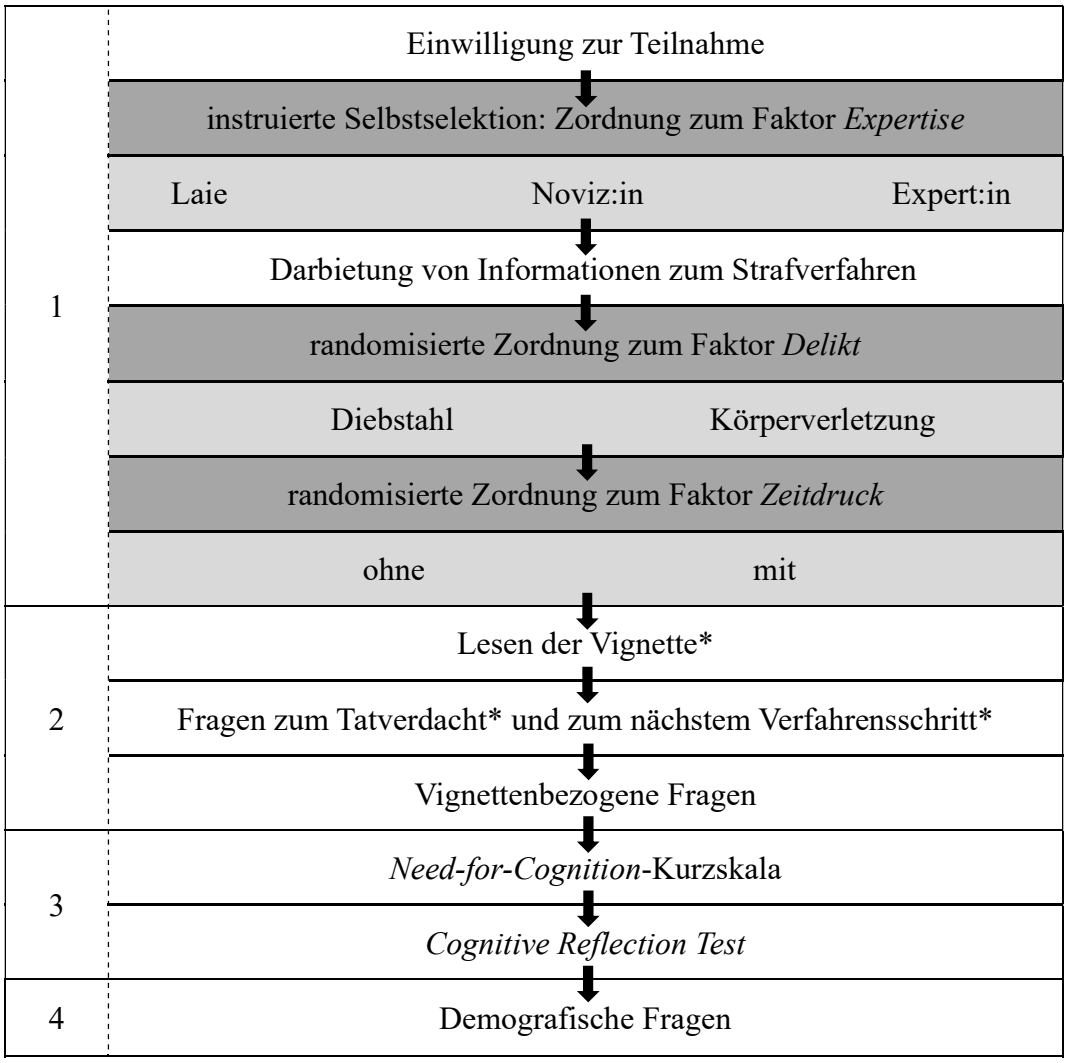

Abbildung 3.1. Schematischer Ablauf der Studie von der Einwilligung zur Teilnahme bis zur Angabe demografischer Daten.
Anmerkungen. Dunkelgrau unterlegte Zellen weisen auf den Zeitpunkt einer Gruppenzuordnung hin. Die jeweiligen Level der Faktoren sind hellgrau unterlegt. Mit Asterisk (*) markierte Aufgaben und Fragen entsprechen Variablen mit Zeitmessung.

Nachdem in die Teilnahme eingewilligt worden war, wurden die Versuchsteilnehmenden aufgefordert, sich einer der drei Expertise-Gruppen zuzuordnen. Das Kriterium für die Selbstselektion war die juristische Ausbildung. Wurde keinerlei juristische Ausbildung absolviert („Ich habe kein abgeschlossenes Studium im Fach

Rechtswissenschaft/Jura und befinde mich derzeit auch nicht in einem solchen Studium.“), wurde die Person als Laie eingestuft. Bei vorhandener juristischer Ausbildung wurde zwischen Noviz:innen („Ich befinde mich derzeit im Studium Rechtswissenschaft/Jura bzw. im Referendariat.“) und Expert:innen („Ich habe ein abgeschlossenes Studium im Fach Rechtswissenschaft/Jura und bin Voll-/Jurist:in. *oder* Ich habe die notwendige Ausbildung absolviert und bin Amtsanwältin/Amtsanwalt.“) unterschieden.[31]

Im Anschluss an die Selbstselektion folgten stark verkürzte und vereinfachte Informationen über das deutsche Straf- und Ermittlungsverfahren. Dabei wurde insbesondere auf die Rolle der Staatsanwaltschaft und auf die vier Kategorien von Indizien und Beweismitteln eingegangen (s. 2.1.6). Alle Expertise-Gruppen erhielten die gleichen Informationen. So konnte sichergestellt werden, dass Laien einen notwendigen Überblick über die für die Studie relevanten juristischen Inhalten bekamen. Dadurch wurde ebenfalls erreicht, dass den Noviz:innen und Expert:innen inhaltliche Grenzen aufgezeigt wurden, sodass diese sich auf relevantes Vorwissen beschränken konnten. Die Entscheidungen über das Vorliegen des Tatverdachtes und über die Wahl des nächsten Verfahrensschrittes sind von zentraler Bedeutung für die Untersuchung (s. 2.5). Daher folgten Informationen darüber, wann diese Entscheidungen im Strafverfahren getroffen werden und welche Optionen zur Verfügung stehen, sollte der Tatverdacht bejaht werden. Auch diese Formulierungen wurden nicht nach dem Grad der Expertise unterschieden, damit alle Gruppen die darauf aufbauenden Fragen auf Grundlage der gleichen Hintergrundinformationen bearbeiten konnten. Die möglichen Entscheidungsoptionen *Erhebung einer Anklage*, *Einstellung unter Auflagen* und *Einstellung wegen Geringfügigkeit* wurden mithilfe einer Grafik zusammengefasst und erklärt (s. A-1). Diese Grafik wurde zweimal unmittelbar hintereinander präsentiert, um möglichst sicherzugehen, dass die Versuchsteilnehmenden sich mit diesen für den Fortlauf der Untersuchung relevanten Inhalten ausreichend auseinandergesetzt haben (s. 2.1.7).

[31] Gemäß dem Gerichtsverfassungsgesetz sind die Mitglieder der Amtsanwaltschaft an den Amtsgerichten tätig (§ 142). Mit Blick auf Nordrhein-Westfalen befähigt die Ausbildung unter bestimmten Voraussetzungen zur Bearbeitung von Strafsachen an ebenjenen Instanzen. Dazu gehören auch die in dieser Studie betrachteten Delikte der Körperverletzung und des Diebstahls, weswegen diese Fachgruppe berücksichtigt wurde.

Es folgte die randomisierte Darbietung einer der beiden Vignetten. Die Versuchsteilnehmenden waren in der Einleitung zur Umfrage bereits informiert worden, dass sie einem der beiden Delikte zugelost werden. Sie erhielten zunächst ein Zitat aus dem StGB, in dem das jeweilige Delikt definiert sowie dessen Strafrahmen skizziert wurde (§ 223 Körperverletzung, § 242 Diebstahl). Nun erfolgte – für die Teilnehmenden unbemerkt – die randomisierte Zuordnung in die Zeitdruck-Gruppen. Diejenigen ohne Zeitdruck erhielten die Instruktion „Bitte nehmen Sie sich Zeit für die Bearbeitung des Falles! Betrachten Sie in Ruhe die Beweislage.“, wohingegen diejenigen mit Zeitdruck folgendermaßen instruiert wurden: „Bitte bearbeiten Sie den Fall zügig! Die Zeit, die Sie dafür benötigen, wird erfasst. Oberhalb der Fallbeschreibung wird eine Uhr eingeblendet, sodass Sie Ihre Bearbeitungszeit nachverfolgen können.“. Diese Instruktionen erinnern an den *speed-accuracy-trade-off*, im Rahmen dessen der Fokus entweder auf die Geschwindigkeit oder die Genauigkeit einer Bearbeitung gelegt wird (s. 2.2.3.1). Unterhalb der jeweiligen Instruktion erschien die Fallbeschreibung. Am Ende der Fallbeschreibung wurden die Proband:innen nochmals erinnert: „Bitte nehmen Sie sich die Zeit, die Sie zur Bearbeitung des Falles benötigen.“ (ohne Zeitdruck) oder „Bitte arbeiten Sie zügig! Denken Sie daran, dass die Zeit, die Sie zur Bearbeitung des Falles benötigen, erfasst wird.“ (mit Zeitdruck). Auf der Folgeseite mussten sich die Versuchsteilnehmenden entscheiden, ob ein Tatverdacht vorliegt oder nicht. Diejenigen ohne Zeitdruck sollten bedacht antworten („Bitte entscheiden Sie nun, nachdem Sie die Beweislage sorgfältig betrachtet haben.“), diejenigen mit Zeitdruck sollten sich spontan äußern („Bitte entscheiden Sie ganz spontan.“). Wurde der Tatverdacht bejaht, mussten sich die Versuchsteilnehmenden für eine der drei möglichen Optionen (Erhebung einer Anklage, Einstellung unter Auflagen oder Einstellung wegen Geringfügigkeit) entscheiden, entweder spontan („Für welche Option würden Sie sich ganz spontan entscheiden?“) oder mit Bedacht („Für welche Option würden Sie sich nach dem sorgfältigen Betrachten der Beweislage entscheiden?“).

Ab diesem Zeitpunkt folgten Fragen mit den gleichen Inhalten und Instruktionen, unabhängig von der Delikt- oder Zeitdruck-Zuordnung. Die Teilnehmenden sollten angeben, wie leicht ihnen die Entscheidung hinsichtlich des Vorliegen des Tatverdachtes fiel, wie sicher sie sich dabei fühlten und wie überzeugt sie von der Rich-

tigkeit ihrer Entscheidung waren. Es konnten optionale qualitative Angaben darüber gemacht werden, woran eine schlechte Entscheidung erkannt würde. Nachdem mittels Schieberegler der Überzeugungsgrad eingeschätzt wurde, sollte angegeben werden, wie viele der insgesamt vier Beweismittel aus der Fallbeschreibung für die Entscheidung über den Tatverdacht nötig waren und welche davon am bedeutsamsten und am wenigsten bedeutsam waren. Zur Auswahl standen das Erscheinungsbild des Beschuldigten (Augenschein), der Auszug aus dem Bundeszentralregister (Urkunde), die Aussage des Zeugen sowie die Aussage des Beschuldigten. Außerdem wurde die subjektiv eingeschätzte Bedeutsamkeit jedes einzelnen der vier Beweismittel erfragt. Es konnten optionale qualitative Angaben darüber gemacht werden, welche Nachermittlungen es zu erheben gelte. Die finalen Fragen zur Vignette bezogen sich auf die Einschätzung der Schwere der Tat und auf den Realitätsbezug der Fallbeschreibung.

Um den inhaltlichen Themenwechsel – sozusagen von Jura zu Psychologie – vorzubereiten, erhielten die Versuchsteilnehmenden den Hinweis, dass es in einem zweiten Teil der Befragung um individuelle Einstellungen und Ansichten gehen würde. Die vier Items der NFC-Kurzskala wurden randomisiert präsentiert. Auf der Folgeseite standen randomisiert die Items des CRT, die als „Rätselaufgaben“ eingeleitet wurden. Abschließend wurden alle Versuchsteilnehmenden gebeten Angaben zu Geschlecht und Alter zu machen. Sie wurden zudem gefragt, ob sie während der Umfrage Pausen eingelegt hatten (wenn ja, wie lange) und welches Endgerät zur Teilnahme genutzt wurde (Smartphone, Tablet, Laptop/PC). Die weiteren Fragen waren auf die jeweilige Expertise-Gruppe abgestimmt. Laien machten Angaben zu ihrem Berufsstand. Studierende gaben zusätzlich ihr Studienfach an, um sicherzustellen, dass sie nicht eigentlich der Noviz:innen-Gruppe hätten zugeordnet werden müssen. Naive und Noviz:innen gaben an, welche Berührungspunkte sie bereits mit dem Strafverfahren hatten, z. B. als Zuschauer:in oder als Geschädigte:r (Mehrfachnennungen möglich). Zum Abgleich, ob die Zuordnung in die Expertise-Gruppe korrekt erfolgt war, wurden Noviz:innen gefragt, ob sie Rechtswissenschaft/Jura studierten oder sich derzeit im Referendariat befänden. Wurde dies bejaht, konnte die Semesteranzahl angegeben werden. Wurde dies verneint, konnte das jeweilige Studienfach spezifiziert werden. Zudem wurden die Noviz:innen nach ihrem Status hinsichtlich Referendariat und bereits erhaltender Abschlüsse befragt.

Die Expert:innen wählten ihre derzeitige Berufsbezeichnung (z. B. Richter:in, Staatsanwältin/-anwalt), ihr Bundesland der beruflichen Tätigkeit und ihr derzeitiges Fachgebiet aus (Zivil-, Strafrecht, Öffentliches Recht). Zudem wurden sie nach den Jahren ihrer Berufserfahrung (insgesamt bzw. im Strafrecht) befragt. Abschließend sollten sie angeben wie vertraut ihnen das Fallbeispiel aus der praktischen Arbeit war und wie hoch sie den Einfluss von Zeitdruck auf juristische Entscheidungen einschätzen.

3.3 Poweranalyse und Zusammensetzung der Stichprobe

Vor Beginn der Datenerhebung wurde die für eine akzeptable Teststärke notwendige Stichprobengröße anhand einer Poweranalyse ermittelt. Die genutzten Materialien sowie der Datensatz sind verfügbar (Ruppenthal, 2022). Notwendige Parameter zur Ermittlung des Umfangs sind neben der Power beziehungsweise der Teststärke (1–β) auch die Effektstärke und das α-Level (Cohen, 1988, 1992). Die A-priori-Poweranalyse wurde für die Hypothese H1 berechnet, da diese im Fokus der Studie steht (s. 2.5). Da keine für diese Analyse erforderlichen Pilotdaten oder vergleichbare, empirisch begründete Schätzungen für die Parameter vorlagen, wurde sich auf Erfahrungswerte berufen und mithilfe der Statistik-Software *R* eine Datensimulation durchgeführt. Dieser Simulation lag gemäß der H1 ein logistisches Regressionsmodell zugrunde. Für die Haupt- und Interaktionseffekte wurde der Regressionskoeffizient *b* mittels der Odds Ratio (OR = 2.48) berechnet, die einem mittleren Effekt entspricht. Basierend auf 1000 Simulationen, einem α-Level von .05, einer Teststärke von .8 und einem geschätzten mittleren Effekt von d = .5 ergab die Analyse einen benötigten Stichprobenumfang von 222 Versuchsteilnehmenden. Mit letztlich 299 Proband:innen wurde diese Zahl überschritten. Der Hauptgrund dafür ist die Dauer des Antragsverfahrens der Autorin beim Ministerium für Justiz. Es konnte nicht eingeschätzt werden, wann welches Gericht oder welche Generalstaatsanwaltschaft informiert beziehungsweise mit welchem Rücklauf teilgenommen würde. Nach dem Überschreiten der erforderlichen Anzahl sollte die Erhebung noch nicht beendet werden, um insbesondere den noch zu rekrutierenden Expert:innen möglichst lange die Gelegenheit zur Teilnahme zu bieten.

Aufgrund des quasi-experimentellen Designs stand bereits vor der Rekrutierung der Versuchsteilnehmenden fest, dass das Merkmal der juristischen Vorbildung bei

ebenjener Rekrutierung von wesentlicher Bedeutung ist, um ausreichend Personen für die jeweilige Expertise-Gruppe zu finden. Um zu erreichen, dass die Teilstichproben der Noviz:innen und Expert:innen möglichst relevant, merkmalsspezifisch repräsentativ und ausreichend groß sind (Pospeschill, 2013; Sonnentag, 2006; Trimmel, 2009), wurden Personen aus diesen Kreisen direkt aufgesucht und bewusst ausgewählt. Für diese beiden Expertise-Gruppen wurden jeweils Flyer beziehungsweise Anschreiben mit Informationen zur Teilnahme entworfen. Diese Flyer und Anschreiben enthielten einen Link sowie einen QR-Code zur Online-Umfrage. Um möglichst viele Studierende und Referendar:innen zu erreichen, wurden Flyer an für diese Personen relevanten Orten aufgehangen (z. B. Bibliotheken der Universität Bonn) oder Anschreiben mit der Bitte um Weiterleitung verschickt (z. B. an Jura-Fachschaften der Universitäten Bonn, Köln, Münster, Bielefeld, Bochum, Osnabrück, Düsseldorf, Kiel, Göttingen, München, Freiburg, Regensburg, Augsburg, Passau und Bremen sowie an andere relevante Institutionen, z. B. das *Kompetenzzentrum für juristisches Lernen und Lehren* der Universität zu Köln oder die *European Law Students' Association* in den Städten Köln und Bonn). Außerdem wurden dem Strafrecht zuzuordnende Lehrende an den Universitäten Bonn und Köln mit der Bitte angeschrieben, die Studierenden in den Lehrveranstaltungen auf die Umfrage aufmerksam zu machen. In den sozialen Medien wurden in für die Noviz:innen relevanten Gruppen Aufrufe zur Teilnahme an der Studie veröffentlicht (z. B. regionale Referendariats- oder Examensgruppen). Die für die Expert:innen bestimmten Anschreiben wurden per Email an Staatsanwaltschaften (Bonn, Köln, Aachen, Koblenz, Siegen, Paderborn, Detmold und Essen), an *Jura Bonn Alumni e.V.* sowie an Kanzleien in den Städten Köln, Bonn, Siegburg und Brühl geschickt. Des Weiteren wurde einem Antrag der Autorin an das Ministerium der Justiz Nordrhein-Westfalen stattgegeben, demnach die Mitarbeitenden der Oberlandesgerichte sowie der Generalstaatsanwaltschaften Köln, Hamm und Düsseldorf um Unterstützung durch deren Teilnahme an der Umfrage gebeten wurden (Erlass vom 25.02.2022, 1410E-II.13/22). Auch die Mitarbeitenden des Oberlandesgerichts Koblenz wurden über die Teilnahme informiert. Bei den Laien handelt es sich um eine Gelegenheitsstichprobe ausgehend vom privaten und beruflichen Netzwerk der Autorin. Es wurde durch direkte Ansprache oder mittels sozialer Medien und Emails versucht, solche Personen zu erreichen, die *nicht* das Merkmal der juristi-

schen Vorbildung aufweisen. Angeschriebene und angesprochenen Personen wurden wiederum gebeten, das jeweils eigene Netzwerk zu rekrutieren – auch im Hinblick auf juristische Noviz:innen und Expert:innen.

Alle Teilnehmenden mussten mindestens 18 Jahre alt sein. Die Erhebung fand im Zeitraum 05.01.–30.03.2022 statt. Die an der Teilnahme interessierten Personen wurden in den Vorabinformationen zur Studie über Freiwilligkeit und Anonymität aufgeklärt. Der Anreiz zur Teilnahme war die freiwillige Option, sich im Anschluss an die Befragung durch Angabe einer Emailadresse für eine Verlosung von 40 Gutscheinen für einen Online-Bücherhandel zu registrieren (im Wert von jeweils 10 Euro). Die Adressen wurden mithilfe einer eigens dafür angelegten Umfrage erhoben, um die Studien- und Kontaktdaten voneinander zu trennen. Psychologie-Studierende der Universität zu Köln erhielten die Möglichkeit sich durch Angabe des universitären Kennnamens 0.5 Versuchspersonenstunden anrechnen zu lassen, die im Laufe des Studiums erworben werden müssen.

Von ursprünglich 403 Interessierten, die mithilfe des Links oder des QR-Codes die Umfrage aufgerufen hatten, schlossen 33 Personen bereits nach dem Lesen der Studieninformationen die Website und eine Person willigte nicht in die Teilnahme ein. Insgesamt wurden 71 Personen aus dem Datensatz entfernt (s. 3.7). Weitere Gründe dafür waren, neben der fehlenden Einwilligung zur Teilnahme ($N = 1$), das Beenden der Umfrage unmittelbar nach der Einwilligung ($N = 17$) beziehungsweise mit Darbietung der Vignette ($N = 47$), die Tatsache, dass sich Expert:innen in den demografischen Fragen als Noviz:innen herausstellten und sich somit in der falschen Expertise-Gruppe befanden ($N = 3$), sowie die Tatsache, dass Noviz:innen ihr aktuelles Semester mit der Zahl „0“ definierten und dadurch keine eindeutige Interpretation des juristischen Vorwissens möglich war ($N = 3$). Es kann vermutet werden, dass es sich bei einem Teil der Personen, die unmittelbar nach der Darbietung der Vignette die Umfrage beendet hatten ($N = 47$), um solche handelte, die in einem ersten Durchgang vor der Beantwortung der Frage nach dem Tatverdacht die Fallbeschreibung erneut lesen wollten. Aus diesem Grund starteten sie womöglich die Umfrage erneut, wurden aber der jeweils anderen Vignette zugelost und somit wurde – vermutlich aus zeitlichen oder motivationalen Gründen – keine zweite Fallbeschreibung bearbeitet und die Umfrage stattdessen beendet. Die Proband:innen waren zuvor nicht informiert worden, dass die Vignette nur einmalig präsentiert und

es keine Möglichkeit zum Nachlesen der Fallinhalte geben würde. Träfe diese Vermutung zu, würde es sich bei dem gesamten Datensatz nicht um 403 verschiedene Personen handeln, sondern um einzelne „Wiederholungstäter:innen".

Letztlich gingen die Datensätze von 299 Personen in die Analysen ein. In Tabelle 3.1 ist dargestellt, wie sich die demografischen Daten sowie die Zuordnungen zu den Experimentalgruppen zusammensetzen. Die Naiven waren leicht unterrepräsentiert. Erwartungsgemäß war die Stichprobe der Noviz:innen am jüngsten. Die Gesamtstichprobe setzte sich in etwa zu zwei Dritteln aus weiblichen Personen zusammen. Aufgrund des bereits beschriebenen Ausschlusses einzelner Datensätze konnte nicht mehr gewährleistet werden, dass die Zuordnung zu den Experimentalgruppen gleich verteilt bleibt. Die Verteilung für die Zeitdruck-Gruppen war nach den Ausschlüssen aber weiterhin nahezu gleich, wohingegen das Delikt „Körperverletzung" leicht überrepräsentiert war.

Tabelle 3.1. Darstellung der Zusammensetzung der Stichprobe hinsichtlich demografischer Angaben sowie der Zuordnung zu den Experimentalgruppen in Abhängigkeit von Expertise

	Demografische Angaben					Zuordnung zu den Expertimentalgruppen			
	Alter		Geschlecht			Delikt		Zeitdruck	
Expertise (*N*)	*M*	*SD*	m	w	n.-b.	DS	KV	ohne	mit
Laien (72)	38.4	13.69	24	48	-	31	59	37	35
Noviz:innen (118)	23.3	4.28	54	64	3	50	60	60	58
Expert:innen (109)	44.6	11.08	44	65	-	58	37	58	51
Gesamt (299)	34.5	13.7	89	150	3	120	156	155	144

Anmerkungen. m = männlich, w = weiblich, n.-b. = nicht-binär/drittes Geschlecht, DS = Diebstahl, KV = Körperverletzung. Eine Person der Noviz:innen-Gruppe machte keine Angaben zum Geschlecht (1%). Die Berechnungen und Anzahlen beziehen sich aufgrund von Dropout nicht in jedem Fall auf die Gesamtstichprobe von $N = 299$, sondern sie stehen in Relation zu der jeweils vorliegenden (Teil-)Stichprobe.

Die von der Software vorausgesagte Dauer der Teilnahme betrug 17.3 Minuten. Diejenigen, die die Umfrage bis zum Schluss bearbeitet und die eine pausenfreie Teilnahme angegeben hatten, benötigten im Schnitt rund 2001 Sekunden, was etwa 33 Minuten entspricht ($M = 2001.7$, $SD = 11830.5$). Der Median ergab allerdings rund 13 Minuten Bearbeitungszeit ($Md = 775$). Insbesondere auf die für die Fallbeispiele benötigte Lesezeit wird in den Ergebnissen zum Manipulationscheck eingegangen (s. 4.1).

Die meisten Personen nahmen über das Smartphone (N = 125, 51.4%) oder einen Laptop/PC (N = 111, 45.7%) an der Studie teil und nur ein kleiner Teil der Stichprobe nutzte ein Tablet (N = 7, 2.9%). Im Verlauf der Studie beendeten einige Versuchsteilnehmende die Umfrage. Allerdings waren diese Dropout-Raten in den Expertise-Gruppen vergleichbar hoch, sodass auch zum Ende der Umfrage keine der Gruppen unverhältnismäßig über- oder unterrepräsentiert war. Die Rate beschreibt den Anteil derjenigen Personen, die in die Teilnahme eingewilligt, aber die letzte Frage der Studie nicht mehr beantwortet hatten. Die Dropout-Rate betrug bei den Naiven 27.8%, bei den Noviz:innen 30.8% und bei den Expert:innen 36% (gruppenübergreifend: 32.5%). Auf wie vielen Versuchsteilnehmenden letztlich die einzelnen Ergebnisse der statistischen Verfahren basieren wird im jeweiligen Abschnitt näher spezifiziert (s. 4). Die hier genannten Dropout-Raten beinhalten auch diejenigen Proband:innen, die bereits als Wiederholungstäter:innen beschrieben wurden, sodass einige der Werte vermutlich auf Dopplungen basieren. Die genaue Erfassung der Stelle im Ablauf, an der die meisten Personen die Umfrage abgebrochen hatten, war nicht relevant.

Die demografischen Daten wurden am Ende der Umfrage erhoben (Block 4; s. Abbildung 3.1), sodass diese bereits vom Dropout beeinflusst sind. Aus diesem Grund sind neben den Anzahlen auch die prozentualen Anteile genannt. Unter den Naiven gaben acht Personen an zu studieren (12.3%), wohingegen der Großteil berufstätig war (Vollzeit: N = 36, 55.4%; Teilzeit: N = 12, 18.5%; Elternzeit: N = 2, 3%). Insgesamt sieben Personen waren nicht berufstätig (10.8%). Der überwiegende Teil der Naiven hatte bisher keine Berührungspunkte mit einem Strafverfahren (N = 42, 65%). Diejenigen, die Erfahrungen hatten, beschrieben diese mit der Funktion als Zeug:in (N = 10, 15.4%), Begleitung/Zuschauer:in (N = 8, 12.3%), Geschädigte:r (N = 5, 7.7%) oder Beschuldige:r/Angeklagte:r, Schöffin/Schöffe, Gerichtsreporter:in und sachkundige:r Zeug:in (jeweils N = 1, jeweils 1.5%). Bei den Noviz:innen zeigte sich erwartbar ein anderes Bild, da die Mehrheit von Erfahrungen berichtete (N = 58, 60%), insbesondere in der Funktion als Begleitung/Zuschauer:in (N = 42, 44%), gefolgt von der Funktion als Zeug:in (N = 12, 12.5%), als Geschädigte:r (N = 6, 6.3%) oder als Beschuldigte:r/Angeklagte:r (N = 4, 4.2%). Zusätzlich wurden in einem optionalen Antwortfeld die Funktionen als Praktikant:in/Referendar:in (N = 9, 9.4%), Sitzungsvertreter:in (N = 2, 2.1%) sowie Mitarbeiter:in

der Verteidigung (N = 1, 1%) genannt, welche aber bereits in der Definition als Noviz:in beinhaltet sind. Im Durchschnitt studierten die Noviz:innen im 6. Semester (M = 6.89, SD = 4.3, Md = 7). Der Großteil besaß zum Zeitpunkt der Umfrage noch keinen Abschluss (N = 72, 77%), wohingegen 21 Personen (23%) bereits das 1. Staatsexamen erreicht hatten. Nur die Minderheit befand sich zum genannten Zeitpunkt bereits im Referendariat (N = 19, 20%). Die Mehrheit der Expert:innen war als Richter:in tätig (N = 48, 58.6%). Insgesamt 16 Personen (19.5%) arbeiteten als Staatsanwältin/-anwalt, 5 Personen als Rechtsanwältin/-anwalt (6.1%) und 2 Personen als Amtsanwältin/-anwalt (2.4%). Insgesamt 11 Personen (13.4%) gaben sonstige Berufsbezeichnungen an, die aber als volljuristische Abschlüsse beschrieben werden können. Nahezu die Hälfte war beruflich im Bundesland Nordrhein-Westfalen tätig (N = 44, 54%), gefolgt von Rheinland-Pfalz (N = 31, 38%), Berlin (N = 3, 4%), Brandenburg (N = 2, 2%) und Bayern oder Hessen (jeweils N = 1, jeweils 1%). Betrachtet man die allgemeine Berufserfahrung, so wird eine sehr erfahrene Stichprobe deutlich: 36 Personen waren bereits seit mindestens 16 Jahren tätig (43.9%), 27 Personen zwischen 6 und 15 Jahren (32.9%) und 19 Personen seit maximal 5 Jahren (23.2%). Bezogen auf das Strafrecht im Besonderen ändert sich die Verteilung: Lediglich 11 Personen gaben nun eine Erfahrungszeit von über 16 Jahren an (13.4%), 17 Personen waren dagegen zwischen 6 und 15 Jahren (20.7%) und 54 Personen seit maximal 5 Jahren tätig (65.9%). Auch wenn die Berufserfahrung in Jahren im Strafrecht mehrheitlich gering war, war das Fachgebiet dennoch stark vertreten: Insgesamt 39 Personen (48%) gaben das Strafrecht als ihr derzeitiges Fachgebiet an (Zivilrecht: N = 37, 45%; Öffentliches Recht: N = 6, 7%). Den Einfluss von Zeitdruck auf juristische Entscheidungen schätzten die Expert:innen als eher niedrig ein (M = 3.46, SD = 1.55; s. Abbildung 3.2). Es wird eine Tendenz zur mittleren Antwort erkennbar.

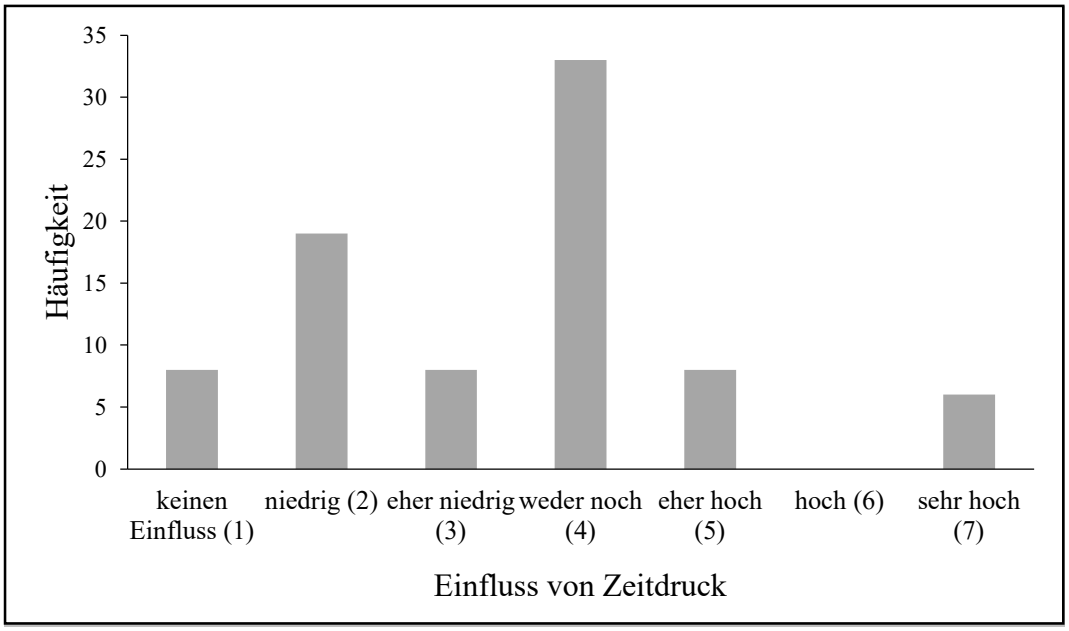

Abbildung 3.2. Einschätzung der Expert:innen über den Einfluss von Zeitdruck auf juristische Entscheidungen auf einer 7-stufigen Likert-Skala.

Aufgrund eines Fehlers in der Skalierung lässt sich die Vertrautheit der Fachpersonen mit den jeweiligen Fallbeschreibungen nicht eindeutig interpretieren. Um Transparenz zu gewährleisten, ist die Verteilung in Abbildung 3.3 dargestellt. Die Option „vertraut“ ist zweimal vertreten, was zwei Interpretationen zulässt. Einerseits könnten sich die Personen an den Zahlenwerten 1–7 als Abstufung orientiert haben („vertraut“ würde einem Wert von 6 entsprechen, weil ein hoher Punktwert für ein hohes Maß der Vertrautheit steht). Andererseits könnten die Teilnehmenden die Abstufungen als Worte wahrgenommen und sich deswegen für die Option „vertraut“ mit dem Wert von 3 entschieden haben, weil diese Option *vor* dem zweiten „vertraut“ mit dem Punktwert von 6 stand, somit zuerst gelesen und als zutreffend befunden wurde.

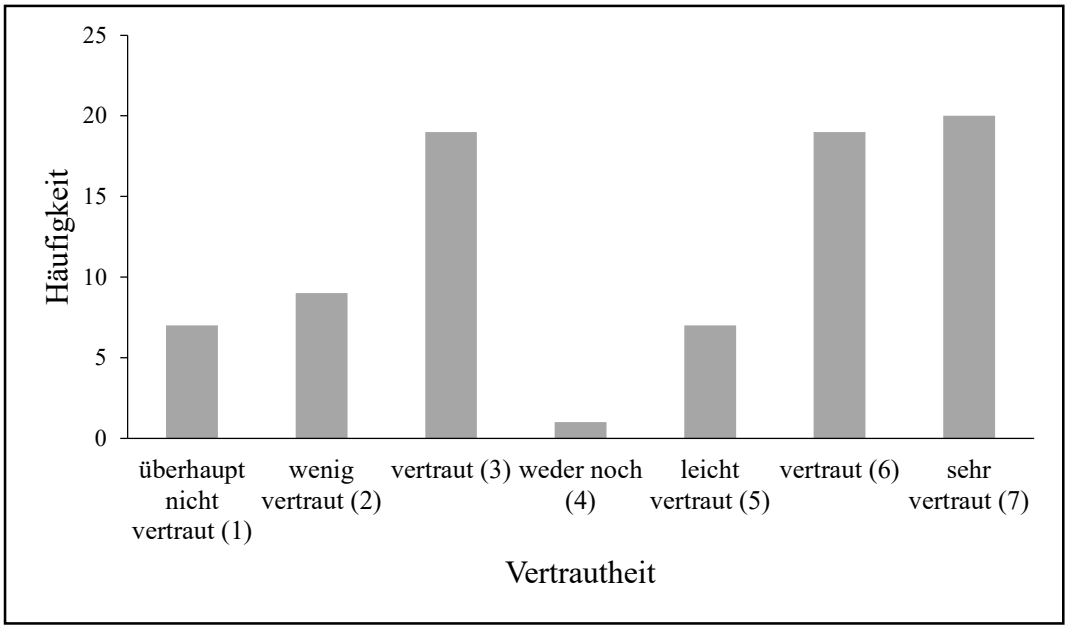

Abbildung 3.3. Einschätzung der Expert:innen über das Vertrautheitsmaß der Delikte in den behandelten Vignetten auf einer 7-stufigen Likert-Skala.
Anmerkungen. Aufgrund des Fehlers in der Skalierung und der dadurch erschwerten Interpretation wurde auf eine zusätzliche Auswertung getrennt nach Delikttypen verzichtet.

Betrachtet man die beiden extremen Pole mit den Punktwerten 1 und 7, lässt sich aber vermuten, dass die Fallbeschreibungen den Expert:innen eher sehr vertraut als überhaupt nicht vertraut waren, da es an diesen beiden Stellen weder zu numerisch noch zu verbal begründeten Unklarheiten gekommen sein sollte.

3.4 Verwendete Materialien

Da es sich um eine fragebogenbasierte Onlineumfrage handelte, wurde bei der Auswahl der Materialien darauf geachtet, Formate und Instruktionen zu wählen, die selbstständig ohne Versuchsleitung bearbeitet werden können. Die Materialien werden in der Reihenfolge beschrieben, in der sie laut Abbildung 3.1 im Versuchsablauf vorkommen: Vignetten (s. 3.4.1), vignettenbezogener Fragebogen (s. 3.4.2), *Need-for-Cognition*-Kurzskala (s. 3.4.3) und *Cognitive Reflection Test* (s. 3.4.4). Auf die Gütekriterien wird in Abschnitt 3.5 eingegangen.

3.4.1 Vignetten

Für diese Studie wurden eigens zwei Fallbeschreibungen konstruiert (s. A-2; A-3). Wichtige Kriterien für die Qualität von Vignetten sind deren Plausibilität und Nähe zur Realität, wobei es darum geht, die Beschreibung gleichzeitig informativ, aber

auch vage zu gestalten (Bieneck, 2009; J. Finch, 1987). Es wurde bei der Konstruktion der Fallbeispiele darauf geachtet, dass diese realistisch, plausibel und von klarer Struktur sind und dass die zu bewertende Handlung nicht durch inhaltliche Nebenstränge aus dem Fokus der Geschichte gerät (Bieneck, 2009). Die Delikte wurden ausgewählt, weil Eigentums- und Vermögensdelikte sowie Straftaten gegen das Leben und die körperliche Unversehrtheit zahlenmäßig stark bei den erledigten Strafverfahren vertreten sind (Statistisches Bundesamt, 2020a, 2020b; s. 2.1.5). Somit kommen Diebstähle und Körperverletzungen in der strafrechtlichen Praxis verhältnismäßig häufig vor und – so eine Vermutung – verleiten zu schnellen Entscheidungen (s. auch Bieneck, 2006). Um zu vermeiden, dass die Natur eines Delikts ausschlaggebend für das Entscheidungsverhalten der Proband:innen ist, wurden *beide* in jeweils einer Fallbeschreibung realisiert. Laut StGB beträgt die Obergrenze des Strafrahmens jeweils maximal 2 Jahre, sodass die beiden, zumindest aus strafrechtlicher Sicht, in ihrer Schwere vergleichbar sind. Aufgrund der Länge der Befragung wurde sich für die randomisierte Zuordnung lediglich einer Vignette pro Person entschieden. Ein häufiges Vorgehen in der Empirie ist das Manipulieren bestimmter Dimensionen innerhalb einer Vignette, wobei diese Dimensionen hinsichtlich ihrer Level variieren (Auspurg, Hinz & Liebig, 2009; s. 2.3.1). Dies ist in der vorliegenden Studie nicht der Fall, denn innerhalb der Vignetten fand keinerlei Randomisierung statt (z. B. Name). Die Randomisierung bezieht sich lediglich auf den Delikttyp und nicht auf Variationen von Dimensionen, um auch „die Wirkung der Befragtenmerkmale auf die Urteile [zu] analysieren“ (Beck & Opp, 2001, S. 301).

Die Vignetten stellen die Zusammenfassung eines (polizeilichen) Ermittlungsberichts dar. Länge und Aufbau der beiden Vignetten sind sehr ähnlich. Sofern dies inhaltlich sinnvoll war, ähneln sich auch gewisse Formulierungen. Einleitend werden Angaben zur Person des Beschuldigten und zum Tatvorwurf gemacht. Es folgt eine Beschreibung des Geschehens aus Sicht des Geschädigten. Im Verlauf wird auf die vier Kategorien von Beweismitteln (s. 2.1.6) eingegangen: Proband:innen erhalten Informationen zur Aussage eines Zeugen, zur Aussage des Beschuldigten, zu einem Foto dessen (Augenschein) sowie zu seinen Vorstrafen (Urkunde). Beide Vignetten beinhalten qualitativ und quantitativ vergleichbare Informationen, die letztlich ein nicht eindeutiges Bild über das Geschehen liefern. Auch wenn sich die

Inhalte der Vignette aufgrund der Verschiedenartigkeit der Delikte nicht ähneln, so wurde darauf geachtet, an möglichst vielen inhaltlichen Stellen ein hohes Maß an Standardisierung und Vergleichbarkeit zu erreichen beziehungsweise den Einfluss möglicher (extralegaler) Störvariablen (z. B. Name) zu reduzieren. In beiden Vignetten ist der Beschuldigte männlich und 23 beziehungsweise 24 Jahre alt. Diese Auswahl ist mit den demografischen Daten der Beschuldigten begründet, laut denen der Großteil männlichen Geschlechts und zwischen 21 und 25 Jahren alt ist (Statistisches Bundesamt, 2020b). Dadurch stellt die Beschreibung des Beschuldigten in den Vignetten hinsichtlich Alter und Geschlecht keine Besonderheit, sondern eine in der Praxis häufig vorkommende Tatsache dar. Des Weiteren wurde sich für die deutsche Staatsbürgerschaft entschieden, um mögliche Vorurteile oder Stereotype zu anderen Ländern und Kulturen zu kontrollieren. Die Vornamen der Beschuldigten (Michael, Matthias) können als zeitlos beschrieben werden (Rudolph et al., 2007; s. auch Nett et al., 2020). Es wird die Wahrscheinlichkeit reduziert, dass die Wahl des Vornamens starke positive oder negative Konnotationen hervorruft. Zudem beginnen beide mit dem gleichen Buchstaben. Die Nachnamen wurden jeweils abgekürzt (B. und P.) und können aufgrund ihres ähnlichen Lautes als vergleichbar eingestuft werden. Die beiden weiteren in der Fallbeschreibung vorkommenden Personen (Geschädigter und Zeuge) sind ebenfalls männlich. Damit sollte vermieden werden, dass das Entscheidungsverhalten zu stark auf möglichen geschlechtsbezogenen Aspekten beruht (z. B. Sympathisieren mit einer weiblichen Geschädigten; Gabriel, 2009), da dies nicht Gegenstand der Untersuchung war. Es wurde sich dagegen entschieden, den Geschädigten und den Zeugen jeweils als geschlechtsneutrale Personen zu beschreiben (um den Einfluss dieser Störvariable stärker zu kontrollieren), da sich dies sprachlich nur schwer hätte umsetzen lassen. Womöglich hätte dies den Lesefluss der Proband:innen beeinflusst, was letztlich für den Manipulationscheck (Zeitmessung beim Lesen) nicht zielführend gewesen wäre.

3.4.2 Vignettenbezogener Fragebogen

Ebenso wie das Präsentieren und Lesen der Vignette befand sich der darauf bezogene Fragebogen im zweiten Block des Versuchsablaufs (s. Abbildung 3.1). Die darin beinhalteten Fragen wurden eigens für diese Studie erstellt (17 Items). Die Items wurden nicht umgepolt oder zu Skalen zusammengefasst, da sie eigenständige Variablen darstellen. Beim Erstellen von neuen Items wurde darauf geachtet,

dass die Skalierung letztlich sinnvoll für die angedachten statistischen Methoden ist. Metrisch-skalierte Skalen sind zielführend, wenn es um die Erfassung des Grades der Zustimmung beziehungsweise der Ablehnung mit einem Item geht (Goddard & Villanova, 2006). Daher wurde sich für eine Operationalisierung und eine Formulierung von Items entschieden, die zu metrisch-skalierten Variablen führen. Es kamen fast ausschließlich geschlossene Fragen zum Einsatz, da diese besser vergleichbar und einfacher in statistische Analysen zu integrieren sind (Goddard & Villanova, 2006). Die eingesetzten Items lassen sich dahingehend differenzieren, ob sie unter dem Eindruck der Zeitdruckmanipulation erhoben wurden (s. 3.4.2.1), ob sie sich auf den Entscheidungsprozess beziehungsweise das Entscheidungserleben (s. 3.4.2.2) oder auf die Beweislage und das Delikt (s. 3.4.2.3) bezogen.

3.4.2.1 Items unter dem Eindruck der Zeitdruckmanipulation

Unmittelbar nach dem Lesen der Vignette folgte die Kernfrage der Studie nach dem Vorliegen des Tatverdachtes. Für dieses Item gab es die Antwortoptionen „Tatverdacht liegt nicht vor." beziehungsweise „Tatverdacht liegt vor.". Wurde der Tatverdacht bejaht, sollte der nächste Verfahrensschritt näher definiert werden: „Anklage", „Einstellung unter Auflagen" oder „Einstellung wegen Geringfügigkeit". Somit antwortete hier nur ein Teil der Stichprobe. Diese nominal-skalierten Items waren im Forced-choice-Format gehalten. Je nachdem, ob es sich um die Gruppe mit oder ohne Zeitdruckmanipulation handelte, unterschieden sich die Instruktionen (s. 3.2), wenngleich die Antwortoptionen identisch waren.

3.4.2.2 Items mit Bezug zum Entscheidungsprozess

Drei Items befragten die Teilnehmenden nach dem Entscheidungsprozess. Die Entscheidung wurde dabei in den Items jeweils als die „Entscheidung über das Vorliegen das Tatverdachtes" definiert. Von Interesse war es, zu ermitteln, wie leicht die Entscheidung fiel und wie sicher sich die Versuchsteilnehmenden dabei waren. Außerdem sollte angegeben werden, wie überzeugt man war, eine gute Entscheidung getroffen zu haben. Das Antwortformat war eine 7-stufige Likert-Skala, wobei die Endpunkte angezeigt wurden: *nicht leicht* bzw. *nicht sicher* bzw. *nicht überzeugt* (jeweils 1) und *sehr leicht* bzw. *sehr sicher* bzw. *sehr überzeugt* (jeweils 7). Die Items waren im Forced-choice-Format gehalten und intervall-skaliert. Es folgte ein

optionales Item im offenen Antwortformat, in dem spezifiziert werden konnte, woran keine gute Entscheidung erkannt würde.

3.4.2.3 Items mit Bezug zur Beweislage und zum Delikt

Das Beweismaß beschreibt den Grad der Überzeugung, den die Beweislage hinsichtlich der Wahrheit des Sachverhaltes liefert (s. 2.1.6.1): Hat sich die Tat wie beschrieben ereignet? Für die Operationalisierung des Beweismaßes wurde sich bei der Itemkonstruktion an Schweizer (2016) orientiert, der Zahlenwerte zur Veranschaulichung nutzt. Da der Überzeugungsgrad nicht dichotom, sondern graduell ist (Schweizer, 2019), wurde ein Schieberegler als Antwortformat gewählt (0–99%). Zudem wurde 99% (= *nahezu vollends überzeugt*) als Maximalwert definiert, da eine absolute Sicherheit von 100% nicht erreicht werden kann (Schweizer, 2019). Das Item war im Forced-choice-Format gehalten und intervall-skaliert. Zudem stellte es eine unipolare Ratingskala dar, bei der der Nullpunkt nicht näher definiert wurde. Es folgten zwei nominal-skalierte Items, die danach fragten, welche Information aus der Fallbeschreibung am bedeutsamsten beziehungsweise am wenigsten bedeutsam für die Entscheidung über das Vorliegen des Tatverdachtes war. Aus den vier Antwortoptionen konnte entweder das Erscheinungsbild des Beschuldigten, der Auszug aus dem Bundeszentralregister, die Aussage des Zeugen oder die Aussage des Beschuldigten gewählt werden. Da es sich um vier inhaltliche Aspekte (bzw. Beweiskategorien; s. 2.1.6) handelte, sollte im folgenden intervall-skalierten Item geschätzt werden, ob „1“, „2“, „3“ oder „4“ dieser Aspekte für das Treffen einer Entscheidung benötigt wurde(n). Im Anschluss wurden die vier Beweiskategorien näher betrachtet. Es galt jeweils deren Bedeutsamkeit für die Entscheidung genauer zu definieren. Für jede Kategorie folgte ein intervall-skaliertes Item mit einer 7-stufigen Likert-Skala. Die Endpunkte waren als *nicht wichtig* (1) und *sehr wichtig* (7) definiert. Sämtliche Items waren im Forced-choice-Format gehalten. Es folgte ein optionales Item im offenen Antwortformat, in dem spezifiziert werden konnte, welche Nachermittlungen erhoben werden sollten. Den Abschluss des vignettenbezogenen Fragebogens bildeten zwei intervall-skalierte Items im Forced-choice-Format, die sich auf das beschriebene Delikt bezogen. So sollte in einer unipolaren Ratingskala mittels eines Schiebereglers die Schwere der Tat definiert wer-

den (100 = *sehr schwere Tat*). Auf einer 7-stufigen Likert-Skala (1 = *nicht realistisch* bis 7 = *sehr realistisch*) galt es abschließend den Realitätsbezug der Fallbeschreibung einzuschätzen.

3.4.3 *Need-for-Cognition*-Kurzskala

Aus zeitökonomischen Gründen wurde sich für die *Need-for-Cognition*-Kurzskala (NFC-K) nach Beißert et al. (2014) entschieden, die vier Items beinhaltet:

1. Es genügt mir einfach die Antwort zu kennen, ohne die Gründe für die Antwort eines Problems zu verstehen.
2. Ich habe es gern, wenn mein Leben voller kniffliger Aufgaben ist, die ich lösen muss.
3. Ich würde kompliziertere Probleme einfachen Problemen vorziehen.
4. In erster Linie denke ich, weil ich muss.

Die Items waren im Forced-choice-Format gehalten und wurden randomisiert präsentiert. Die Teilnehmenden gaben auf einer 7-stufigen Likert-Skala an, inwiefern die Items im Allgemeinen auf sie persönlich zutreffen. Die Endpunkte waren als *trifft überhaupt nicht zu* (1) und *trifft voll zu* (7) definiert. Die Items mit der Nummer 1 und 4 der obigen Auflistung wurden zur Auswertung umgepolt. Pro Person wurden die Werte addiert, sodass ein hoher Punktwert von maximal 28 einer hohen NFC-Ausprägung entspricht. Laut Cronbachs α ist die Reliabilität der Kurzskala mit α = .54 als niedrig einzustufen (s. auch Field, 2013). Die Interkorrelation des CRT und der NFC-Kurzskala weist auf einen positiven, nicht signifikanten Zusammenhang hin (r = .046, p = .5).

3.4.4 Cognitive Reflection Test

Zum Einsatz kamen die drei Items des *Cognitive Reflection Test* nach Frederick (2005) in deutscher Übersetzung (s. 2.2.3.3). Diese Items wurden als „Rätselaufgaben“ angekündigt, die nur mit ganzzahligen Angaben (ohne Einheit) beantwortet werden konnten (s. Tabelle 3.2).

Tabelle 3.2. Ins Deutsche übersetzte CRT-Items nach Frederick (2005) sowie deren als ganze Zahl ausgedrückte Lösung ohne Nennung einer Einheit

Item	Lösung
Ein Schläger und ein Ball kosten insgesamt 1,10 Euro. Der Schläger kostet 1,00 Euro mehr als der Ball. Wie viel Cent kostet der Ball?	5
Wenn 5 Maschinen 5 Minuten benötigen, um 5 Gegenstände herzustellen, wie viele Minuten würden 100 Maschinen benötigen, um 100 Gegenstände herzustellen?	5
In einem See befinden sich Seerosenblätter. Jeden Tag verdoppeln sich die Seerosenblätter. Wenn es 48 Tage dauern würde, bis die Seerosenblätter den gesamten See bedecken, wie viele Tage würde es dauern, bis die Hälfte des Sees bedeckt ist?	47

Die Items waren im Forced-choice-Format gehalten und wurden randomisiert präsentiert. Richtige Antworten wurden jeweils mit einem Punkt bewertet, falsche Antworten erhielten dagegen keinen Punkt. Die Itemwerte wurden für jede Person addiert, sodass der Gesamtscore die Skala *kognitive Reflexion* ergab. Somit reicht die Spanne der Punktwerte von 0 bis 3, wobei ein höherer Wert mit einer stärker ausgeprägten kognitiven Reflexion einhergeht. Laut Cronbachs α ist die Reliabilität der Skala mit α = .56 als niedrig einzustufen (s. auch Field, 2013). Für insgesamt 22 Personen wurde kein Gesamtwert berechnet, da diese während der Bearbeitung des CRT die Studie beendet und somit nicht die notwendigen drei Items beantwortet hatten.

3.5 Gütekriterien der verwendeten Materialien

Bei Messinstrumenten bezieht sich die Objektivität auf die Unabhängigkeit der Ergebnisse von der untersuchenden Person (Pospeschill, 2013). Aufgrund der Tatsache, dass es sich um eine onlinebasierte, „versuchsleitungsfreie“ Untersuchung mit genau fest gelegten Instruktionen handelte, ist die Durchführungsobjektivität gegeben (Trimmel, 2009). Hinsichtlich der geschlossenen Items im vignettenbezogenen Fragebogen, im CRT und in der NFC-K können die Auswertungs- und Interpretationsobjektivität als gegeben angesehen werden (Trimmel, 2009). Lediglich bei der Auswertung und auch der Interpretation der qualitativen Items im vignettenbezogenen Fragebogen ist die Objektivität eingeschränkt (optionale Angaben zur Qualität der Entscheidung beziehungsweise zu Nachermittlungen). In Abschnitt 4.12 wird dargestellt, wie bei der Auswertung versucht wurde, möglichst objektive Standards zu gewährleisten. Die Reliabilität eines Messinstruments beschreibt, wie genau und zuverlässig ein Merkmal gemessen wird; das besagte Instrument gilt dann

als valide, wenn es das Merkmal misst, was es vorgibt zu messen (Peters & Dörfler, 2019a). Da der empirisch fundierte CRT sowie die NFC-K genutzt wurden, wird für nähere Angaben zu Reliabilität und Validität auf die jeweilige Literatur verwiesen (CRT: Frederick, 2005; NFC: Beißert et al., 2014; Bless et al., 1994; Cacioppo et al., 1996). Der Reliabilitätskoeffizient Cronbachs α sowie die Interkorrelation zwischen den beiden Testverfahren wurden für die vorliegende Studie in den Abschnitten 3.4.3 und 3.4.4 berechnet. Nähere Ausführungen zur internen und externen Validität der Studie und ihrer Schlussfolgerungen folgen in Abschnitt 5.3.

3.6 Pilotstudie zur Auswahl der Vignetten und zur Überprüfung der Zeitdruckmanipulation

Im Vorfeld dieser Studie wurde eine Pilotstudie mit folgenden Fragestellungen durchgeführt:

- Waren die Instruktionen zur Manipulation des Zeitdrucks in den Gruppen *mit Zeitdruck* und *ohne Zeitdruck* ausreichend?
- Welche Vignetten sollten aufgrund bestimmter Kriterien für die Hauptstudie ausgewählt werden?

An der Pilotstudie, die im Zeitraum 06.–15.04.2021 durchgeführt wurde, nahmen 33 Personen aus dem beruflichen und privaten Umfeld der Autorin teil. Davon wurden 15 Personen der Gruppe *mit Zeitdruck* und 18 Personen der Gruppe *ohne Zeitdruck* randomisiert zugeordnet. Ursprünglich standen vier Vignetten zur Auswahl, aus denen anhand der Kriterien „Realismus“ und „Varianz in den Entscheidungen“ zwei ausgewählt werden sollten. Die Reihenfolge der präsentierten Vignetten war randomisiert, wobei jeder Person alle Szenarien präsentiert wurden. Neben den beiden in dieser Studie eingesetzten Fallbeispielen (s. 3.4.1) standen zudem eine Vignette über einen Fall von Diebstahl eines Rucksacks und eine Vignette über einen Fall von Körperverletzung in einem Waldgebiet zur Verfügung. Über die gesamte Studie hinweg war der Unterschied in der Bearbeitungsdauer (Sekunden) zwischen den Zeitdruck-Gruppen signifikant (ohne Zeitdruck: mittlerer Rang = 20.71; mit Zeitdruck: mittlerer Rang = 10.29; $U = 199.00$, $p = .001$).[32] Auf deskriptiver Ebene zeigt sich für jede der vier Vignetten, dass diejenigen mit Zeitdruck eine schnellere

[32] Dies entsprach gerundet 19 bzw. 54 Minuten. Der Median betrug gerundet 16 bzw. 38 Minuten.

Bearbeitungsdauer in den Vignetten aufwiesen als diejenigen, die ohne Zeitstress arbeiten konnten. Dieser Unterschied war nur für die Fälle *Kopfhörer* (ohne Zeitdruck: mittlerer Rang = 20.22; mit Zeitdruck: mittlerer Rang =13.13; U = 193.00, p = .036) sowie *Wald* (ohne Zeitdruck: mittlerer Rang = 19.06; mit Zeitdruck: mittlerer Rang = 11.43; U = 169.00, p = .017) statistisch signifikant. Ausgehend davon wurden Ideen zur Verstärkung des Zeitdrucks abgeleitet, z. B. das Einbauen eines Hinweises auf die Zeitmessung in der Instruktion (s. 3.2). Die Frage nach dem Realitätsbezug einer Vignette erwies sich aufgrund der geringen Mittelwertsunterschiede auf deskriptiver Ebene als unzureichendes Auswahlkriterium. In Kombination mit der Betrachtung der Varianz in den Anzahlen der ausgewählten Entscheidungsoptionen erwiesen sich die Vignetten *Kopfhörer* und *Sportplatz* als am geeignetsten, da diese keine eindeutigen Antworttendenzen der Teilnehmenden „provozierten“. In den Vignetten *Rucksack* und *Wald* entschieden sich beispielsweise jeweils rund zwei Drittel der Teilnehmenden für eine Einstellung (kein Tatverdacht), wohingegen die Antworten in den letztlich ausgewählten Fallbeispielen ausgewogener waren und somit die angestrebte Nichteindeutigkeit der Beweislage untermauerten.

3.7 Inspektion und Bereinigung der Daten

Vor der Berechnung von Tests wurde das Datenset hinsichtlich möglicher Ausreißer inspiziert, um eine Verzerrung der Ergebnisse zu vermeiden (s. auch Peters & Dörfler, 2019a; Tabachnick & Fidell, 2013). Auf weiterführende Inspektionen der Normalverteilung oder der Homogenität der Varianzen wird – getrennt nach Hypothesen und Fragestellungen – im Abschnitt 3.8 eingegangen.

Zunächst wurde das Datenset von solchen Teilnehmenden befreit, deren Angaben für die weitere Studie unbrauchbar waren.[33] Zur Ermittlung von Ausreißern in den drei Zeitmessungen wurden Boxplots, Histogramme und z-Werte analysiert (Field, 2013; Peters & Dörfler, 2019a). Deren Ermittlung fand für jede der drei Zeitmessungen zunächst separat statt (Lesezeit für die Vignette, Zeit zur Entscheidung über den Tatverdacht, Zeit zur Entscheidung über den nächsten Verfahrensschritt). Dazu

[33] Dabei handelte es sich beispielsweise um Personen, die sich der falschen Expertise-Gruppe zugeordnet hatten. Eine ausführliche Begründung der Entfernung einzelner Personen aus dem Datensatz findet sich in Abschnitt 3.3.

wurden die drei Variablen in ihre beiden Zeitdruck-Level unterteilt, sodass insgesamt sechs Untergruppen im Fokus der Analyse standen (s. auch Tabachnick & Fidell, 2013). Für die Histogramme und Boxplots wurden die Annahmen des Statistikprogramms SPSS zugrunde gelegt und für die *z*-Werte galt ein Wert $z \geq 3.29$ als auffällig ($p \leq .001$; Field, 2013). Letztlich wurden die Werte der Personen entfernt, die in allen drei Betrachtungen als Ausreißer galten. Für die Lesezeit der Vignetten betraf dies in der Gruppe *mit Zeitdruck* insgesamt zwei Fälle und in der Gruppe *ohne Zeitdruck* vier Fälle, für die Zeit zur Entscheidung über den Tatverdacht betraf es fünf beziehungsweise drei Fälle und für die Entscheidung über den nächsten Verfahrensschritt wurden einer beziehungsweise zwei Fälle entfernt. Galten Personen in den Histogrammen und Boxplots als Extreme, nicht aber hinsichtlich ihres *z*-Wertes (weil $z \leq 3.29$), so wurden diese nicht entfernt, da es durchaus realistisch ist, dass manche Teilnehmende längere Zeit zum Lesen oder Entscheiden benötigen als andere.

Für die kategorischen Variablen stellen Ausreißer kein Problem im eigentlichen Sinne dar. Die Identifikation von Ausreißern fand für die metrischen Variablen *kognitive Reflexion, Need for Cognition, Beweismaß, Schweregrad des Delikts, Anzahl der Beweismittel* sowie *Leichtigkeit, Sicherheit* und *Überzeugung* statt. Die drei Items des CRT und die vier Items der NFC-K wurden – sofern sie vollständig bearbeitet worden waren – jeweils zu einem Gesamtscore zusammengefasst. Es wurden ebenfalls Histogramme, Boxplots und *z*-Werte erstellt und untersucht. Aufgrund der kleinen Spanne zwischen möglichem Minimum und Maximum für kognitive Reflexion war nicht mit Ausreißern zu rechnen, was durch die Analysen bestätigt wurde. Für NFC lag die mögliche Spanne zwischen 4 und 28 Punkten. Laut Histogramm und Boxplot waren neun Fälle als Ausreißer zu identifizieren. Betrachtet man zusätzlich die *z*-Werte, fand sich nur ein Fall, der in allen drei Kriterien auffällig war. Dieser wurde für die weitere Betrachtung von NFC ausgeschlossen. Für das Beweismaß erwies sich kein Wert als auffällig. Hinsichtlich des Schweregrades wurde ein Fall identifiziert, der in allen drei Bereichen auffällig war. Personen, die aufgrund ihrer Werte nur in zwei der drei Kategorien angezeigt wurden, verblieben für diese Variable im Datenset, da eine interindividuelle Unterschiedlichkeit in der Einschätzung des Schweregrades durchaus realistisch ist. Für die Va-

riablen *Leichtigkeit, Sicherheit* und *Überzeugung* war jeweils aufgrund der geringen Spanne zwischen 1 und 7 Punkten nicht mit Ausreißern zu rechnen. Dies wurde durch die Analysen bestätigt. Gleiches galt für die Anzahl der Beweismittel (Punktwertspanne zwischen 1 und 4).

Aufgrund des Forced-choice-Formats (s. 3.4) waren keine fehlenden Werte im eigentlichen Sinn zu erwarten und diese entstanden nur dann, wenn eine Person ihre Teilnahme an der Umfrage beendete und den Fragebogen somit nicht vollständig ausfüllen konnte. Die Teilantworten wurden dennoch gespeichert. Dementsprechend basieren Berechnungen immer auf der (Teil-)Stichprobe, die bei der Beantwortung eines jeweiligen Items noch aktiv beteiligt war. Lediglich die optionalen offenen Fragen wurden von einzelnen Personen nicht ausgefüllt (s. 4.12), was aber nicht als fehlende Werte aufgefasst wurde.

3.8 Statistische Verfahren und Prüfung der Voraussetzungen

In Abschnitt 2.5 werden die für diese Studie relevanten Hypothesen und Fragestellungen formuliert. Zur genauen Beschreibung einzelner Variablen wird auf Abschnitt 3.4 verwiesen. Ausgehend von der Hypothese beziehungsweise der Fragestellung sowie der Variablenskalierung wurden bestimmte statistische Testverfahren zur Überprüfung herangezogen (s. Tabelle 3.3). Als signifikant gilt in dieser Studie ein Wert von $p \leq .05$ (Field, 2013). Sämtliche Hypothesen H1–H4 sind gerichtet, sodass einseitige Signifikanzwerte betrachtet werden. Für die Analysen wurde das Programm *IBM SPSS Statistics* (Version 28) eingesetzt.

Tabelle 3.3. Auflistung der eingesetzten Testverfahren für die Überprüfung und Untersuchung der Hypothesen (H) und Fragestellungen (F)

Testverfahren	Hypothese/Fragestellung
Logistische Regression (binär)	H1, H3, F3, F6
ANOVA	H2, H4
MANOVA	F1
Kruskal-Wallis-Test	F2
Multiple Regression	F4
Logistische Regression (multinominal)	F5

Anmerkungen. Die Auswahl der Testverfahren basiert auf Field (2013), Peters und Dörfler (2019a) sowie Tabachnick und Fidell (2013). Die Nummerierung der Hs und Fs bezieht sich auf die Auflistung in Abschnitt 2.5.

In den sich darauf beziehenden Abschnitten wird dargestellt, welche Voraussetzungen für das jeweilige Testverfahren überprüft und welche Maßnahmen bei Bedarf ergriffen wurden (s. 3.8.1–3.8.5). Zur Übersichtlichkeit findet diese Darstellung sortiert nach Testverfahren, aber getrennt nach Hypothesen und Fragestellungen statt. Eine Übersicht über die eingesetzten Effektstärken wird in Abschnitt 3.8.6 bereitgestellt.

3.8.1 Binäre logistische Regression

Die Hypothesen H1 und H3 sowie die Fragestellungen F3 und F6 wurden mittels binärer logistischer Regression untersucht, da sich deren AV „Tatverdacht“ auf zwei Level aufteilt. Für dieses Testverfahren sind insbesondere das Vorhandensein von Ausreißern (bei metrischen Variablen) und von kleinen beziehungsweise leeren Zellen relevant (Eid et al., 2017; Field, 2013). Die notwendige Unabhängigkeit der Fehler für F3 und F6 war gegeben. Im Rahmen des Manipulationschecks wurden Personen identifiziert, die als Ausreißer in den Zeitmessungen galten (s. 4.1). Da die hier relevante AV „Tatverdacht“ unter dem Einfluss der Manipulation erhoben worden war, wurden die Angaben ebenjener Personen für diese Analysen nicht berücksichtigt.[34]

3.8.1.1 Hypothese H1

Mittels Kontingenztabelle konnte festgestellt werden, dass keine der Zellen leer stand beziehungsweise nicht mehr als 20% der Zellen einen Wert < 5 aufwiesen. In lediglich zwei Zellen kam es zu kleineren Anzahlen. Die Voraussetzungen für eine binäre logistische Regression waren somit erfüllt.

3.8.1.2 Hypothese H3

Ursprünglich teilte sich die AV „nächster Verfahrensschritt“ auf drei Level auf. Da die Frage nach dem nächsten Verfahrensschritt nur von der Teilstichprobe beantwortet wurde, die zuvor den Tatverdacht bejaht hatte (128 von 299 Personen), blieben aufgrund der Anzahl an Prädiktoren einzelne Zellen der Kontingenztabelle leer beziehungsweise mehr als 20% der Zellen wiesen Werte < 5 auf. Um dennoch die Voraussetzung für das Testverfahren zu erfüllen, wurden folgende Level der AV

[34] Für die darüberhinausgehenden Analysen blieben diese Personen integriert, da die folgenden untersuchten AVs unabhängig von der Zeitdruckmanipulation erhoben wurden.

zusammengefügt: „Einstellung unter Auflagen“ und „Einstellung wegen Geringfügigkeit“. Da beide Optionen mit einer Einstellung einhergehen (und die Anklage dagegen mit der Eröffnung eines Hauptverfahrens), ist dies inhaltlich gut begründet. Folglich findet ein Vergleich hinsichtlich der Entscheidung „Anklage“ und „Einstellung“ statt, sodass es sich um das Verfahren einer binären logistischen Regression handelt.

Nach dem Zusammenfügen zweier Level blieb weiterhin eine Zelle wertfrei. Insgesamt lag in 11 von 24 Zellen der Wert unterhalb von 5 (45.8%). Die weitere Kombination von Variablen oder das Nacherheben von Daten war inhaltlich nicht sinnvoll oder umsetzbar (s. auch Eid et al., 2017). Auch eine Zusammenlegung von Noviz:innen und Expert:innen wäre aufgrund der grundlegenden Rolle, die die Unterscheidung von Expertise-Ausprägungen in der Studie spielt, nicht zielführend. Es wurde sich für Bootstrapping entschieden, um den problematischen Zellgrößen mit dieser Schätzmethode entgegen zu wirken (Field, 2013; Langeheine et al., 1996; Lin et al., 2015).

3.8.1.3 Fragestellung F3

Mittels Kontingenztabelle konnte für die kategorischen Prädiktoren festgestellt werden, dass keine der Zellen leer stand beziehungsweise nicht mehr als 20% der Zellen einen Wert < 5 aufwiesen. Die metrischen Prädiktoren *kognitive Reflexion, Need for Cognition* und *Lesezeit* wurden bereits von Ausreißern befreit, sofern dies notwendig war (s. 3.7). Die Überprüfung der Linearität bezieht sich auf die drei metrischen Prädiktoren und deren lineare Beziehung zum Logit der AV (Eid et al., 2017; Field, 2013). Dazu wurde für jeden der Prädiktoren die jeweilige Ln-Variable erstellt und die Interaktion des Prädiktors mit der eigenen Logit-Variable auf Signifikanz hin überprüft. Da keine der Interaktionen den Wert von $p \leq .05$ aufwies, war die Voraussetzung für die Linearität des Logits erfüllt (Field, 2013). Homoskedastizität wurde mittels Levene-Test untersucht und konnte für die drei Prädiktoren angenommen werden: Lesezeit, $F(1, 285) = 0.017$, $p = .895$, kognitive Reflexion, $F(1, 235) = 0.009$, $p = .926$, sowie NFC, $F(1, 245) = 1.166$, $p = .281$. Die Multikollinearitätsdiagnose ergab, dass die Toleranz- und die VIF-Statistiken für jeden Prädiktor unauffällig waren (Field, 2013). Die Toleranz-Statistiken lagen zwischen 0.956 und 0.991, die VIF-Statistiken waren nahezu 1 (zwischen 1.009 und

1.046). Betrachtet man zudem die Varianzanteile der jeweiligen Prädiktoren, so erklärten keine zwei (oder mehr) der Prädiktoren hohe Varianzanteile in der AV. Daher wurde nicht von (Multi-)Kollinearität zwischen den Prädiktoren ausgegangen. Die Voraussetzungen für eine binäre logistische Regression waren erfüllt.

3.8.1.4 Fragestellung F6

Zur besseren Überschaubarkeit wurden die Kontingenztabellen für das Beweismittel mit niedrigster beziehungsweise höchster Relevanz getrennt betrachtet. Für das Beweismittel mit niedrigster Relevanz waren nicht alle Kombinationen der Prädiktoren *Expertise*, *Zeitdruck* und *Delikt* für beide Level der AV vertreten, sodass letztlich 83 Zellen (anstatt der insgesamt möglichen 96 Zellen) inspiziert wurden. Mittels Kontingenztabelle konnte festgestellt werden, dass insgesamt 13 der Zellen leer standen beziehungsweise insgesamt 76% der Zellen einen Wert < 5 aufwiesen. Für das Beweismittel mit der höchsten Relevanz zeichnete sich ein ähnliches Bild ab. Betrachtet man die vorhandenen 77 Zellen (anstatt 96), so konnte festgestellt werden, dass insgesamt 13 der Zellen leer standen beziehungsweise insgesamt 75% der Zellen einen Wert < 5 aufwiesen. Die Tatsache, dass nicht alle Kombinationen vertreten waren beziehungsweise die Werte einzelner Zellen klein blieben, war aufgrund der großen Anzahl an möglichen Subgruppen zu erwarten (Agresti, 2007). Mit ausreichender Stichprobengröße hätten leere Zellen gefüllt beziehungsweise kleine Werte vergrößert werden können, denn generell war jede Zellenkombination möglich gewesen (Eid et al., 2017; Oyeyemi & Mbaeyi, 2018). Ebenso wie bei der Berechnung der H3 wurde sich für Bootstrapping entschieden, da eine Zusammenlegung von Subgruppen inhaltlich nicht sinnvoll erschien (s. 3.8.1.2).

3.8.2 ANOVA und Kruskal-Wallis-Test

Die Hypothesen H2 und H4 wurden mittels unabhängiger 2x2x3-faktorieller ANOVA untersucht. Für dieses Testverfahren sind das Vorhandensein von Ausreißern, die Normalverteilung der Daten, die Homogenität der Varianzen sowie die Unabhängigkeit der Variablen relevant (Field, 2013). Letzteres war für beide Hypothesen gegeben. Zur Überprüfung der Normalverteilung kamen die visuelle Inspektion von Histogrammen und Q-Q-Diagrammen, die Berechnung von Schiefe und Kurtosis und von deren *z*-Werten sowie der Kolmogorov-Smirnov-Test zum

Einsatz. Für die Berechnung der *z*-Werte wurden die Werte der Schiefe und Kurtosis durch ihre jeweiligen Standardfehler geteilt (Field, 2013; Peters & Dörfler, 2019a). Ein Wert von $z \geq 1.96$ galt als auffällige, signifikante Abweichung ($p \leq .05$; Field, 2013). Ein nicht signifikanter Kolmogorov-Smirnov-Test weist auf die Normalverteilung der Daten hin und ein nicht signifikanter Levene-Test spricht für die Homogenität der Varianzen (Field, 2013).

3.8.2.1 Hypothese H2

An anderer Stelle wurde bereits angemerkt, dass bei der Inspektion der Variable „Beweismaß" keine auffälligen Werte identifiziert werden konnten (s. 3.7). Mit Blick auf die verschiedenen Untergruppen der UVs wiesen die meisten der Analysen auf nicht normalverteile Daten hin. Die Histogramme und die Q-Q-Diagramme zeigten Abweichungen für Naive, Noviz:innen, Expert:innen, Teilnehmende ohne und mit Zeitdruck sowie für beide Vignetten. Der Kolmogorov-Smirnov-Test war für die Expert:innen zwar nur knapp auffällig ($p = .048$), deutete aber für die übrigen Gruppen signifikante Abweichungen an. Die *z*-Werte der Kurtosis waren nur für Expert:innen und Teilnehmende mit Zeitdruck auffällig. Die ANOVA ist robust gegen Abweichungen der Normalität (Lakens & Caldwell, 2021; Nguyen et al., 2019). Die Homogenität der Varianzen konnte laut dem Levene-Test angenommen werden, $F(11, 245) = 1.531$, $p = .121$. Die Voraussetzungen für eine faktorielle ANOVA waren ausreichend erfüllt.

3.8.2.2 Hypothese H4

In Abschnitt 3.7 wurde bereits angemerkt, dass die Variable „Schweregrad" von einem Ausreißer befreit wurde. Histogramme, Q-Q-Diagramme und der Kolmogorov-Smirnov-Test deuteten für alle Untergruppen auf nicht normalverteilte Daten hin. Laut Lakens und Caldwell (2021) und Nguyen et al. (2019) ist die ANOVA aber robust gegen Abweichungen der Normalität. Der *z*-Wert der Schiefe war nur für die Laien unauffällig. Der *z*-Wert der Kurtosis überschritt bei den Expert:innen sowie bei der Gruppe ohne Zeitdruck den kritischen Wert von 1.96. Der Levene-Test ergab das Vorhandensein heterogener Varianzen, $F(11, 244) = 1.917$, $p = .038$.

Obwohl die Voraussetzungen demnach nicht erfüllt waren, wurde eine faktorielle ANOVA berechnet.[35]

3.8.2.3 Fragestellung F2 (Kruskal-Wallis-Test)

Im Rahmen der Präregistrierung wurde die Überprüfung der F2 mittels MANOVA angegeben (Ruppenthal, 2022). Allerdings wurde sich aufgrund der konzeptuell nicht allzu stark ausgeprägten Nähe der beiden Konstrukte *kognitive Reflexion* und *Need for Cognition* für eine separate Varianzanalyse der jeweiligen AV entschieden (s. auch Field, 2013) – zumal es sich hierbei um eine exploratorische Untersuchung handelt.

Für die Skala der kognitiven Reflexion wurde kein Ausreißer identifiziert (s. 3.7). Die Analysen deuteten für sämtliche Untergruppen auf nicht normalverteilte Daten hin. Aufgrund des geringen Punktbereichs zeigten sich sehr rechtssteile Verteilungen, für die der Kolmogorov-Smirnov-Test hochsignifikant war (ps < .001). Lediglich der z-Wert der Kurtosis für die Fachpersonen war unauffällig. Die Varianzgleichheit konnte angenommen werden, $F(2, 217) = 0.549$, $p = .578$. Aufgrund der nicht normalverteilten Daten wurde sich für das Durchführen des Kruskal-Wallis-Tests entschieden.

Im Zusammenhang mit der Dateninspektion wurde die *Need-for-Cognition*-Kurzskala bereits von Ausreißern bereinigt (s. 3.7). Die Q-Q-Diagramme zeigten nur geringe Abweichungen. Ebenso waren die Histogramme eher unauffällig, wenngleich bei den Laien und Expert:innen die Verteilung leicht linksschief war. Der Kolmogorv-Smirnov-Test deutete für alle Untergruppen auf nicht normalverteilte Daten hin. Die Homogenität der Varianzen war gegeben, $F(2, 250) = 0.652$, $p = .522$. Auch hier wurde sich für den Kruskal-Wallis-Test entschieden.

[35] Da die Voraussetzungen für die faktorielle ANOVA nicht vollständig gegeben waren, wurde mittels der Software *R* eine robuste Analyse nach Mair und Wilcox (2020) gerechnet, um zu überprüfen, inwiefern sich die Ergebnisse unterscheiden. Die Expertise war weiterhin als einziger Faktor signifikant ($p = .001$, einseitig). Da somit qualitativ keine Unterschiede in den Signifikanzen festgestellt werden konnten, wurde sich für die Berechnung der faktoriellen ANOVA entschieden.

3.8.3 MANOVA

Die Fragestellung F1 wurde mittels 2x2x3-faktorieller MANOVA untersucht. Für dieses Testverfahren sind die multivariate Normalität der Residuen sowie die Homogenität der Kovarianzmatrizen relevant (Field, 2013). Die zudem notwendige Unabhängigkeit der Fehler war gegeben. Für keine der drei AVs (Sicherheit, Leichtigkeit, Überzeugung) lagen Ausreißer vor (s. 3.7). Um die Normalität der standardisierten Residuen zu überprüfen, wurden diese separat für jede AV analysiert (Histogramme, Q-Q-Diagramme sowie Kolmogorov-Smirnov-Test; s. 3.8.2). Die Histogramme stellten eher breite Verteilungen dar. Die Q-Q-Diagramme zeigten leichte Abweichungen. Auch der Kolmogorov-Smirnov-Test deutete auf nicht normalverteilte Residuen hin (ps < .05). Laut Levene-Test konnte von homogenen Varianzen ausgegangen werden: Leichtigkeit, $F(11, 255) = 0.441$, $p = .936$, Sicherheit, $F(11, 255) = 0.349$, $p = .973$, und Überzeugung, $F(11, 255) = 0.908$, $p = .533$. Die Homogenität der Kovarianzmatrizen wurde mittels Box-Test überprüft und konnte ebenfalls angenommen werden ($p = .246$). Da besagte Homogenität gegeben war und die Teststatistiken der MANOVA auch bei nicht gegebener Normalität robust sind (Field, 2013; H. Finch, 2005), wurde sich für keine weiteren Maßnahmen, sondern für den Einsatz der Teststatistik Pillai-Spur entschieden.

3.8.4 Multiple Regression

Die Fragestellung F4 wurde mittels multipler Regression untersucht. Für dieses Testverfahren sind insbesondere das Vorhandensein von Ausreißern und von fehlenden Werten relevant (Field, 2013). Ausreißer waren lediglich für die metrischen Prädiktoren *kognitive Reflexion, Need for Cognition* und *Lesezeit* zu erwarten. Diese wurden bereits überprüft (s. 3.7), sodass diesem Modell der bereits bereinigte Datensatz zugrunde lag. Mittels Kontingenztabelle wurden die Zellgrößen für die kategorischen Prädiktoren im Hinblick auf die AV untersucht. In rund 48% der Fälle standen die Zellen leer beziehungsweise wiesen einen Wert < 5 auf. Ebenso wie in den Fällen der binären logistischen Regression (H3 und F6; s. 3.8.1.2; 3.8.1.4) war diese Unterbelegung einzelner Zellen zu erwarten, sodass sich bei der Berechnung der F4 für Bootstrapping entschieden wurde. Um sicherzustellen, dass die Fehlerterme unabhängig waren und keine Autokorrelation vorlag, wurde der Durbin-Watson-Test durchgeführt. Da diese Statistik mit einem Wert von 1.998

nahe dem Wert von 2 war, konnte von der Unabhängigkeit der Fehlerterme ausgegangen werden (Field, 2013). Homoskedastizität konnte für die metrischen Prädiktoren angenommen werden: Lesezeit, $F(3, 249) = 0.406$, $p = .749$, kognitive Reflexion, $F(3, 216) = 0.687$, $p = .561$, und NFC, $F(3, 249) = 0.341$, $p = .796$. Die Multikollinearitätsdiagnose ergab, dass die Toleranz- und die VIF-Statistiken für jeden Prädiktor unauffällig waren (Field, 2013). Die Toleranz-Statistiken lagen zwischen 0.612 und 0.984 und somit weit über einem kritischen Wert von 0.1. Die durchschnittliche VIF-Statistik lag mit einem Wert von 1.2 nahe dem Richtwert 1. Zudem zeigte die Korrelationsmatrix keine auffällig hohen Korrelationen zwischen den Prädiktoren ($r > .9$; Field, 2013). Betrachtet man zudem die Varianzanteile der jeweiligen Prädiktoren, so erklärten keine zwei (oder mehr) der Prädiktoren zeitgleich hohe Varianzanteile in der AV. Daher wird nicht von einer (Multi-)Kollinearität zwischen den Prädiktoren ausgegangen. Die Voraussetzungen für eine multiple Regression waren erfüllt.

3.8.5 Multinominale logistische Regression

Die Fragestellung F5 wurde mittels multinominaler logistischer Regression untersucht, da sich die kategorische AV „Beweismittel“ auf vier Level aufteilt. Aufgrund der Tatsache, dass es sowohl das Beweismittel mit höchster als auch mit niedrigster Relevanz zu untersuchen galt, wurden zwei Regressionsberechnungen durchgeführt. In beiden Fällen ist die Unabhängigkeit der Fehler erfüllt. Für dieses Testverfahren sind insbesondere das Vorhandensein von Ausreißern (bei metrischen Variablen) und von fehlenden Werten relevant (Field, 2013). Da es sich ausschließlich um kategorische Variablen handelt, stellen Ausreißer für die F5 kein Problem im eigentlichen Sinne dar. Mittels Kontingenztabelle konnte für das Beweismittel der höchsten Relevanz festgestellt werden, dass insgesamt fünf der Zellen leer standen beziehungsweise insgesamt rund 46% der Zellen einen Wert < 5 aufwiesen. Für das Beweismittel mit der niedrigsten Relevanz zeichnete sich das gleiche Bild ab. Mittels Kontingenztabelle konnte festgestellt werden, dass insgesamt fünf der Zellen leer standen beziehungsweise insgesamt rund 69% der Zellen einen Wert < 5 aufwiesen. Mit der gleichen Begründung wie für die Analyse der H3, F4 und F6 wurde in beiden Regressionsrechnungen auf Bootstrapping zurückgegriffen (s. 3.8.1.2; 3.8.4; 3.8.1.4).

3.8.6 Effektstärken

Effektstärken beschreiben ein Ausmaß, in dem ein Phänomen in einer Population vorhanden ist (Cohen, 1988): Je höher der Wert, desto größer ist das Ausmaß. Effektstärken ermöglichen eine Einschätzung der praktischen Relevanz eines Effektes, unabhängig von der statistischen Signifikanz (Ellis, 2010; Field, 2009; Fritz et al., 2012). Deren Standardisierung ermöglicht den Vergleich über Studien hinweg (Field, 2009; Peters & Dörfler, 2019b). Aus diesem Grund wurde auch in dieser Studie auf standardisierte Größen zurückgegriffen (s. Tabelle 3.4).

Tabelle 3.4. Auflistung der Effektstärken in Abhängigkeit des jeweiligen Testverfahrens

Testverfahren	Hypothese/ Fragestellung	Effektstärke	Interpretation		
			klein	mittel	groß
Logistische Regression (binär)	H1, H3, F3, F6	*OR*/Exp(B)	1.4	3.1	4.3
		Nagelkerkes R^2	.2	.4	.5
ANOVA	H2, H4	Partielles η^2 (η_p^2)	.01	.06	.14
MANOVA	F1	partielles η^2 (η_p^2)	.01	.06	.14
Kruskal-Wallis-Test	F2	η^2_H	.01	.06	.14
Multiple Regression	F4	f^2	.02	.15	.35
Logistische Regression (multinominal)	F5	Nagelkerkes R^2	.2	.4	.5
t-Test / Mann-Whitney-Test	Manipulations check	*r*	.1	.3	.5

Anmerkungen. OR = Odds Ratio. Es werden die jeweils relevanten Hypothesen (H) oder Fragestellungen (F) angegeben. Deren Nummerierung bezieht sich auf die Auflistung in Abschnitt 2.5. Die Auswahl der Effektstärken basiert auf Bortz und Schuster (2010), Cohen (1988, 1992), Ellis (2010), Field (2013), Lenhard und Lenhard (2017) sowie Leonhart (2017).

Um die Relevanz des Effekts – unabhängig von der Stichprobengröße – besser einschätzen zu können, wurde auch bei nicht signifikanten Ergebnissen dessen Stärke berechnet, sofern dies möglich war (Ellis, 2010; Fritz et al., 2012; Peters & Dörfler, 2019b). Die Ermittlung der Effektstärke steht in Abhängigkeit zum jeweiligen Testverfahren, wenngleich auch verschiedene Größen konvertiert und somit vergleichbar gemacht werden können (Fritz et al., 2012; Leonhart, 2017). Bei der Auswahl der Effektstärken wurde dahingehend von der Präregistrierung abgewichen, als dass aufgrund der Vielzahl an statistischen Analyseverfahren auf mehr Effektmaße zurückgegriffen wurde, als ursprünglich angedacht war (Ruppenthal, 2022). Dies wird aber als eine methodische Verbesserung gesehen.

4 Ergebnisse

Die Darstellung der Ergebnisse beginnt mit einem Manipulationscheck für den Faktor „Zeitdruck“ (s. 4.1). Darauf folgen die deskriptiven und inferenzstatistischen Ergebnisse der Hypothesen H1–H4 (s. 4.2– 4.5) sowie der Fragestellungen F1–F6 (s. 4.6–4.11).[36] Abschließend wird in einem Exkurs in Abschnitt 4.12 auf die Auswertung der qualitativen Daten zu Nachermittlungen und zur Entscheidungsqualität eingegangen.

4.1 Manipulationscheck für den Faktor „Zeitdruck“

Die Einschätzung der Normalverteilung erfolgte für die drei von Ausreißern befreiten Zeitmessungen (Lesezeit für die Vignette, Zeit zur Entscheidung über den Tatverdacht, Zeit zur Entscheidung über den nächsten Verfahrensschritt) jeweils getrennt nach den beiden Zeitdruck-Gruppen (sechs Untergruppen; s. 3.7). Es wurden die visuelle Inspektion von Histogrammen und Q-Q-Diagrammen, die Berechnung von Schiefe und Kurtosis und von deren z-Werten sowie der Kolmogorov-Smirnov-Test eingesetzt (Field, 2013). Zur Ermittlung der z-Werte wurden die Werte der Schiefe und Kurtosis durch ihre jeweiligen Standardfehler geteilt (Field, 2009; Peters & Dörfler, 2019a). Werte von $z \leq 1.96$ galten als nicht signifikante Abweichung, wohingegen Werte von $z \geq 1.96$ so interpretiert wurden, dass Schiefe oder Kurtosis signifikant von der Norm abweichen ($p \leq .05$). Für die sechs Untergruppen wurde deutlich, dass lediglich für die Lesezeit in beiden Zeitdruck-Gruppen eine nahezu normale Verteilung angenommen werden konnte. Auch wenn der Kolmogorov-Smirnov-Test jeweils signifikant ausfiel ($ps < .05$), wurde dies im Zusammenhang mit den anderen Analysen nicht als bemerkenswerte Abweichung der Norm angesehen (s. auch Peters & Dörfler, 2019a). Im Gegensatz dazu deuteten die Ergebnisse für die Zeit zur Entscheidung über den Tatverdacht sowie für die Zeit zur Entscheidung über den nächsten Verfahrensschritt in beiden Zeitdruck-Gruppen jeweils auf eine deutlichere Abweichung von der Normalverteilung hin. Hier war es einheitlich, dass die Verteilungen jeweils eine (starke) rechtsseitige Schiefe und

[36] Während im Methodenteil bei der Überprüfung der Voraussetzungen die Reihenfolge der Hypothesen und Fragestellungen mittels der eingesetzten Testverfahren bestimmt wurde (s. 3.8), so erfolgt die Darstellung der Ergebnisse wieder in der ursprünglichen Reihenfolge, die am Ende des Theorieteils abgeleitet worden war (s. 2.5).

eine (sehr) spitze Form aufwiesen, die laut den ermittelten *z*-Werten signifikant von der Norm abwichen. Dies bestätigte jeweils auch der Kolmogorov-Smirnov-Test (*p*s < .05). Die Mehrheit der Proband:innen benötigte wenig Zeit um Entscheidungen zu treffen, sodass sich derartige rechtsschiefe, linkssteile und spitze Verteilungen ergaben. Dies ist ein typisches Verteilungsmuster für Reaktionszeiten (Peters & Dörfler, 2019a) – sofern man die Zeit für das Treffen einer Entscheidung als eine solche Reaktion wertet. Die Untersuchungen zur Homogenität der Varianzen bezogen sich ebenfalls auf die von Ausreißern befreiten Zeitmessungen. Für die Lesezeit, $F(1, 129) = 2$, $p = .16$, sowie für die Zeit zur Entscheidung über den nächsten Verfahrensschritt, $F(1, 129) = 0.002$, $p = .96$, war von homogen verteilten Varianzen auszugehen (Field, 2013). Für die Entscheidungszeit über den Tatverdacht war dies laut Levene-Test nicht der Fall, $F(1, 129) = 4.5$, $p = .04$.

Für die Lesezeit waren die Normalverteilung und die Homogenität der Varianzen gegeben, sodass zur Überprüfung der Gruppenunterschiede ein unabhängiger *t*-Test durchgeführt wurde. Für die beiden anderen Zeitmessungen erfolgte wegen der für einen *t*-Test nicht erfüllten Voraussetzungen ein Mann-Whitney-Test. Aufgrund der starken Abweichung von der Normalverteilung betraf dies auch die gemessene Zeit zur Entscheidung über den nächsten Verfahrensschritt, obwohl deren Varianzverteilung als unbedenklich galt. Betrachtet man die deskriptiven Werte (Sekunden), so unterscheiden sich die Experimentalgruppen erwartungsgemäß: Personen mit Zeitdruck ($M = 103.24$, $SD = 38.67$) benötigten für das Lesen weniger Zeit als Personen ohne Zeitdruck ($M = 109.46$, $SD = 52.57$). Dieser Unterschied, der im Mittel lediglich rund 6 Sekunden betrug, war bei gleichen Varianzen nicht signifikant, $t(291) = -1.15$, $p = .13$. Das *p*-Level bezieht sich hierbei auf eine gerichtete Hypothese (einseitig). Die Effektgröße $r = .07$ war als klein einzustufen (s. 3.8.6). Personen unter Zeitstress (mittlerer Rang = 130.01, $M = 6.8$, $SD = .36$) benötigten signifikant weniger Zeit, um über das Vorliegen des Tatverdachts zu entscheiden als Personen, die ohne Zeitdruck (mittlerer Rang = 160.82, $M = 9.75$, $SD = .73$) agierten, $U = 12808.00$, $p = .001$ (einseitig). Die anhand des *z*-Wertes und der Stichprobengröße ermittelte Effektgröße $r = .18$ war als klein einzustufen. Mit Blick auf die Teilstichprobe ($N = 136$) derjenigen, die sich für den nächsten Verfahrensschritt entscheiden sollten, zeigt sich auf deskriptiver Ebene, dass diejenigen

unter Zeitdruck (mittlerer Rang = 66.03, M = 11.77, SD = 1.08) weniger Zeit benötigten als diejenigen, die nicht unter Zeitdruck handelten (mittlerer Rang = 71.37, M = 12.57, SD = 1.02). Dieser Unterschied war aber nicht signifikant (U = 2480.00, p = .22, einseitig) und auch dieser Effekt erwies sich als klein (r = .08). Die Zeitdruckmanipulation war somit nur für die für diese Studie zentrale Entscheidung über den Tatverdacht erfolgreich, aber nicht für die Lesezeit der Vignette oder für die Entscheidung über den weiteren Verfahrensschritt, wenngleich die deskriptiven Werte jeweils in die erwartete Richtung gingen.

4.2 H1 (binäre logistische Regression)

Laut H1 zeigt sich ein Haupteffekt von Expertise auf die Entscheidung hinsichtlich des Tatverdachts in Interaktion mit Zeitdruck, unabhängig vom Delikttyp: Laien bejahen den Tatverdacht eher als Noviz:innen und Expert:innen, insbesondere unter Zeitdruck.

4.2.1 Deskriptive Auswertung der H1

Die Tabelle 4.1 liefert deskriptive Erkenntnisse hinsichtlich der Verteilung der Entscheidungen über das Vorliegen des Tatverdachts. Betrachtet man das Delikt „Diebstahl“, so zeigten sich konträre Ergebnisse für die Laien und Expert:innen: Während rund 65.52% der befragten Laien den Tatverdacht bejahten, so verneinten ihn rund 73.68% der Expert:innen. Die Entscheidung der Noviz:innen hielt sich nahezu die Waage: Rund 55.1% verneinten und standen damit auf der Seite der Expert:innen. Unterscheidet man nach Zeitdruck-Gruppen, so zeigte sich ein ähnliches Bild: Während Expert:innen mit teils deutlicher Mehrheit den Tatverdacht verneinten, bestätigten Laien ihn, und zwar in beiden Zeitdruckmanipulationen. Die Noviz:innen verteilten sich insbesondere in der Manipulationsgruppe „mit Zeitdruck“ nahezu gleich auf „Ja“ und „Nein“. Während Laien und Expert:innen unabhängig vom Zeitdruck jeweils eine (konträre) Tendenz aufwiesen, war dies bei den Noviz:innen nicht erkennbar, wenngleich deren Tendenz zur Verneinung den Fachpersonen ähnelte.

Tabelle 4.1. Kontingenztabelle für die Beantwortung der Frage nach dem Vorliegen des Tatverdachtes in Abhängigkeit von Expertise, Zeitdruck und Delikt

			Tatverdacht (*N*)		
			Nein	Ja	Gesamt
Delikt	Zeitdruck	Expertise			
Diebstahl	ohne	L	6	10	16
		N	13	9	22
		E	14	4	18
	mit	L	4	9	13
		N	14	13	27
		E	14	6	20
	Gesamt	L	10	19	29
		N	27	22	49
		E	28	10	38
		Gesamt	65	51	116
Körperverletzung	ohne	L	5	11	16
		N	11	19	30
		E	17	10	27
	mit	L	5	15	20
		N	10	20	30
		E	20	10	30
	Gesamt	L	10	26	36
		N	21	39	60
		E	37	20	57
		Gesamt	68	85	153
Gesamt	ohne	L	11	21	32
		N	24	28	52
		E	31	14	45
	mit	L	9	24	33
		N	24	33	57
		E	34	16	50
	Gesamt	L	20	45	65
		N	48	61	109
		E	65	30	95
		Gesamt	133	136	269

Anmerkungen. L = Laien, N = Noviz:innen, E = Expert:innen.

Betrachtet man das Delikt „Körperverletzung“, setzte sich der Trend der Antworten der Laien und Expert:innen fort: Während rund 72.22% der Laien den Tatverdacht bejahten, so verneinten ihn rund 64.91% der Expert:innen. Dieser Trend zur Zwei-Drittel-Mehrheit fand sich jeweils auch bei Unterscheidung nach Zeitdruck-Gruppen. Anders als beim Delikt „Diebstahl“ verhielt es sich hier allerdings bei den Noviz:innen: Die Mehrheit bejahte den Tatverdacht und beschreibt somit ein Antwortmuster, das den Laien ähnelte. Auch hier war es in etwa ein Zwei-Drittel-Unterschied (Bejahen: 65%), der sich auch in den Zeitdruck-Gruppen zeigte.

Insgesamt wurde der Tatverdacht im Fall der Körperverletzung (knapp) mehrheitlich bejaht (55.56%), wohingegen er im Fall des Diebstahls (knapp) mehrheitlich verneint wurde (56.03%). Fasst man allerdings beide Delikttypen zusammen und betrachtet die Gesamtheit der Ergebnisse auf Grundlage der Expertise und des Zeitdrucks, verfestigt sich das bereits beschriebene Bild. Rund zwei Drittel der Laien bejahten den Tatverdacht (69.23%), wobei sich die Tendenz unter Zeitdruck noch verstärkte (72.73%; ohne Zeitdruck: 65.63%). Im Gegensatz dazu verneinte eine vergleichbare Mehrheit an Expert:innen diesen (68.42%). Die Zeitdruckmanipulation führte deskriptiv zu keinerlei Unterschieden (jeweils rund 69%). Auf die Gesamtheit der Noviz:innen bezogen bejahten knapp die Mehrheit (55.96%) deliktübergreifend den Tatverdacht und ähnelten dabei der Entscheidung der Laien. Die Unterscheidung nach Zeitdruck-Gruppen zeigte eine leichte Verstärkung der Tendenz zur Bejahung für diejenigen mit Zeitdruck (57.89%) im Vergleich zu denjenigen ohne Zeitdruck (53.85%).

4.2.2 Inferenzstatistische Auswertung der H1

Die Voraussetzungen für die binäre logistische Regression waren erfüllt (s. 3.8.1.1). Als Referenzgruppen für die Vergleiche je Prädiktor wurden die Laien, der Diebstahl sowie die Manipulation „ohne Zeitdruck“ ausgewählt.[37] Die Prädiktoren wurden mittels „Einschluss“-Variante dem Modell hinzugefügt. Zunächst wurden die Residuen untersucht, um das Vorhandensein einflussreicher Fälle und die Passung zwischen Daten und Modell zu überprüfen. Für eine gute Passung wurden folgende Kriterien angelegt (Field, 2013):

- Einflussstatistik nach Cook < 1,
- Hebelwerte der einzelnen Proband:innen sind nicht dreimal höher als der durchschnittliche Hebelwert,
- standardisierte Residuen ≤ 1.96 und
- DFBeta-Werte < 1.

Die Einflussstatistik nach Cook sowie die DFBeta-Werte für die Konstante sowie das Modell lagen für alle Fälle jeweils unter einem Wert von 1. Der durchschnittlich zu erwartende Hebelwert betrug den Wert .015. Der Maximalwert in der Stichprobe lag bei .035, sodass der Wert in keinem Fall der dreifachen Höhe des Durchschnitts

[37] Die Laien wurden als Referenzkategorie gewählt, da sie aufgrund nicht vorhandener juristischer Vorbildung eine Art Kontrollgruppe darstellen. Diese Begründung gilt auch für die weiteren Analysen.

entsprach. Der maximale Wert der standardisierten Residuen betrug 1.78 und galt als unauffällig. Anhand dieser Kriterien konnte sichergestellt werden, dass keine Ausreißer oder andere stark beeinflussenden Fälle im Regressionsmodell vorhanden waren (Field, 2013).

Fügte man die Prädiktoren und deren Interaktionen dem Modell hinzu, so ließ sich eine signifikante Verbesserung ablesen ($p < .001$). Die anfängliche -2-Log-Likelihood reduzierte sich von 372.88 im Null-Modell auf 342.41 im Modell 1. Der Gesamtprozentsatz der Richtigen stieg von 50.4% im Nullmodell auf 65.4% in Modell 1. Dabei wurden in 69.2% der Fälle das Verneinen des Tatverdachts und in 61.8% der Fälle das Bejahen korrekt vorhergesagt. Der Hosmer-Lemeshow-Test war aufgrund der Nicht-Signifikanz von $p = 1$ ein Hinweis darauf, dass Modell und Daten gut zueinander passten.

Sowohl für Noviz:innen als auch für Expert:innen zeigte sich im Vergleich mit den Laien ein negativer Regressionskoeffizient (s. Tabelle 4.2). Da das Bejahen des Tatverdachts mit 1 kodiert wurde, kann abgeleitet werden, dass die Zugehörigkeit zu einer dieser beiden Expertise-Gruppen signifikant mit einer sinkenden Wahrscheinlichkeit einherging, den Tatverdacht zu bejahen (ps $< .05$). Das 95%-Konfidenzintervall für die Expert:innen blieb unter 1, wohingegen das Intervall der Noviz:innen den Wert von 1 überschritt. Letzterer Prädiktor konnte demnach nicht eindeutig angeben, welche Entscheidung getroffen wurde. Beide Effekte waren klein (s. auch Nagelkerkes R^2).

Der Prädiktor „Zeitdruck“ war nicht signifikant ($p = .242$) und änderte wenig am Entscheidungsverhalten der Proband:innen. Stand man unter Zeitdruck, so war laut Regressionskoeffizient die Wahrscheinlichkeit größer, den Tatverdacht zu bejahen, aber die Richtung des Entscheidungsergebnisses konnte nicht eindeutig angeben werden (Konfindenzintervall).

Der Prädiktor „Delikt“ erreichte das Signifikanzniveau ($p = .026$). Für den Delikttyp lässt sich laut Regressionskoeffizient ableiten, dass die Wahrscheinlichkeit für das Bejahen des Tatverdachts stieg, wenn es sich um das Delikt „Körperverletzung“ handelte. Das Konfidenzintervall lag mit dem unteren Wert zwar nahezu bei 1, aber trotzdem knapp darunter. Streng genommen konnte auch dieser Prädiktor keine eindeutige Entscheidungsrichtung vorgeben. Die Effektstärke war zudem klein.

Tabelle 4.2. Regressionsergebnisse der H1 für die Prädiktoren sowie deren Interaktion

Prädiktoren	*b*	SE *b*	*p**	Konfidenzintervall (95%) Unterer Wert	Oberer Wert	Exp(B)
Expertise(1)	-.608	.361	.046	.268	1.104	.544
Expertise(2)	-1.524	.389	<.001	.102	.467	.218
Zeitdruck(1)	.232	.331	.242	.66	2.411	1.261
Delikt(1)	.625	.321	.026	.997	3.507	1.868
Delikt(1)*Expertise(1)*Zeitdruck(1)	.075	.566	.447	.355	3.272	1.078
Delikt(1)*Expertise(2)*Zeitdruck(1)	-.395	.58	.248	.216	2.101	.674

Anmerkungen. R^2 = .14 (Nagelkerke), **p* einseitig, Expertise(1) = Noviz:innen, Expertise(2) = Expert:innen, Zeitdruck(1) = mit, Delikt(1) = Körperverletzung.

Als Referenzgruppe für die Interaktionen diente *Diebstahl*Laie*ohne Zeitdruck*. Für die Interaktion *Körperverletzung*Noviz:in*mit Zeitdruck* stieg die Wahrscheinlichkeit, den Tatverdacht zu bejahen, nur marginal. Im Gegensatz dazu sank die Wahrscheinlichkeit für die Gruppe *Körperverletzung*Expert:in*mit Zeitdruck*. Keine dieser Interaktionen erreichte das Signifikanzniveau (*p*s > .05) und in beiden Fällen konnte eine Entscheidungsrichtung nicht eindeutig anzugeben werden (Konfindenzintervalle).

Die H1 kann nur teilweise bestätigt werden. Es zeigte sich der erwartete Haupteffekt von Expertise, denn im Vergleich zu Laien bejahten Noviz:innen und Expert:innen mit geringerer Wahrscheinlichkeit den Tatverdacht – aber nicht in Interaktion mit Zeitdruck. Außerdem wurde entgegen der H1 ein Haupteffekt des Deliktтyps festgestellt, da die Wahrscheinlichkeit für das Bejahen des Tatverdachts im Fall der Körperverletzung größer war.

4.3 H2 (ANOVA)

Laut H2 zeigt sich ein Haupteffekt von Expertise auf das Beweismaß, unabhängig von Zeitdruck und Deliktтyp: Die Beweislage überzeugt Laien mehr als Noviz:innen und Expert:innen.

4.3.1 Deskriptive Auswertung der H2

Wird die Gesamtstichprobe betrachtet, so zeigte sich hinsichtlich des Beweismaßes eine Tendenz zur Mitte (s. Tabelle 4.3). Dabei lag auf deskriptiver Ebene kein Unterschied zwischen den Delikten vor, sofern der Zeitdruck unberücksichtigt bleibt.

Beim Fehlen von Zeitdruck waren die Teilnehmenden im Fall der Körperverletzung minimal überzeugter von der Beweislage.

Tabelle 4.3. Mittelwerte und Standardabweichungen des Beweismaßes (0-99) in Abhängigkeit von Expertise, Zeitdruck und Delikt

			Beweismaß	
			M	*SD*
Expertise	Zeitdruck	Delikt (*N*)		
Laien	ohne	DS (16)	60.81	22.91
		KV (16)	65.44	24.74
	mit	DS (15)	61.53	26.53
		KV (19)	52.59	20.52
	Gesamt	DS (31)	61.16	24.31
		KV (35)	58.51	23.12
		Gesamt (66)	59.76	23.54
Noviz:innen	ohne	DS (22)	49.05	24.46
		KV (23)	53.13	22.92
	mit	DS (28)	50.5	21.06
		KV (29)	57.07	23.24
	Gesamt	DS (50)	49.86	22.39
		KV (52)	55.37	22.96
		Gesamt (102)	52.65	22.74
Expert:innen	ohne	DS (18)	55.00	27.31
		KV (24)	57.71	30.21
	mit	DS (21)	46.71	33.36
		KV (26)	48.31	31.34
	Gesamt	DS (39)	50.54	30.61
		KV (50)	52.82	30.85
		Gesamt (89)	51.82	30.59
Gesamt	ohne	DS (56)	54.32	25.01
		KV (63)	58.00	26.34
	mit	DS (64)	51.84	27.05
		KV (74)	52.87	25.72
	Gesamt	DS (120)	53.00	26.04
		KV (137)	55.23	26.06
		Gesamt (257)	54.19	26.02

Anmerkungen. DS = Diebstahl, KV = Körperverletzung.

Die beiden Gruppen mit juristischer Vorbildung schätzten das Beweismaß ähnlich ein, wohingegen die Laien etwas überzeugter waren. Betrachtet man nur ein Beweismaß von ≥ 90%, so gaben 13 Laien, 2 Noviz:innen und 9 Expert:innen einen solchen Wert an. Für die Expert:innen hatte der Delikttyp keine Auswirkungen, allerdings waren diejenigen unter Zeitdruck von beiden Beweislagen insgesamt weniger überzeugt als die Vergleichsgruppen ohne Zeitdruck. Zudem lagen in der

Gruppe der Expert:innen relativ große Standardabweichungen vor. Die Noviz:innen in beiden Zeitdruck-Gruppen gaben für die Körperverletzung ein etwas höheres Beweismaß an. Dies war auch für die Laien ohne Zeitdruck der Fall, wohingegen Laien unter Zeitdruck vom Fall des Diebstahls überzeugter waren.

4.3.2 Inferenzstatistische Auswertung der H2

Es zeigten sich keine signifikanten Unterschiede zwischen den Gruppen oder in den Interaktionen (*p*s > .05). Die Expertise verpasste das einseitige Signifikanzniveau nur knapp, $F(2, 245) = 2.251$, $p = .054$. Es wurden einfache Kontraste gerechnet, da laut Hypothese bereits Annahmen über die Unterschiede in den Gruppen vorlagen. Auch wenn das Signifikanzniveau nicht erreicht wurde, so deuteten sich nahezu signifikante Unterschiede zwischen den Laien und Expert:innen an ($p = .056$, 95% CI [-16.571, .202]; Laien und Noviz:innen: $p = .064$, 95% CI [-15.827, .465]). Die Kontraste für das Delikt und den Zeitdruck waren weit vom geforderten Niveau entfernt. Die ermittelten Effektstärken waren durchgehend als klein zu bewerten und lagen zwischen $\eta_p^2 = .001$ (Delikt) und $\eta_p^2 = .018$ (Expertise).

Die H2 kann somit nicht bestätigt werden, wenngleich sich – unabhängig von Zeitdruck und Delikttyp – ein Haupteffekt von Expertise in die erwartete Richtung andeutete: Laien waren von der Beweislage überzeugter als die Vergleichsgruppen.

4.4 H3 (binäre logistische Regression)

Laut H3 zeigt sich – wenn der Tatverdacht bejaht wurde – ein Haupteffekt von Expertise auf die Entscheidung hinsichtlich des nächsten Verfahrensschritts in Interaktion mit Zeitdruck, unabhängig vom Delikttyp: Laien entscheiden sich eher für eine Anklage als Noviz:innen und Expert:innen, insbesondere unter Zeitdruck.

4.4.1 Deskriptive Auswertung der H3

Die Teilstichprobe setzte sich aus 42 Laien, 59 Noviz:innen und 27 Expert:innen zusammen (s. Tabelle 4.4). Das Delikt „Diebstahl“ war in 51 Fällen vertreten, das Delikt „Körperverletzung“ in 77 Fällen. Die Manipulationsgruppen unterteilten sich in 73 Fälle mit und 55 Fälle ohne Zeitdruck.

Tabelle 4.4. Kontingenztabelle für die Beantwortung der Frage nach dem nächsten Verfahrensschritt in Abhängigkeit von Expertise, Zeitdruck und Delikt (Teilstichprobe)

			Nächster Verfahrensschritt		
			Einstellung	Anklage	Gesamt
Delikt	Zeitdruck	Expertise			
Diebstahl	ohne	L	7	2	9
		N	7	2	9
		E	2	2	4
	mit	L	4	5	9
		N	9	5	14
		E	3	3	6
	Gesamt	L	11	7	18
		N	16	7	23
		E	5	5	10
		Gesamt	32	19	51
Körperverletzung	ohne	L	7	2	9
		N	10	6	16
		E	8	0	8
	mit	L	12	3	15
		N	13	7	20
		E	5	4	9
	Gesamt	L	19	5	24
		N	23	13	36
		E	13	4	17
		Gesamt	55	22	77
Gesamt	ohne	L	14	4	18
		N	17	8	25
		E	10	2	12
	mit	L	16	8	24
		N	22	12	34
		E	8	7	15
	Gesamt	L	30	12	42
		N	39	20	59
		E	18	9	27
		Gesamt	87	41	128

Anmerkungen. L = Laien, N = Noviz:innen, E = Expert:innen.

Beim Delikt „Diebstahl“ entschieden sich die Proband:innen insgesamt mehrheitlich für eine Einstellung (63%). Ohne Zeitdruck setzte sich diese Tendenz fort, allerdings war die Entscheidung bei denjenigen mit Zeitdruck nahezu gleichverteilt (Einstellung: 55%). Aufgeschlüsselt nach Expertise-Gruppen wird deutlich, dass Expert:innen kein eindeutiges Entscheidungsmuster zeigten, unabhängig vom Zeitdruck (jeweils 50%). Laien und Noviz:innen entschieden sich ohne Zeitdruck mehrheitlich für die Einstellung (jeweils 78%); mit Zeitdruck unterschieden sich die Entscheidungen (44% der Laien beziehungsweise 64% der Noviz:innen bevorzugten eine Einstellung).

Für das Delikt „Körperverletzung“ zeigten sich ähnliche Muster, unabhängig von Expertise und Zeitdruck: Die Einstellung wurde in insgesamt 71% der Fälle gewählt. Dabei entschieden sich alle Expertise-Gruppen jeweils mehrheitlich für die Einstellung (Laien: 80%, Noviz:innen: 64%, Expert:innen: 77%). Die Aufschlüsselung nach Zeitdruck-Gruppen änderte an dieser Tendenz der Laien und Noviz:innen wenig. Mit Blick auf die Expert:innen zeigte sich allerdings nahezu eine Gleichverteilung, wenn diese unter Zeitdruck agierten (Einstellung: 56%), im Vergleich zum Agieren ohne Zeitdruck (Einstellung: 100%).

Unabhängig vom Delikt verdeutlichen sich die bereits beschriebenen Entscheidungstendenzen: Insgesamt 68% entschieden sich für eine Einstellung. Diese Mehrheit änderte sich weder für die Laien noch die Noviz:innen, unabhängig davon, ob sie unter Zeitdruck standen oder nicht. Standen Expert:innen dagegen unter Zeitdruck, verschob sich die Mehrheit, sodass sich lediglich 53% für eine Einstellung entschieden (ohne Zeitdruck: 83%).

4.4.2 Inferenzstatistische Auswertung der H3

Die Gruppe der Laien, die Manipulation „ohne Zeitdruck“ sowie das Delikt „Diebstahl“ galten wie bei der Berechnung der H1 als Referenzgruppen. Die Einstellung des Verfahrens wurde als Referenzkategorie festgelegt. Die Prädiktoren wurden gemäß der „Einschluss“-Methode dem Modell hinzugefügt. Zunächst wurden die Residuen untersucht, wofür die gleichen Kriterien wie bei der Berechnung der H1 eingesetzt wurden (s. 4.2.2). Die Einflussstatistik nach Cook sowie die DFBeta-Werte für die Konstante sowie das Modell lagen für alle Fälle jeweils unter 1. Daraus kann geschlossen werden, dass keine Fälle das Modell übermäßig beeinflussten. Der maximale Wert der standardisierten Residuen betrug 2.02 und galt somit als auffällig. Insgesamt acht Fälle überschritten den Wert von 1.96, was einem Anteil von 6.4% entspricht. Laut Field (2013) spricht die Nähe zu einem 5%-Wert aber dafür, dass die Passung zwischen Modell und Daten noch angemessen ist. Der durchschnittlich zu erwartende Hebelwert betrug den Wert .031. Insgesamt 30 Fälle überschritten einen Hebelwert von 0.093 (was der dreifachen Höhe des Durchschnitts entspricht). Diese Fälle schienen somit Ausreißer in Bezug auf die Prädiktoren zu sein. Laut Stevens (2002) müssen Ausreißer aber nicht zwingend aus dem Datensatz entfernt werden (s. auch Field, 2009). Aufgrund der bereits niedrigen Werte in einzelnen

Zellen wurde auf das Bereinigen der Daten verzichtet, da dies das Problem der kleinen Zellgrößen weiter verstärkt hätte. Die Analysen wiesen demnach darauf hin, dass weiterhin Ausreißer im Regressionsmodell vorhanden waren, wenngleich eine gewisse Passung zwischen Modell und Daten durchaus angenommen werden konnte.

Fügte man die oben genannten Prädiktoren und deren Interaktion dem Modell hinzu, so ließ sich keine signifikante Verbesserung des Modells ablesen (p = .66). Die anfängliche -2-Log-Likelihood reduzierte sich minimal von 160.54 im Null-Modell auf 156.412 im Modell 1. Der Gesamtprozentsatz der Richtigen blieb bei 68%. Dabei wurde jeweils in 100% der Fälle die Einstellung korrekt vorhergesagt. Der Hosmer-Lemeshow-Test war aufgrund der Nicht-Signifikanz (p =.636) ein Hinweis darauf, dass Modell und Daten gut zueinander passten.

Da die Voraussetzungen für das Testverfahren teilweise verletzt waren (s. 3.8.1.2), wurde sich für den Einsatz von Bootstrapping entschieden.[38] Kein Prädiktor und keine Interaktion erwiesen sich als signifikant (ps > .05; s. Tabelle 4.5).

Tabelle 4.5. Regressionsergebnisse der H3 für die Prädiktoren sowie deren Interaktion (1000 Bootstrapping-Stichproben)

Prädiktoren	*b*	SE *b*	*p**	BCa Konfidenzintervall (95%) Unterer Wert	Oberer Wert
Expertise(1)	.208	.575	.348	-.934	1.527
Expertise(2)	-.035	1.709	.47	-1.384	.904
Zeitdruck(1)	.399	.528	.207	-.66	1.499
Delikt(1)	-.567	.545	.13	-1.618	.422
Delikt(1)*Expertise(1)*Zeitdruck(1)	.183	1.253	.402	-1.57	1.823
Delikt(1)*Expertise(2)*Zeitdruck(1)	.822	3.258	.192	-1.43	3.34

Anmerkungen. R^2 = .044 (Nagelkerke), **p* einseitig, Expertise(1) = Noviz:innen, Expertise(2) = Expert:innen, Zeitdruck(1) = mit, Delikt(1) = Körperverletzung.

Mit Blick auf die Regressionskoeffizienten zeigte sich, dass – im Vergleich zu Laien – die Wahrscheinlichkeit für eine Anklage bei den Noviz:innen stieg, aber bei den Expert:innen (gemäß der Hypothese) leicht sank. Diejenigen mit Zeitdruck

[38] Bei wiederholtem Durchführen der Regression ändern sich die Werte gewisser Parameter aufgrund des Bootstrappings minimal (z. B. Standardfehler, p-Wert), da mit jeder neuen Durchführung neue Stichproben gezogen werden. Dies beeinträchtigt nicht die Robustheit der Methode (Field, 2013).

entschieden sich (erwartungsgemäß) mit größerer Wahrscheinlichkeit für eine Anklage als diejenigen ohne Zeitdruck. Für die Körperverletzung sank die Wahrscheinlichkeit für das Erheben einer Anklage im Vergleich zum Diebstahl.

Für die Interaktionen *Körperverletzung*Noviz:in*mit Zeitdruck* und *Körperverletzung*Expert:in*mit Zeitdruck* stieg die Wahrscheinlichkeit der Entscheidung für eine Anklage. Die Tatsache, dass die unteren und oberen Werte der BCa-Konfidenzintervalle (95%) für die Prädiktoren (außer für die Expert:innen und das Delikt) und jede Interaktion den Wert von 1 überschritten, spricht dafür, dass die Richtung der Entscheidung nicht eindeutig vorgegeben werden konnte. Die Effektstärken waren im kleinen Bereich zu verorten (s. auch Nagelkerkes R^2).

Die H3 kann somit nicht bestätigt werden. Laien äußerten auf deskriptiver wider Erwarten eine Tendenz zur Einstellung und hoben sich dadurch nicht von den Gruppen mit juristischer Vorbildung ab.

4.5 H4 (ANOVA)

Laut H4 zeigt sich ein Haupteffekt von Expertise auf den eingeschätzten Schweregrad der Delikte, unabhängig von Zeitdruck: Laien stufen den Schweregrad für das Delikt „Körperverletzung“ höher ein als den Schweregrad für das Delikt „Diebstahl“, wohingegen Noviz:innen und Expert:innen diese Tendenz nicht zeigen.

4.5.1 Deskriptive Auswertung der H4

Laut Gesamtstichprobe handelte es sich bei beiden Fällen um Delikte geringer Schwere (s. Tabelle 4.6). Unter Zeitdruck galt die Körperverletzung als etwas schwerer, wohingegen dies beim Fehlen des Zeitdrucks umgekehrt war, wenngleich in nur geringerem Ausmaß.

Tabelle 4.6. Mittelwerte und Standardabweichungen des Schweregrades (0-100) in Abhängigkeit von Expertise, Zeitdruck und Delikt

			Schweregrad	
			M	*SD*
Expertise	Zeitdruck	Delikt (*N*)		
Laien	ohne	DS (16)	30.13	14.67
		KV (16)	32.44	20.9
	mit	DS (15)	25.67	20.58
		KV (19)	38.53	21.62
	Gesamt	DS (31)	27.97	17.62
		KV (35)	35.74	21.21
		Gesamt (66)	32.09	19.84
Noviz:innen	ohne	DS (21)	27.43	17.18
		KV (23)	27.3	15.00
	mit	DS (28)	31.39	16.19
		KV (29)	33.41	17.09
	Gesamt	DS (49)	29.69	16.56
		KV (52)	30.71	16.34
		Gesamt (101)	30.22	16.37
Expert:innen	ohne	DS (18)	28.28	20.97
		KV (24)	20.08	10.57
	mit	DS (21)	24.33	18.87
		KV (26)	23.31	14.32
	Gesamt	DS (39)	26.15	19.7
		KV (50)	21.76	12.64
		Gesamt (89)	23.69	16.17
Gesamt	ohne	DS (55)	28.49	17.57
		KV (63)	25.86	15.87
	mit	DS (64)	27.73	18.18
		KV (74)	31.18	18.32
	Gesamt	DS (119)	28.04	17.83
		KV (137)	28.73	17.38
		Gesamt (256)	28.43	17.56

Anmerkungen. DS = Diebstahl, KV = Körperverletzung.

Naive und Noviz:innen schätzten den Schweregrad insgesamt ähnlich ein, aber die Expert:innen gaben geringere Werte an. Für die Fachpersonen war der Diebstahl das schwerere Delikt. Diese Tendenz zeigte sich deutlicher beim Fehlen von Zeitdruck. Für die Noviz:innen machten weder der Delikttyp noch die Zeitdruckmanipulation einen Unterschied aus. Insgesamt und insbesondere unter Zeitdruck schätzten die Laien die Körperverletzung als die schwerere Tat ein.

4.5.2 Inferenzstatistische Auswertung der H4

Für die verschiedenen Expertise-Gruppen ergaben sich signifikante Unterschiede, $F(2, 244) = 4.398$, $p = .007$, $\eta_p^2 = .035$ (s. Abbildung 4.1). Da laut Hypothese bereits

Annahmen über die Unterschiede der Gruppen vorlagen, wurden einfache Kontraste gerechnet. Laien und Expert:innen unterschieden sich signifikant in ihrer Einschätzung des Schweregrades (p = .007, 95% CI [-13.24, -2.136]). Für Laien und Noviz:innen galt dies nicht (p = .512, 95% CI [-7.211, 3.603]).

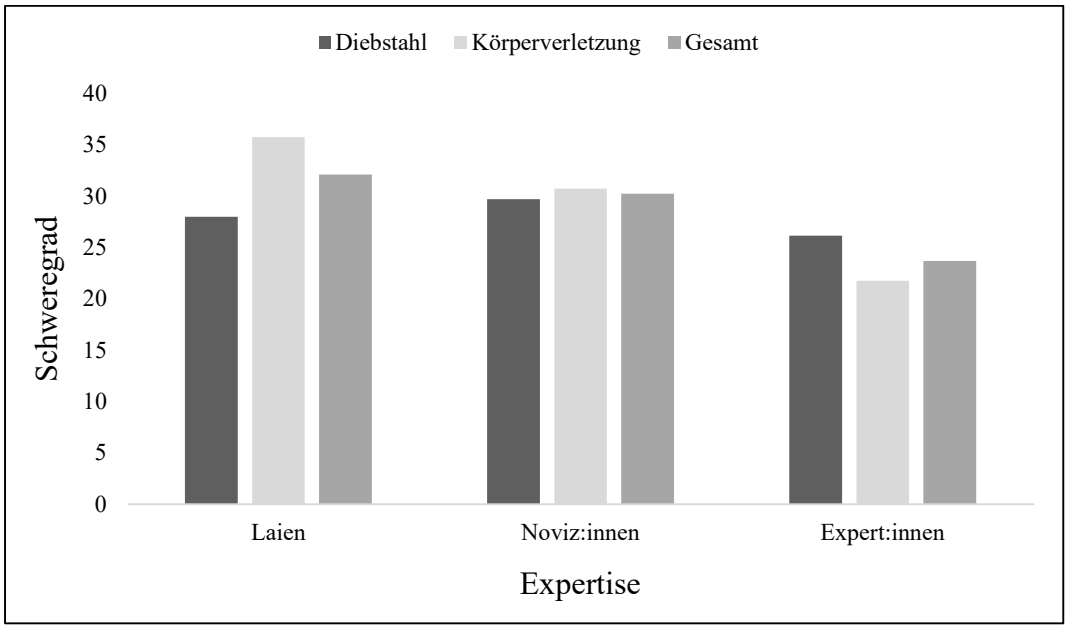

Abbildung 4.1. Darstellung der Mittelwerte der Expertise-Gruppen hinsichtlich des Schweregrades der Delikte (0-100).

Auch die Interaktion *Expertise*Delikt* erreichte noch knapp das einseitige Niveau, $F(2, 244) = 2.343$, $p = .049$, $\eta_p^2 = .019$ (s. Abbildung 4.2). Das Delikt, der Zeitdruck und die weiteren Interaktionen zeigten einseitige ps > .05. Die Effektstärken waren klein und lagen zwischen $\eta_p^2 = .001$ (Delikt) und $\eta_p^2 = .009$ (*Expertise*Delikt*).

Zusätzlich wurde die Korrelation des Schweregrades und des Beweismaßes ermittelt, welche signifikant und positiv war ($r = .126$, $p = .043$). Das Beweismaß wurde mit steigender Schwere des Delikts höher eingeschätzt beziehungsweise ein Delikt wurde als schwerer definiert, wenn das Beweismaß höher war.

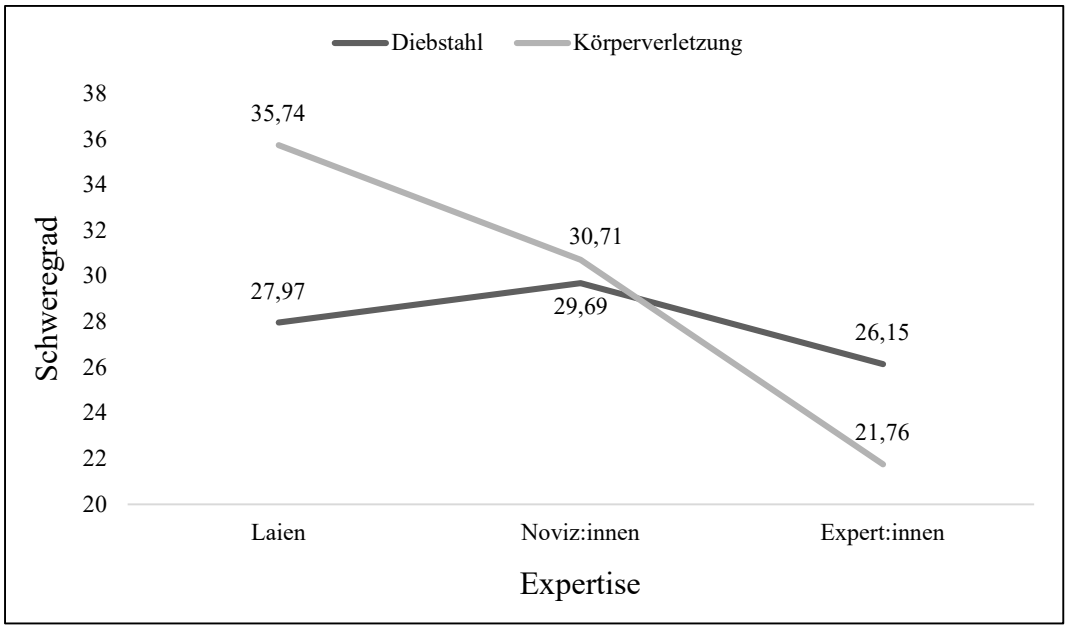

Abbildung 4.2. Darstellung der Interaktion *Expertise*Delikt* hinsichtlich des Schweregrades der Delikte (0-100).

Die H4 kann teilweise bestätigt werden. Es zeigte sich zwar – unabhängig vom Zeitdruck – die Interaktion von Expertise und Delikt dahingehend, dass Laien die Körperverletzung als schwerer erachteten als die Fachpersonen, allerdings galt dies nicht für Laien und Noviz:innen. Für das Delikt des Diebstahls ließen sich erwartungsgemäß keine Unterschiede feststellen.

4.6 F1 (MANOVA)

Die F1 befasst sich mit der explorativen Fragestellung, ob sich Laien, Noviz:innen und Expert:innen in den Prozessmerkmalen *Leichtigkeit*, *Sicherheit* und *Überzeugung* unterscheiden. Dabei ist der Prozess der Entscheidung über das Vorliegen des Tatverdachtes gemeint.

4.6.1 Deskriptive Auswertung der F1

Auf deskriptiver Ebene zeigte sich, dass Expert:innen jedes der drei Prozessmerkmale am höchsten bewerteten: Sie empfanden die Entscheidung als leichter, fühlten sich sicherer und überzeugter als Laien und Noviz:innen (s. Tabelle 4.7). Insbesondere zu den Laien waren vergleichsweise große Unterschiede in den Mittelwerten erkennbar.

Tabelle 4.7. Mittelwerte und Standardabweichungen der drei Prozessmerkmale auf einer 7-stufigen Likert-Skala in Abhängigkeit von Expertise, Zeitdruck und Delikt

Prozessmerkmal	Gruppenzuordnung (*N*)	*M*	*SD*
Leichtigkeit	Laien (67)	4.03	1.64
	Noviz:innen (106)	4.56	1.59
	Expert:innen (94)	5.00	1.69
	ohne Zeitdruck (124)	4.61	1.66
	mit Zeitdruck (143)	4.56	1.69
	Diebstahl (120)	4.68	1.68
	Körperverletzung (147)	4.5	1.67
	Gesamt (267)	4.58	1.67
Sicherheit	Laien (67)	3.51	1.6
	Noviz:innen (106)	4.57	1.62
	Expert:innen (94)	5.21	1.67
	ohne Zeitdruck (124)	4.63	1.8
	mit Zeitdruck (143)	4.44	1.71
	Diebstahl (120)	4.61	1.71
	Körperverletzung (147)	4.46	1.79
	Gesamt (267)	4.53	1.75
Überzeugung	Laien (67)	4.06	1.47
	Noviz:innen (106)	4.96	1.39
	Expert:innen (94)	5.78	1.48
	ohne Zeitdruck (124)	5.01	1.54
	mit Zeitdruck (143)	4.9	1.55
	Diebstahl (120)	5.12	1.5
	Körperverletzung (147)	4.82	1.58
	Gesamt (267)	4.95	1.55

Die Mittelwertunterschiede in den Zeitdruckmanipulationen waren marginal. Betrachtet man die Delikttypen, so wird deutlich, dass die drei Variablen für den Diebstahl jeweils höher eingeschätzt wurden. Gruppenübergreifend waren das Empfinden von Leichtigkeit und Sicherheit sowie das Gefühl von Überzeugung insgesamt vergleichbar stark ausgeprägt.

4.6.2 Inferenzstatistische Auswertung der F1

Weder der Zeitdruck noch der Delikttyp noch die Interaktionen erwiesen sich in den multivariaten Tests auf Grundlage der Pillai-Spur als signifikant (*p*s > .05). Es ergaben sich kleine Effektstärken zwischen $\eta_p^2 = .006$ (Zeitdruck) und $\eta_p^2 = .016$ (Delikt). Allerdings zeigte sich ein signifikanter, großer Effekt der Expertise auf die Prozessmerkmale, $V = 0.925$, $F(3, 253) = 1044.98$, $p < .001$, $\eta_p^2 = .085$ (s. Abbildung 4.3). Laut der ANOVA-Statistiken bezog sich der Effekt auf die Leichtigkeit, $F(2, 255) = 6.57$, $p = .002$, auf die Sicherheit, $F(2, 255) = 21.52$, $p < .001$, und auf die Überzeugung, $F(2, 255) = 21.01$, $p < .001$. Die Effektstärken waren mittel bis

groß (Leichtigkeit: $\eta_p^2 = .049$; Sicherheit: $\eta_p^2 = .144$; Überzeugung: $\eta_p^2 = .141$). Keine der anderen UVs und Interaktionen in den Zwischensubjekteffekten erreichte das Signifikanzniveau (ps > .05) und die Effektstärken lagen zwischen $\eta_p^2 = .001$ und $\eta_p^2 = .012$.

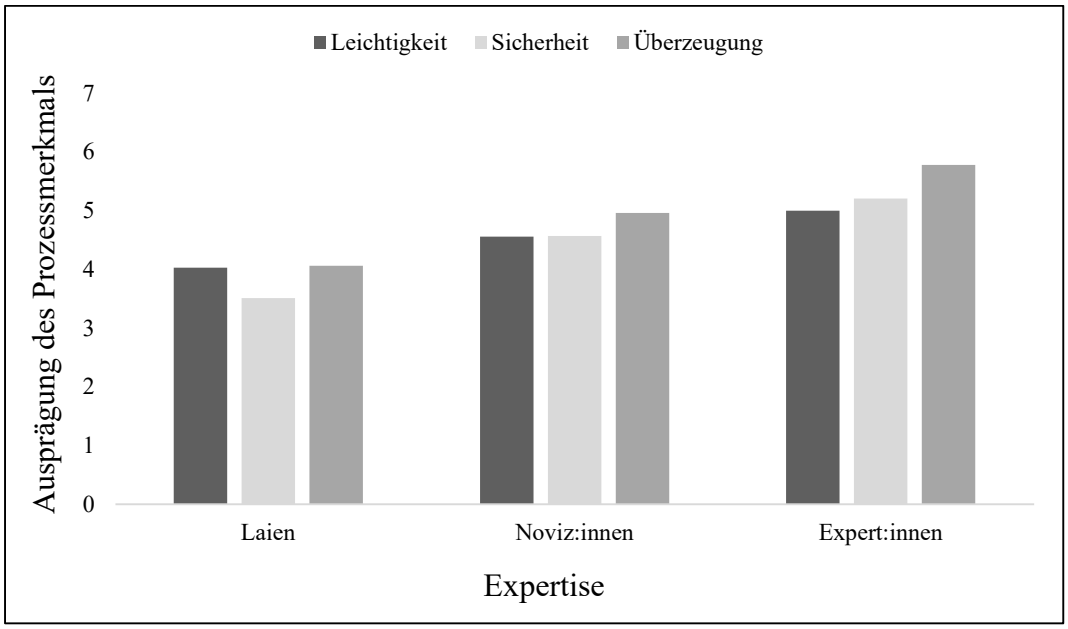

Abbildung 4.3. Darstellung der Mittelwerte der Expertise-Gruppe hinsichtlich der Prozessmerkmale auf einer 7-stufigen Likert-Skala.

Da die Gleichheit der Gruppengrößen sowie die multivariate Normalität nicht gegeben waren, aber von der Homogenität der Kovarianzmatrizen ausgegangen werden konnte (s. 3.8.3), wurde sich für den post-hoc-Test GT2 nach Hochberg entschieden (Field, 2013). Hinsichtlich der Leichtigkeit unterschieden sich die Laien signifikant von den Expert:innen ($p < .001$). In ihrer Sicherheit unterschieden sich die Laien signifikant von den Noviz:innen und Expert:innen (ps < .001), wobei dies auch auf den Vergleich von Noviz:innen und Expert:innen zutraf ($p = .017$). Gleiches galt für das Gefühl der Überzeugung, in dem sich alle Expertise-Gruppe im Vergleich zur jeweils anderen Gruppe unterschieden (ps < .05).

Die F1 kann dahingehend beantwortet werden, dass sich die Expertise-Gruppen in ihren Einschätzungen zur Leichtigkeit, Sicherheit und Überzeugung signifikant und mit teils großer Effektstärke unterschieden. Expert:innen fiel die Entscheidung leichter als Laien. Expert:innen fühlten sich sicherer und waren überzeugter als Laien und Noviz:innen, wobei sich letztere ebenfalls als sicherer und überzeugter

einschätzten als Nicht-Fachpersonen. Sowohl das Delikt als auch der Zeitdruck übten nicht signifikante und kleine Effekte aus.

4.7 F2 (Kruskal-Wallis-Test)

Die F2 befasst sich mit der explorativen Fragestellung, ob sich Laien, Noviz:innen und Expert:innen in ihrer Fähigkeit zur kognitiven Reflexion und ihrem Need for Cognition unterscheiden.

4.7.1 Deskriptive Auswertung der F2

Die Mittelwerte der einzelnen Expertise-Gruppen sowie der gesamten Stichprobe deuteten an, dass im Schnitt zwei der drei Items des CRT korrekt beantwortet wurden (s. Tabelle 4.8). Das Item 1 (Schläger und Ball) wurde von 59.47% der Teilnehmenden richtig gelöst. Dabei schnitten die Expert:innen (64.2%) am besten ab, gefolgt von den Noviz:innen (61.11%) und Laien (50%). Beim Item 2 (Maschinen) schnitten die Teilnehmenden insgesamt besser ab (74.14%). Auch hierbei lagen die Expert:innen (81.48%) vorne, wobei Noviz:innen (70.65%) und Laien (69.5%) vergleichbar abschnitten. Am einfachsten fiel das Item 3 (Seerosen), das von insgesamt 81.44% gelöst wurde. Hierbei änderte sich die Reihenfolge, wenn auch nur knapp, da die Noviz:innen (83.87%) den höchsten Lösungsanteil hatten, gefolgt von den Expert:innen (82.93%) und den Laien (75.81%). Unterscheidet man die Reflexionsfähigkeit nach Geschlechtern, so schnitten Männer (M = 2.35, SD = .81) etwas besser ab als nichtbinäre Personen (M = 2.33, SD = 1.15) sowie Frauen (M = 2.1, SD = 1.03). Rund 54.12% der Männer, 44.19% der Frauen und 66.67% der nichtbinären Personen lösten dabei alle Items korrekt. Etwa 2.35% der Männer und 12.04% der Frauen erzielten keinen Punkt.

Tabelle 4.8. Mittelwerte und Standardabweichungen der metrischen Prädiktoren „kognitive Reflexion“ und „Need for Cognition“ in Abhängigkeit von Expertise

	Metrische Prädiktoren					
	Kognitive Reflexion			Need for Cognition		
Expertise	*N*	*M*	*SD*	*N*	*M*	*SD*
Laien	53	2.06	.97	65	19.94	3.4
Noviz:innen	89	2.15	1.01	101	21.15	3.05
Expert:innen	78	2.31	.89	87	19.93	3.43
Gesamt	220	2.18	.98	253	20.41	3.32

Mit Blick auf das Kognitionsbedürfnis zeigten sich nur minimale Mittelwertsunterschiede zwischen den Expertise-Leveln. Geht man vom Maximalwert von 28 aus, wurden gruppenübergreifend sämtliche Items im Sinne eines hohen Kognitionsbedürfnisses beantwortet, wenngleich das Ausmaß der Zustimmung durchaus variierte. So war die Bereitschaft sich eher mit komplizierten als einfachen Problem zu beschäftigen nicht allzu stark ausgeprägt (M = 3.94, SD = 1.45). Dagegen waren sich die Teilnehmenden einig, dass es nicht ausreicht die Antwort zu kennen, ohne das Problem zu verstehen (M = 6.06, SD = 1.13). Die Mittelwerte für die Vorliebe für knifflige Aufgaben (M = 4.47, SD = 1.34) sowie für die allgemeine Bereitschaft zum Denken (M = 5.9, SD = 1.33) sprachen für eine eher hohe Motivation zum Denken.

4.7.2 Inferenzstatistische Auswertung der F2

Da die Voraussetzungen für eine ANOVA nicht erfüllt waren (s. 3.8.2.3), wurde der Kruskal-Wallis-Test eingesetzt. Die Expertise-Gruppen unterschieden sich nicht signifikant in ihrer Fähigkeit zur kognitiven Reflexion, $H(2) = 2.511$, $p = .285$. Die mittleren Rangsummen betrugen für die Expert:innen 117.96, für die Noviz:innen 109.3 und für die Laien 101.54. Für das Kognitionsbedürfnis zeigte sich ein signifikanter Unterschied, $H(2) = 6.843$, $p = .033$. Die mittleren Rangsummen betrugen für die Expert:innen 117.18, für die Noviz:innen 141.7 und für die Laien 117.3. Die paarweisen Vergleiche (mit adjustierten p-Werten) wiesen jedoch auf keine signifikanten Unterschiede bestimmter Gruppen hin (ps > .05). Effektstärken für die einzelnen, nicht signifikanten paarweisen Vergleiche waren nicht zu berechnen. Zur Ermittlung der gesamten Effektstärke diente die Formel $\eta^2_H = H / (N - 1)$ nach Serlin et al. (1982, zitiert nach Leonhart, 2017). Für die kognitive Reflexion ($\eta^2_H = .011$) sowie für NFC ($\eta^2_H = .028$) ergaben sich kleine Effekte.

Die F2 kann dahingehend beantwortet werden, dass die Expertise einen kleinen, aber signifikanten Effekt auf die Ausprägung des Kognitionsbedürfnisses, aber nicht auf die kognitive Reflexion hatte. Auf deskriptiver Ebene war das Kognitionsbedürfnis der Noviz:innen größer als das der Vergleichsgruppen.

4.8 F3 (binäre logistische Regression)

Die F3 befasst sich mit der explorativen Fragestellung, ob die kognitive Reflexion, Need for Cognition, die Expertise, der Zeitdruck, der Delikttyp oder die genutzte Lesezeit die Entscheidung hinsichtlich des Tatverdachtes vorhersagen.

4.8.1 Deskriptive Auswertung der F3

Für deskriptive Informationen über die kategorischen Prädiktoren wird auf die Kontingenztabelle der H1 verwiesen (s. Tabelle 4.1). Die Mittelwerte und Standardabweichungen für das Kognitionsbedürfnis sowie die kognitive Reflexion wurden bereits im Zusammenhang mit der F2 berechnet (s. Tabelle 4.8). Für das Lesen der Vignette (Sekunden) benötigten die Teilnehmenden durchschnittlich weniger als zwei Minuten (M = 106.45, SD = 46.39). Dabei waren die Expert:innen durchschnittlich am schnellsten (M = 99.74, SD = 47.42), gefolgt von den Noviz:innen (M = 108.63, SD = 44.8) und den Laien (M = 113.02, SD = 46.65).

4.8.2 Inferenzstatistische Auswertung der F3

Die Voraussetzungen für die binäre logistische Regression waren erfüllt (s. 3.8.1.3). Ebenso wie bei den Berechnungen der H1 und der H3 galten auch hier die Laien, das Delikt „Diebstahl“ sowie die Manipulation „ohne Zeitdruck“ als Referenzgruppen (s. 4.2.2; 4.4.2). Die Prädiktoren wurden mittels „Einschluss“-Methode dem Modell hinzugefügt. Die Untersuchung der Residuen erfolgte nach den bekannten Kriterien (s. 4.2.2). Das Maximum der Einflussstatistik nach Cook sowie die DFBeta-Werte für die Konstante sowie das Modell lagen durchgehend jeweils unter 1. Der durchschnittlich zu erwartende Hebelwert betrug einen Wert von .029. Der Maximalwert in der Stichprobe betrug .077, wonach kein Fall über der dreifachen Höhe des Durchschnitts lag. Der maximale Wert der standardisierten Residuen war mit einem Wert von 2.12 auffällig. Da lediglich drei Personen (1.29%) – und somit weniger als 5% der Stichprobe – einen Wert ≤ 1.96 erreichten, ist der Einfluss ausreißender Werte als gering anzusehen (Field, 2013).

Fügte man die Prädiktoren und deren Interaktionen dem Modell hinzu, so ließ sich eine signifikante Verbesserung des Modells ablesen ($p < .001$) und die anfängliche

-2-Log-Likelihood reduzierte sich von 323.97 im Null-Modell auf 286.63 im Modell 1. Der Gesamtprozentsatz der Richtigen stieg auf 66.2%. Laut dem Hosmer-Lemeshow-Test passten Daten und Modell gut zusammen (p = .652).

Noviz:innen und Expert:innen wiesen im Vergleich zur Referenzgruppe einen negativen Regressionskoeffizienten auf: Die Wahrscheinlichkeit, den Tatverdacht zu bejahen, sank in beiden Gruppen (s. Tabelle 4.9). Das Ergebnis der Noviz:innen verfehlte knapp das Signifikanzniveau (p = .052). In dieser Gruppe überschritt das Konfidenzintervall zudem leicht den Wert von 1, sodass eine Richtung der Entscheidung nicht eindeutig angegeben werden konnte. Das Ergebnis der Expert:innen war dagegen signifikant (p < .001).

Tabelle 4.9. Regressionsergebnisse der F3 für die Prädiktoren

Prädiktoren	*b*	SE *b*	*p*	Konfidenzintervall (95%) Unterer Wert	Konfidenzintervall (95%) Oberer Wert	Exp(B)
Expertise(1)	-.759	.391	.052	.218	1.006	.468
Expertise(2)	-1.913	.42	< .001	.065	.336	.148
Zeitdruck(1)	.559	.31	.071	.953	3.207	1.748
Delikt(1)	.529	.305	.083	.933	3.084	1.697
Kognitive Reflexion	-.193	.157	.22	.606	1.123	.825
Need for Cognition	-.032	.049	.512	.880	1.066	.969
Lesezeit	.005	.004	.183	.998	1.012	1.005

Anmerkungen. R^2 = .199 (Nagelkerke), Expertise(1) = Noviz:innen, Expertise(2) = Expert:innen, Zeitdruck(1) = mit, Delikt(1) = Körperverletzung.

Die Zeitdruckmanipulation erwies sich nicht als signifikanter Prädiktor (p = .071). Laut Regressionskoeffizient stieg die Wahrscheinlichkeit, den Tatverdacht unter Zeitdruck zu bejahen, obwohl die Richtung gemäß dem unteren und oberen Wert des Konfidenzintervalls nicht eindeutig war. Auch das Delikt war kein signifikanter Prädiktor (p = .083). Für das Delikt „Körperverletzung“ stieg im Vergleich zum Diebstahl die Wahrscheinlichkeit, den Tatverdacht zu bejahen. Eine Richtung war aber nicht eindeutig anzugeben. Die Effektstärken waren klein, wobei die nicht signifikanten Prädiktoren „Delikt“ und „Zeitdruck“ größere Werte erzielten als der signifikante Prädiktor „Expertise“ (s. auch Nagelkerkes R^2).

Die metrischen Prädiktoren erwiesen sich als nicht signifikant (ps > .05). Laut Regressionskoeffizienten gingen hohe Ausprägungen der kognitiven Reflexion und

des Need for Cognition mit einer niedrigeren Wahrscheinlichkeit einher den Tatverdacht zu bejahen, wohingegen für eine längere Lesezeit die Wahrscheinlichkeit (minimal) stieg. Das (knappe) Überschreiten des Wertes von 1 bei den Bereichen der Konfidenzintervalle weist daraufhin, dass eine Richtung nicht eindeutig angegeben werden konnte.

Die F3 kann dahingehend beantwortet werden, dass nur der Prädiktor „Expertise" das Vorliegen des Tatverdachtes im Vergleich mit der Referenzgruppe der Laien vorhersagte: Die Wahrscheinlichkeit des Bejahens eines Tatverdachtes sank sowohl für die Noviz:innen (nur knapp nicht signifikant) als auch für die Expert:innen.

4.9 F4 (multiple Regression)

Die F4 befasst sich mit der explorativen Fragestellung, ob die kognitive Reflexion, Need for Cognition, die Expertise, der Zeitdruck, der Delikttyp oder die genutzte Lesezeit die Anzahl der Beweismittel vorhersagen, die zur Entscheidung hinsichtlich des Tatverdachtes herangezogen wird.

4.9.1 Deskriptive Auswertung der F4

Für deskriptive Angaben zu den metrischen Prädiktoren „Need for Cognition" und „kognitive Reflexion" wird auf Abschnitt 4.7.1 und für Angaben zur Lesezeit wird auf Abschnitt 4.8.1 verwiesen. Die Mittelwerte und Standardabweichungen der genutzten Anzahlen von Beweismitteln sind in Tabelle 4.10 nach Prädiktoren aufgeschlüsselt.

Tabelle 4.10. Mittelwerte und Standardabweichungen der Anzahl der genutzten Beweismittel in Abhängigkeit von Expertise, Zeitdruck und Delikt

		Anzahl der Beweismittel	
Prädiktor	Gruppenzuordnung (*N*)	*M*	*SD*
Expertise	Laien (66)	2.82	.76
	Noviz:innen (102)	3.03	.79
	Expert:innen (89)	2.85	.9
Zeitdruck	ohne (119)	2.84	.8
	mit (138)	2.98	.84
Delikt	Diebstahl (120)	3.00	.83
	Körperverletzung (137)	2.84	.82
	Gesamt (257)	2.91	.83

Es wird deutlich, dass sich die Untergruppen der einzelnen Prädiktoren nicht wesentlich in der Anzahl der genutzten Beweismittel unterschieden. Da maximal vier Beweismittel für das Treffen der Entscheidung genutzt werden konnten, lässt sich ableiten, dass sich die Proband:innen im Durchschnitt auf über die Hälfte der vorhandenen Informationselemente bezogen.

4.9.2 Inferenzstatistische Auswertung der F4

Als Referenzgruppen galten weiterhin das Delikt „Diebstahl“ und die Manipulation „ohne Zeitdruck“. Für die Noviz:innen und Expert:innen wurden Dummy-Variablen erstellt, sodass Vergleiche mit den Laien stattfanden. Die Prädiktoren wurden mittels „Einschluss“-Methode dem Modell hinzugefügt.

Es wurden die Residuen überprüft, um die Passung zwischen Daten und Modell sowie das Vorhandensein von einflussreichen Fällen zu untersuchen (s. 4.2.2). Das Maximum der Einflussstatistik nach Cook sowie die DFBeta-Werte für die Konstante sowie das Modell lagen durchgehend jeweils unter dem Wert von 1. Der durchschnittlich zu erwartende Hebelwert betrug .029, aber da der Maximalwert in der Stichprobe bei einem Wert von .07 lag, war nicht vom Vorhandensein einflussreicher Fälle auszugehen. Der maximale Wert der standardisierten Residuen war mit einem Wert von 1.8 unauffällig. Rund 5% der Fälle einer Stichprobe dürfen standardisierte Residuen von ± 2 aufweisen (Field, 2013). In der hiesigen Stichprobe entspräche dies in etwa 11 Proband:innen. Da lediglich acht Fälle auffällige Werte > 2 aufwiesen, lag die Anzahl möglicher extremer Fälle im Rahmen.

Da die Voraussetzungen für das Testverfahren teilweise verletzt waren (s. 3.8.4), wurde sich für den Einsatz von Bootstrapping entschieden.[39] Das Modell mit den ausgewählten Prädiktoren erklärte nur einen kleinen Teil der Varianz (R^2 = .076), wobei sich dieser Anteil noch weiter verkleinerte, wenn man das Modell von der vorliegenden Stichprobe auf die Gesamtpopulation übertragen würde (korrigiertes R^2 = .045; s. Tabelle 4.11). Dennoch war das Ergebnis laut ANOVA signifikant, $F(7, 208) = 2.446$, $p = .02$.

[39] Siehe Fußnote 38.

Tabelle 4.11. Regressionsergebnisse der F4 für die Prädiktoren (1000 Bootstrapping-Stichproben)

				BCa Konfidenzintervall (95%)	
Prädiktoren	*b*	SE *b*	*p*	Unterer Wert	Oberer Wert
Expertise (1)	.157	.128	.215	-.093	.403
Expertise(2)	.006	.152	.969	-.261	.311
Zeitdruck	.128	.108	.242	-.078	.34
Delikt	-.153	.116	.197	-.385	.058
Kognitive Reflexion	.132	.06	.037	.018	.26
Need for Cognition	.02	.018	.275	-.016	.052
Lesezeit	.003	.001	.016	.001	.005

Anmerkungen. R^2 = .076, korrigiertes R^2 = .045, Expertise(1) = Noviz:innen, Expertise(2) = Expert:innen.

Die Regressionskoeffizienten zeigen, dass die Expertise (Noviz:in) den stärksten Prädiktor darstellte: Noviz:in zu sein erhöhte die Wahrscheinlichkeit, eine größere Anzahl an Informationen zu nutzen. Es folgten in abnehmender Stärke das Delikt, die kognitive Reflexion und der Zeitdruck. Eine hohe Ausprägung von kognitiver Reflexion sowie das Vorhandensein von Zeitdruck erhöhten die Wahrscheinlichkeit, wohingegen das Delikt „Körperverletzung" sie verringerte. Sehr geringe Stärken wiesen Need for Cognition (verringert leicht), die Expertise (Expert:in; verringert leicht) sowie die Lesezeit (erhöht sehr leicht) auf. Allerdings waren nur die Ergebnisse für die kognitive Reflexion (p = .037) und die Lesezeit (p = .016) signifikant. Die mit der Formel $f^2 = R^2 / (1 - R^2)$ ermittelte Effektstärke war allerdings als klein einzustufen (f^2 = .01).[40]

Die BCa-Konfidenzintervalle (95%) für die signifikanten Prädiktoren waren eng gefasst, überschritten nicht den Wert von 0 und blieben jeweils positiv. Die Intervalle anderer Prädiktoren waren zwar auch eng gefasst, aber sie überschritten (teilweise knapp) den Wert von 0 und deuteten somit auf ein schlecht passendes Modell hin, da die Richtung der Vorhersage nicht eindeutig beschrieben werden konnte. Für die kognitive Reflexion betrug die Korrelation Nullter Ordnung r = .161 (Pearsons Korrelationskoeffizient) und für die Lesezeit ergab sich ein Wert von r = .136. Dies waren gleichzeitig auch die stärksten Korrelation zwischen einem Prädiktor und der AV. Die schwächste Korrelation bestand mit dem Zeitdruck (r = .052).

Die F4 kann dahingehend beantwortet werden, dass nur die kognitive Reflexion sowie die Lesezeit die Anzahl der Beweismittel vorhersagten, die zur Entscheidung

[40] Siehe Bortz und Schuster (2010, S. 359).

über das Vorliegen des Tatverdachtes genutzt wurde: Hohe Werte in kognitiver Reflexion sowie eine lange Lesezeit gingen mit der Nutzung zahlreicher Informationen einher.

4.10 F5 (multinominale logistische Regression)

Die F5 befasst sich mit der explorativen Fragestellung, ob sich Laien, Noviz:innen und Expert:innen darin unterscheiden, wie sie die Relevanz einzelner Beweismittel für die Entscheidung über das Vorliegen des Tatverdachtes einschätzen.

4.10.1 Deskriptive Auswertung der F5

Zur besseren Überschaubarkeit werden die Beweismittel mit niedrigster beziehungsweise höchster Relevanz getrennt betrachtet. Da sich aufgrund der Vielzahl an Zellen und der daraus resultierenden Kombinationsmöglichkeiten sehr komplexe Kontingenztabellen ergeben, wird an dieser Stelle auf diese Darstellungsform verzichtet. Da die Antworten mitunter auf sehr kleinen Zellgrößen basieren und daher eher als Antworttendenzen zu werten sind, werden anstatt der Prozentzahlen die jeweiligen Anzahlen in Relation zur Gesamtzahl angegeben, um ein besseres Bild zu vermitteln.

Zunächst werden die deskriptiven Beschreibungen für das Beweismittel mit der *niedrigsten* Relevanz dargestellt. Auf die Gesamtstichprobe bezogen zeigte sich in den Häufigkeiten die folgende Reihenfolge hinsichtlich der Relevanz: Auszug aus dem Bundeszentralregister als Urkundenbeweis (161/257), Erscheinungsbild als Augenscheinbeweis (49/257), Aussage des Beschuldigten (32/257) und Zeugenaussage (15/257). Für das Delikt „Diebstahl“ war die gleiche Reihenfolge der (niedrigen) Relevanz erkennbar. Am stärksten zeigte sich die Tendenz für die Wahl des Bundeszentralregisterauszuges für die Expert:innen (32/39), wohingegen die Laien diesem (13/31) eine ähnlich niedrige Relevanz beimaßen wie dem Erscheinungsbild des Beschuldigten (12/31). Das Fehlen oder Vorhandensein von Zeitdruck machte für die Entscheidungstendenz der Expert:innen und Noviz:innen keinen Unterschied. Laien ohne Zeitdruck entschieden sich mehrheitlich (8/16) für das Erscheinungsbild des Beschuldigten, wohingegen Laien unter Zeitdruck zum Urkundenbeweis tendierten (8/15). Beim Delikt „Körperverletzung“ blieb das Bild bei den Expert:innen eindeutig (Bundeszentralregisterauszug: 40/50), wobei sich die

Antworten der Laien und Noviz:innen mehr verteilten. So tendierten die meisten Proband:innen zum Bundeszentralregisterauszug, aber bei den Laien erhielten das Erscheinungsbild (7/35) sowie die Aussage des Beschuldigten (9/35) ähnlich viele Stimmen wie bei den Noviz:innen (8/52 bzw. 7/52). Unabhängig vom Delikt beeinflusste die Zeitdruckmanipulation nicht die Wahl des Bundeszentralregisterauszuges als Beweismittel mit der niedrigsten Relevanz (84/138). Unabhängig vom Zeitdruck zeigte sich weiterhin, dass diese Tendenz für die Expert:innen stärker ausgeprägt war als für die Laien und Noviz:innen, deren Wahl durchaus auf das Erscheinungsbild oder die Aussage des Beschuldigten fiel – sowohl mit als auch ohne Zeitdruck. Expert:innen sprachen sich dagegen mehrheitlich für die geringe Bedeutsamkeit des Urkundenbeweises aus (72/89).

Nun folgen die deskriptiven Beschreibungen für das Beweismittel mit der *höchsten* Relevanz. Auf die Gesamtstichprobe bezogen zeigte sich in den Häufigkeiten die folgende Reihenfolge hinsichtlich der Relevanz: Zeugenaussage (131/257), Erscheinungsbild (61/257) beziehungsweise Aussage des Beschuldigten (55/257) und Auszug aus dem Bundeszentralregister (10/257). Dies passt zu der bereits genannten Reihenfolge zur Einschätzung der niedrigsten Relevanz, wenngleich die Positionen 2 und 3 vertauscht sind. Beim Delikt „Diebstahl“ erwies sich Expertise-übergreifend die Aussage des Zeugen (64/120) am relevantesten, gefolgt von der Aussage (31/120) und dem Erscheinungsbild (19/120) des Beschuldigten. Für lediglich sechs Personen war der Bundeszentralregisterauszug das entscheidende Beweismittel für die Entscheidung über das Vorliegen des Tatverdachtes (6/120). Für Laien und Expert:innen zeigte sich dieses Muster auch in der jeweiligen Untergruppe, wohingegen für die Noviz:innen das Erscheinungsbild (12/50) sowie die Aussage des Beschuldigten (11/50) vergleichbar relevant waren. Unterscheidet man zudem nach Zeitdruck, so verschiebt sich das Antwortmuster für die jeweiligen Beweismittel leicht: Blieb mit (32/64) beziehungsweise ohne (32/56) Zeitdruck die Aussage des Zeugen an erster Stelle der Relevanz, so gaben diejenigen unter Zeitdruck eine ähnliche Bedeutsamkeit für das Erscheinungsbild (15/64) und die Aussage (14/64) des Beschuldigten an, wohingegen diejenigen ohne Zeitdruck der Aussage des Beschuldigten (17/54) eine wesentlich höhere Bedeutung beimaßen als dem Erscheinungsbild (4/54). Unterscheidet man auf Zeitdruck-Ebene nun auch nach Expertise, tendierten die meisten Personen aller Expertise-Gruppen ohne Zeitdruck

dazu, die Aussage des Zeugen am bedeutsamsten einzuschätzen. Die Aussage des Beschuldigten wurde am zweithäufigsten genannt. Unter Zeitdruck verteilten sich die Angaben zur Relevanz: Für mehr Noviz:innen unter Zeitdruck erwies sich das Erscheinungsbild des Beschuldigten bedeutsamer als dessen Aussage. Auch für das Delikt „Körperverletzung“ wurde die Aussage des Zeugen (67/137) am häufigsten als relevant für die Entscheidung hinsichtlich des Vorliegens des Tatverdachtes eingeschätzt, in abnehmender Häufigkeit gefolgt vom Erscheinungsbild (42/137), von der Aussage des Beschuldigten (24/137) sowie vom Auszug aus dem Bundeszentralregister (4/137). Noviz:innen und Expert:innen ähnelten sich dabei in den Antwortmustern (wenngleich Noviz:innen das Erscheinungsbild verhältnismäßig häufiger als am relevantesten einschätzten). Laien nutzten am häufigsten die Informationen des Zeugen für ihre Entscheidung (18/35), gefolgt vom Erscheinungsbild (14/35). Unter Zeitdruck waren das Erscheinungsbild (29/74) sowie die Aussage des Zeugen (31/74) ähnlich häufig relevant. Ohne Zeitdruck erschienen dagegen das Erscheinungsbild (13/63) beziehungsweise die Aussage (14/63) des Beschuldigten ähnlich relevant (Aussage des Zeugen: 36/63). Laien und Noviz:innen unter Zeitdruck ähnelten sich in ihren Antwortmustern, da dem Erscheinungsbild des Beschuldigten beziehungsweise der Aussage des Zeugen am häufigsten die größte Relevanz beigemessen wurde; Expert:innen verteilten sich dagegen recht gleichmäßig auf diese beiden Beweismittel sowie zusätzlich auf die Aussage des Beschuldigten. Ohne Zeitdruck unterschieden sich die Antworten der Laien von den anderen beiden Gruppen: Noviz:innen und Expert:innen wählten häufiger das Erscheinungsbild sowie die Aussage des Beschuldigten (wenngleich die Aussage des Zeugen für alle Gruppen am häufigsten ausgewählt wurde).

Zusammenfassend wurde sowohl für den Diebstahl als auch für die Körperverletzung die gruppenübergreifende Tendenz deutlich, die Aussage des Zeugen am häufigsten als relevantestes Beweismittel einzustufen. Unterscheidet man allerdings nach Zeitdruck- beziehungsweise Expertise-Gruppen, so verschoben sich die Relevanzen teilweise. Unabhängig vom Delikt wurde deutlich, dass sich Expert:innen in den Angaben zum relevantesten Beweismittel auf deskriptiver Ebene unterschieden, je nachdem, ob sie unter Zeitdruck standen oder nicht. Verteilten sich die Einschätzungen zur Relevanz unter Zeitdruck auf die Aussage des Zeugen (19/47), die Aussage des Beschuldigten (14/47) beziehungsweise dessen Erscheinungsbild

(11/47), so zeigte sich für die Fachpersonen ohne Zeitdruck die gleiche Reihenfolge, aber eine stärkere Tendenz (27/42, 10/42, 5/42). Noviz:innen mit und ohne Zeitdruck unterschieden sich auf deskriptiver Ebene nur darin, welches Beweismittel am zweithäufigsten als am relevantesten für die Entscheidung über den Tatverdacht benannt wird: Unter Zeitdruck war das Erscheinungsbild häufiger relevant (21/57) als die Aussage des Beschuldigten (6/57), aber ohne Zeitdruck ist die Aussage (15/45) häufiger relevant als das Erscheinungsbild (8/45). Laien unter Zeitdruck stuften in erster Linie die Aussage des Zeugen als relevant ein (17/34), am zweithäufigsten das Erscheinungsbild (12/34). Laien ohne Zeitdruck maßen dagegen der Aussage des Beschuldigten (6/32) ähnlich häufig (bzw. selten) die größte Relevanz bei (4/32). Unabhängig von Delikt, Zeitdruck und Expertise zeigte sich durchgehend, dass dem Bundeszentralregisterauszug am seltensten die höchste Bedeutsamkeit zugesprochen wurde.

4.10.2 Exkurs: Bedeutsamkeit der einzelnen Beweismittel sowie Bewertung der Vignetten

Zusätzlich lässt sich auf deskriptiver Ebene betrachten, wie die einzelnen Gruppen die jeweilige Relevanz der einzelnen Informationselemente einschätzten. Da diese Messungen nicht weiter inferenzstatistisch untersucht wurden, werden sie an dieser Stelle lediglich berichtet (s. Tabelle 4.12). Mit Blick auf die Expertise waren sich die einzelnen Gruppen in ihrer Einschätzung der Bedeutsamkeit recht ähnlich, mit Ausnahme des Urkundenbeweises (Bundeszentralregisterauszug). Dieser wurde von den Expert:innen vergleichsweise unbedeutsam eingeschätzt. Der Einfluss von Zeitdruck auf die Einschätzung der Aussagen des Zeugen und des Beschuldigten war deskriptiv sehr gering. Ein solcher Einfluss lässt sich am ehesten für den Augenschein erkennen, der unter Zeitdruck als bedeutsamer für die Entscheidung nach dem Vorliegen des Tatverdachtes eingeschätzt wurde. Unterschiede zwischen den Delikttypen lassen sich nur für den Augenschein ausmachen, da dieser beim Delikt der Körperverletzung als relevanter eingestuft wurde. Über die Gruppen hinweg zeigte sich eine starke Tendenz, die Aussage des Zeugen als bedeutsam einzustufen, wohingegen dem Urkundenbeweis nur eine geringe Bedeutsamkeit beigemessen wurde (s. auch 4.10.1).

Tabelle 4.12. Mittelwerte und Standardabweichungen zur Einschätzung der Bedeutsamkeit der Beweismittel auf einer 7-stufigen Likert-Skala in Abhängigkeit von Expertise, Zeitdruck und Delikt

		Bedeutsamkeit der Beweismittel	
Prädiktor	Gruppenzuordnung (*N*)	*M*	*SD*
Augenschein	Laien (66)	4.17	1.92
	Noviz:innen (102)	4.64	1.78
	Expert:innen (89)	4.18	1.87
	ohne Zeitdruck (119)	4.01	1.92
	mit Zeitdruck (138)	4.66	1.75
	Diebstahl (120)	3.95	1.81
	Körperverletzung (137)	4.72	1.84
	Gesamt (257)	4.36	1.86
Urkunde	Laien (66)	3.36	1.76
	Noviz:innen (102)	3.1	1.74
	Expert:innen (89)	2.16	1.59
	ohne Zeitdruck (119)	2.62	1.74
	mit Zeitdruck (138)	3.03	1.77
	Diebstahl (120)	3.1	1.85
	Körperverletzung (137)	2.61	1.65
	Gesamt (257)	2.84	1.76
Aussage des Zeugen	Laien (66)	5.32	1.18
	Noviz:innen (102)	5.35	1.38
	Expert:innen (89)	5.55	1.62
	ohne Zeitdruck (119)	5.34	1.39
	mit Zeitdruck (138)	5.26	1.43
	Diebstahl (120)	5.35	1.36
	Körperverletzung (137)	5.47	1.47
	Gesamt (257)	5.41	1.42
Aussage des Beschuldigten	Laien (66)	4.36	1.45
	Noviz:innen (102)	4.83	1.47
	Expert:innen (89)	5.03	1.67
	ohne Zeitdruck (119)	4.91	1.55
	mit Zeitdruck (138)	4.67	1.55
	Diebstahl (120)	4.91	1.53
	Körperverletzung (137)	4.67	1.57
	Gesamt (257)	4.78	1.55

Anmerkungen. Augenschein = Erscheinungsbild des Beschuldigten, Urkunde = Auszug aus dem Bundeszentralregister.

Neben diesen Einschätzungen zu den Inhalten der Vignetten (Beweismittel) lassen sich auch die Angaben der Proband:innen zum Realismus der Fallbeschreibungen betrachten (in Abhängigkeit von Expertise; s. Tabelle 4.13).

Tabelle 4.13. Mittelwerte und Standardabweichungen zur Einschätzung der Vignetten hinsichtlich deren Realismus auf einer 7-stufigen Likert-Skala in Abhängigkeit von Expertise

		Realismus der Vignette	
Delikt	Gruppenzuordnung (*N*)	*M*	*SD*
Diebstahl	Laien (31)	5.71	1.31
	Noviz:innen (50)	5.6	1.43
	Expert:innen (39)	5.2	1.75
	Gesamt (120)	5.49	1.48
Körperverletzung	Laien (35)	5.29	1.38
	Noviz:innen (52)	5.46	1.36
	Expert:innen (50)	4.88	1.72
	Gesamt (137)	5.2	1.52

Beide Fallbeschreibungen wurden vergleichbar realistisch eingestuft. In beiden Fällen war die Einschätzung der Expert:innen am „kritischsten", wenngleich die deskriptiven Mittelwertsunterschiede nur marginal waren. Die Standardabweichungen der Expert:innen waren jeweils größer als die Expertise-übergreifenden Werte.

4.10.3 Inferenzstatistische Auswertung der F5

Da die Voraussetzungen für das Testverfahren teilweise verletzt waren (s. 3.8.5), wurde sich für den Einsatz von Bootstrapping entschieden.[41] Aufgrund der geringen Zellgrößen kam bei dieser Maßnahme auch das Ziehen von Stichproben mit leeren Zellen vor. Laut dem eingesetzten Statistikprogramm war die „Gültigkeit der Modellanpassung" ungewiss. Die folgenden zwei Modellrechnungen sollten daher unter Vorbehalt betrachtet und interpretiert werden (s. auch Eid et al., 2017). Dafür sprechen auch die teils sehr weit gefassten BCa-Konfidenzintervalle (95%).

Zur besseren Übersichtlichkeit erfolgten die beiden Modellrechnungen getrennt nach dem Beweismittel mit niedrigster und höchster Relevanz. In beiden Modellen wurden die Prädiktoren mittels „Einschluss"-Methode dem Modell hinzugefügt. Das Fehlen von Zeitdruck sowie das Delikt „Diebstahl" galten weiterhin als Referenzgruppen (s. 4.2.2). Da sich der Faktor „Expertise" auf drei Level aufteilt, wurden für die Noviz:innen und Expert:innen die Dummy-Variablen eingesetzt, die zuvor für die F4-Berechnung genutzt wurden (s. 4.9.2). Da sich die AV „Beweismittel" je Modell auf vier Level aufteilt, musste auch hier eine Referenzkategorie ge-

[41] Siehe Fußnote 38.

wählt werden. Es wurde der Augenschein, also das Erscheinungsbild des Beschuldigten, festgelegt. Da keine theoretische Ableitung darüber vorlag, welches der vier Beweismittel eine gute Referenzgruppe darstellt, wurde mit dem Beweis begonnen, der auch im Versuchsablauf an erster Stelle abgefragt worden war (s. 3.4.2.3).

Die erste Modellrechnung befasste sich mit dem Beweismittel mit *niedrigster* Relevanz. Das Modell erwies sich als signifikant ($p < .001$). Die anfängliche -2-Log-Likelihood reduzierte sich von 141.817 im Null-Modell auf 100.943 im Modell 1, was darauf hindeutet, dass im ursprünglichen Null-Modell mehr Varianz unerklärt blieb. Mit Blick auf die Güte der Anpassung zeigte sich, dass sowohl die Pearson- als auch die Abweichungsstatistik nicht signifikant waren (ps > 0.05). Dies spricht für eine gute Passung zwischen Daten und Modell.

Von den Prädiktoren erwies sich lediglich die Expertise der Expert:innen im Modell als signifikant ($\chi^2 = 25.72$, $p < .001$; Noviz:innen: $\chi^2 = 5.36$, $p = .149$), wenngleich das Delikt das Niveau nur sehr knapp verfehlte ($\chi^2 = 7.667$, $p = .053$). Der Zeitdruck war dagegen kein signifikanter Prädiktor ($\chi^2 = 7.201$, $p = .066$). Die Effekte waren klein (s. auch Nagelkerkes R^2). Im Folgenden wird auf die individuellen Parameterschätzungen und Regressionskoeffizienten eingegangen (s. Tabelle 4.14). Die Wahrscheinlichkeit, sich für den Auszug des Bundeszentralregisters als am wenigsten bedeutsames Beweismittel zu entscheiden (im Vergleich zum Erscheinungsbild des Beschuldigten),

- sank für die Laien im Vergleich zu den Noviz:innen und Expert:innen,
- sank ohne Zeitdruck,
- sank beim Delikt „Diebstahl“.

Diese Ergebnisse erwiesen sich für die Expert:innen und das Delikt als signifikant (ps < .05). Betrachtet man die unteren und oberen Werte der BCa-Konfidenzintervalle (95%) der signifikanten Prädiktoren, so konnte die Richtung für Expert:innen und das Delikt eindeutig vorgegeben werden.

Die Wahrscheinlichkeit, sich für die Aussage des Zeugen als am wenigsten bedeutsames Beweismittel zu entscheiden (im Vergleich zum Erscheinungsbild des Beschuldigten),

- sank für die Laien im Vergleich zu den Noviz:innen und Expert:innen,
- sank ohne Zeitdruck,
- sank beim Delikt „Diebstahl“.

Die Expertise-Prädiktoren hatten dabei keinen signifikanten Beitrag, im Gegensatz zu den Ergebnissen des Zeitdrucks (p = .017) und des Delikts (p = .033). Die Ober- und Untergrenze der BCa-Konfidenzintervalle (95%) der beiden signifikanten Prädiktoren blieben < 1, sodass die Richtung eindeutig vorgegeben werden konnte.

Tabelle 4.14. Regressionsergebnisse der F5 für die Prädiktoren hinsichtlich des Beweismittels mit der niedrigsten Relevanz (1000 Bootstrapping-Stichproben)

					BCa Konfidenzintervall (95%)	
Beweismittel	Prädiktoren	*b*	SE *b*	*p*	Unterer Wert	Oberer Wert
Urkunde	Expertise(1)	-.682	.429	.101	-1.577	.202
	Expertise(2)	-1.669	.507	< .001	-2.700	-.799
	Zeitdruck	-.278	.361	.434	-1.026	.398
	Delikt	-.736	.355	.031	-1.427	-.117
Aussage des	Expertise(1)	-.204	1.624	.806	-16.653	1.075
Zeugen	Expertise(2)	-.18	4.336	.845	-2.468	17.805
	Zeitdruck	-1.596	4.605	.017	-19.851	-.54
	Delikt	-1.256	1.152	.033	-2.252	-.451
Aussage des	Expertise(1)	.182	.567	.722	-.887	1.352
Beschuldigten	Expertise(2)	.321	1.578	.649	-1.312	2.425
	Zeitdruck	-.733	.49	.117	-1.755	.202
	Delikt	-1.06	.51	.027	-2.045	-.203

Anmerkungen. R^2 = .168 (Nagelkerke), Urkunde = Auszug aus dem Bundeszentralregister, Expertise(1) = Noviz:innen, Expertise(2) = Expert:innen.

Die Wahrscheinlichkeit, sich für die Aussage des Beschuldigten als am wenigsten bedeutsames Beweismittel zu entscheiden (im Vergleich zum Erscheinungsbild des Beschuldigten),

- stieg für die Laien im Vergleich zu den Noviz:innen und Expert:innen,
- sank ohne Zeitdruck,
- sank beim Delikt „Diebstahl“.

Lediglich das Ergebnis des Delikts erwies sich als signifikant (p = .027). Das Nicht-Überschreiten des Wertes von 1 im BCa-Konfidenzintervall (95%) bestätigte die Richtung der Wahrscheinlichkeit.

Die zweite Modellrechnung befasste sich mit dem Beweismittel mit *höchster* Relevanz. Das Modell erwies sich als signifikant (p = .003). Die anfängliche -2-Log-Likelihood reduzierte sich von 138.371 im Null-Modell auf 108.561 im Modell 1, was darauf hindeutet, dass im ursprünglichen Null-Modell mehr Varianz unerklärt

blieb. Sowohl die Pearson- als auch die Abweichungsstatistik war nicht signifikant (*p*s > .05). Dies spricht für eine gute Passung zwischen Daten und Modell.

Zwei der drei Prädiktoren waren signifikant: „Zeitdruck“ ($\chi^2 = 13.77$, $p = .003$) und „Delikt“ ($\chi^2 = 9.93$, $p = .019$). Der Prädiktor „Expertise“ verfehlte im Modell das Signifikanzniveau sowohl für die Noviz:innen ($\chi^2 = 2.21$, $p = .531$) als auch für die Expert:innen ($\chi^2 = 3.89$, $p = .274$). Die Effekte waren als klein einzustufen (s. auch Nagelkerkes R^2). Im Folgenden wird auf die individuellen Parameterschätzungen und Regressionskoeffizienten eingegangen (s. Tabelle 4.15). Die Wahrscheinlichkeit, sich für den Auszug des Bundeszentralregisters als bedeutsamstes Beweismittel zu entscheiden (im Vergleich zum Erscheinungsbild des Beschuldigten),

- sank für die Laien im Vergleich zu den Noviz:innen und Expert:innen,
- stieg ohne Zeitdruck,
- stieg beim Delikt „Diebstahl“.

Dies gilt allerdings vorrangig auf deskriptiver Ebene, da nur der Prädiktor „Delikt“ das Signifikanzniveau erreichte ($p = .048$). Die Richtung der Entscheidung konnte aber nicht eindeutig vorgeben werden.

Tabelle 4.15. Regressionsergebnisse der F5 für die Prädiktoren hinsichtlich des Beweismittels mit der höchsten Relevanz (1000 Bootstrapping-Stichproben)

					BCa Konfidenzintervall (95%)	
Beweismittel	Prädiktoren	*b*	SE *b*	*p*	Unterer Wert	Oberer Wert
Urkunde	Expertise(1)	-.458	7.011	.581	-18.973	14.127
	Expertise(2)	-.285	6.287	.732	-18.545	1.477
	Zeitdruck	.149	4.49	.834	-18.306	1.553
	Delikt	1.21	3.361	.048	-.945	19.611
Aussage des Zeugen	Expertise(1)	.374	.395	.327	-.419	1.29
	Expertise(2)	-.257	.431	.532	-1.145	.576
	Zeitdruck	1.047	.354	.002	.385	1.816
	Delikt	.809	.35	.011	.122	1.683
Aussage des Beschuldigten	Expertise(1)	-.157	.523	.753	-1.287	.86
	Expertise(2)	-.969	.55	.055	-2.17	-.004
	Zeitdruck	1.254	.413	< .001	.424	2.148
	Delikt	1.138	.417	.002	.303	2.074

Anmerkungen. $R^2 = .122$ (Nagelkerke), Urkunde = Auszug aus dem Bundeszentralregister, Expertise(1) = Noviz:innen, Expertise(2) = Expert:innen.

Die Wahrscheinlichkeit, sich für die Aussage des Zeugen als bedeutsamstes Beweismittel zu entscheiden (im Vergleich zum Erscheinungsbild des Beschuldigten),

- stieg für die Laien im Vergleich zu den Noviz:innen, aber sank im Vergleich zu den Expert:innen,
- stieg ohne Zeitdruck,
- stieg beim Delikt „Diebstahl".

Die Expertise-Prädiktoren hatten dabei keinen signifikanten Beitrag, im Gegensatz zu den Ergebnissen des Zeitdrucks (p = .002) und des Delikts (p = .011). Die Ober- und Untergrenze der BCa-Konfidenzintervalle (95%) der beiden signifikanten Prädiktoren überschritten den Wert von 1, sodass keine eindeutige Entscheidungsrichtung vorgegeben werden konnte.

Die Wahrscheinlichkeit, sich für die Aussage des Beschuldigten als bedeutsamstes Beweismittel zu entscheiden (im Vergleich zum Erscheinungsbild des Beschuldigten),

- sank für die Laien im Vergleich zu den Noviz:innen und Expert:innen,
- stieg ohne Zeitdruck,
- stieg beim Delikt „Diebstahl".

Der Zeitdruck (p < .001) und das Delikt (p = .002) erwiesen sich auch hier als signifikant. Das Überschreiten des Wertes von 1 der BCa-Konfidenzintervalle (95%) ließ aber keine eindeutige Richtung vorgeben. Für die Expert:innen wurde das Signifikanzniveau knapp verpasst (p = .055).

Mit Blick auf das Beweismittel mit der *niedrigsten* Relevanz kann die Fragestellung F5 dahingehend beantwortet werden, dass nur die Expertise (Expert:innen) einen signifikanten Prädiktor für das gesamte Modell darstellte. Hinsichtlich der einzelnen Parameter galt dies nur teilweise. Folgende Aussagen können getroffen werden:

- Die Wahrscheinlichkeit, die Urkunde als am wenigsten relevant einzuschätzen, sank für die Laien (im Vergleich zu den Expert:innen) sowie beim Delikt „Diebstahl".
- Die Wahrscheinlichkeit, die Zeugenaussage als am wenigsten relevant einzuschätzen, sank ohne Zeitdruck sowie beim Delikt „Diebstahl.
- Die Wahrscheinlichkeit, die Aussage des Beschuldigten als am wenigsten relevant einzuschätzen, sank beim Delikt „Diebstahl".

Betrachtet man die *höchste* Relevanz, so ist festzustellen, dass der Zeitdruck sowie das Delikt einen signifikanten Beitrag für das gesamte Modell leisteten. Hier können für die einzelnen Parameter folgende Aussagen gemacht werden:

- Die Wahrscheinlichkeit, die Urkunde als am relevantesten einzuschätzen, stieg beim Delikt „Diebstahl".
- Die Wahrscheinlichkeit, die Zeugenaussage als am relevantesten einzuschätzen, stieg ohne Zeitdruck und beim Delikt „Diebstahl".
- Die Wahrscheinlichkeit, die Aussage des Beschuldigten als am relevantesten einzuschätzen, stieg ohne Zeitdruck und beim Delikt „Diebstahl", wohingegen das Ergebnis der Expert:innen das Signifikanzniveau knapp verfehlte.

4.11 F6 (binäre logistische Regression)

Die F6 befasst sich mit der explorativen Fragestellung, ob die Beweismittel mit der höchsten und niedrigsten Relevanz die Entscheidung von Laien, Noviz:innen und Expertinnen hinsichtlich des Tatverdachts vorhersagen.

4.11.1 Deskriptive Auswertung der F6

Aufgrund der hohen Komplexität der Kontigenztabellen – begründet in der Vielzahl der Zellen und der daraus resultierenden Kombinationsmöglichkeiten – wird an dieser Stelle auf eine derartige Darstellung verzichtet und sich lediglich auf Kernbeobachtungen fokussiert. Zur besseren Überschaubarkeit werden die Beweismittel mit niedrigster beziehungsweise höchster Relevanz getrennt betrachtet. Da die Antworttendenzen mitunter auf sehr kleinen Zellgrößen basieren und lediglich als Beschreibungen dieser Tendenzen betrachtet werden sollten, werden anstatt der Prozentzahlen die jeweiligen Anzahlen in Relation zur Gesamtzahl angegeben, um ein besseres Bild zu vermitteln (s. auch 4.10.1).

Gruppenübergreifend wurde dem Bundeszentralregisterauszug als Urkundenbeweis am häufigsten die *niedrigste* Relevanz beigemessen (159/251), in absteigender Häufigkeit gefolgt vom Erscheinungsbild des Beschuldigten als Augenscheinbeweis (46/251), der Aussage des Beschuldigten (32/251) sowie der Aussage des Zeugen (14/251).

Wurde der Bundeszentralregisterauszug als Beweismittel mit der niedrigsten Bedeutsamkeit angegeben,

- so war die gruppenübergreifende Verteilung der Entscheidung über den Tatverdacht nahezu gleichverteilt („Nein“: 80/159, „Ja“: 79/159),
- so tendierten Expert:innen (49/71) insgesamt eher zur Verneinung des Tatverdachts, im Gegensatz zu Laien (6/27) und Noviz:innen (25/61),
- so blieben diese unterschiedlichen Tendenzen von Expert:innen, im Gegensatz zu Noviz:innen und Laien, sowohl unabhängig vom Zeitdruck als auch vom Delikttyp bestehen.

Wurde das Erscheinungsbild des Beschuldigten als Beweismittel mit der niedrigsten Bedeutsamkeit angegeben,

- so entschieden sich Proband:innen eher für die Verneinung des Tatverdachts (28/46),
- so wiesen Laien (9/17 vs. 8/17) und Expert:innen (4/8 vs. 4/8) im Gegensatz zu den Noviz:innen (Verneinen: 15/21 vs. 6/21) keine klare Tendenz auf,
- so führte Zeitdruck zu keiner klareren Entscheidungstendenz bei den Expert:innen (jeweils 2/4) , wohingegen sich Laien (7/11) und Noviz:innen (9/11) ohne Zeitdruck eher gegen das Vorliegen des Tatverdachts entschieden und lediglich Laien mit Zeitdruck zum Bejahen tendierten (4/6),
- so wurde sich gruppenübergreifend beim Diebstahl gegen das Vorliegen des Tatverdachts entschieden (19/29), während dies bei der Körperverletzung weniger eindeutig war (9/17).

Wurde die Aussage des Beschuldigten als Beweismittel mit der niedrigsten Bedeutsamkeit angegeben,

- so entschieden sich Proband:innen eher für das Bejahen des Tatverdachtes (20/32),
- so entschieden sich sowohl Laien (11/14) als auch Noviz:innen (9/13) für, aber Expert:innen (5/5) gegen das Vorliegen des Tatverdachtes,
- so änderte die Zeitdruckmanipulation nur die Entscheidungstendenz der Noviz:innen ohne Zeitdruck, die nun eine 50/50-Verteilung aufwiesen,
- so schien der Delikttyp nicht die Laien und Expert:innen zu beeinflussen (Tatverdacht wurde mehrheitlich verneint), aber die Noviz:innen, die den Tatverdacht bei der Körperverletzung geschlossen bejahten (7/7) und beim Diebstahl verneinten (4/6).

Wurde die Aussage des Zeugen als Beweismittel mit der niedrigsten Bedeutsamkeit angegeben,

- so lag insgesamt keine klare Antworttendenz vor (7 vs. 7),
- so gaben Expert:innen und Laien konträre Antworten (3/3 Expert:innen verneinten, 1/5 Laien verneinten), wohingegen Noviz:innen keine klare Tendenz hatten (3/6),

- so ähnelten Noviz:innen unter Zeitdruck (3/4) mit dem Verneinen eher den Expert:innen (2/2), im Gegensatz zu den Laien (2/5),
- so entschieden sich Noviz:innen ohne Zeitdruck (2/2) für das Vorliegen des Tatverdachtes, im Vergleich zur Einzelstimme der Expert:innen, die verneinte (1/1),
- so lagen weder beim Delikt „Diebstahl“ (2/4) noch bei der Körperverletzung (4/8) klare Entscheidungstendenzen vor.

Gruppenübergreifend wurde der Aussage des Zeugen (129/251) als Beweismittel am häufigsten die *höchste* Relevanz beigemessen, in absteigender Häufigkeit gefolgt vom Erscheinungsbild (60/251) und der Aussage des Beschuldigten (53/251) sowie dem Bundeszentralregisterauszug (9/251). Diese Häufigkeiten, wenngleich die Ränge 2 und 3 vertauscht sind, passen gut zu der Verteilung der Beweismittel, denen die niedrigste Relevanz zugeschrieben wurde.

Wurde die Aussage des Zeugen als Beweismittel mit der höchsten Bedeutsamkeit angegeben,

- so zeigte sich gruppenübergreifend keine klare Antworttendenz, da nur die Hälfte der Personen den Tatverdacht verneinte (67/129),
- so wiesen Laien und Expert:innen eindeutige, aber konträre Tendenzen in ihren Entscheidungen auf (11/37 Laien verneinten, 35/45 Expert:innen verneinten), wobei Noviz:innen eher den Laien ähnelten und verneinten (21/47),
- so blieben diese Richtungen der Entscheidungen für alle Gruppen unabhängig vom Zeitdruck bestehen, außer für die Noviz:innen ohne Zeitdruck, die zu einer 50/50-Entscheidung kamen,
- so beeinflusste der Delikttyp „Diebstahl“ nur die Noviz:innen, indem sie sich nun mehrheitlich gegen den Tatverdacht entschieden (13/23) und den Expert:innen ähnelten.

Wurde das Erscheinungsbild des Beschuldigten als Beweismittel mit der höchsten Bedeutsamkeit angegeben,

- so entschieden sich Proband:innen gruppenübergreifend für das Vorliegen des Tatverdachtes (43/60),
- so tendierten alle Expertise-Gruppen zur Entscheidung für das Vorliegen des Tatverdachtes (Laien: 14/16, Noviz:innen: 19/28, Expert:innen: 10/16),
- so führte (fehlender) Zeitdruck nicht zu einer Änderung der Entscheidungsrichtung (wenngleich sich die relativen Mehrheitsanteile mitunter änderten),
- so beeinflusste auch der Delikttyp nicht die Entscheidungsrichtung.

Wurde die Aussage des Beschuldigten als Beweismittel mit der höchsten Bedeutsamkeit angegeben,

- so entschieden sich Proband:innen gruppenübergreifend gegen das Vorliegen des Tatverdachtes (37/53),
- so wiesen alle Expertise-Gruppen diese Antworttendenz auf (Laien: 5/8, Noviz:innen: 14/21, Expert:innen: 18/24),
- so beeinflusste der Zeitdruck nur die Entscheidungsrichtung der Laien, die sich unter Zeitdruck mehrheitlich für das Vorliegen des Tatverdachtes entschieden (2/3),
- so beeinflusste der Delikttyp für keine der Gruppen die Entscheidungsrichtung.

Wurde der Bundeszentralregisterauszug als Beweismittel mit der höchsten Bedeutsamkeit angegeben,

- so tendierten die Proband:innen gruppenübergreifend eher dazu den Tatverdacht zu verneinen (6/9),
- so wiesen die Laien (1/2) keine Tendenz auf, im Gegensatz zu den Noviz:innen (3/5) und den Expert:innen (2/2), die den Tatverdacht mehrheitlich verneinten,
- so war die allgemeine Tendenz derjenigen ohne Zeitdruck, den Tatverdacht zu bejahen (2/3), im Vergleich zu denjenigen mit Zeitdruck (1/6),
- so konnte für die Delikttypen kein aussagekräftiger Vergleich gemacht werden, da zu viele Zellen auf Seiten der Körperverletzung leer geblieben sind.

4.11.2 Inferenzstatistische Auswertung der F6

Die Gruppe der Laien, die Manipulation „ohne Zeitdruck“ sowie das Delikt „Diebstahl“ galten weiterhin als Referenzgruppen (s. 4.2.2). Für die Beweismittel mit niedrigster beziehungsweise höchster Relevanz wurde der Augenschein, also das Erscheinungsbild des Beschuldigten, als Referenzgruppe festgelegt. Da keine theoretische Ableitung darüber vorlag, welches der vier Beweismittel eine gute Referenzgruppe darstellt, wurde aufgrund der Einheitlichkeit mit dem Mittel begonnen, das auch im Rahmen der Studie an erster Stelle abgefragt worden war (s. auch 4.10.3). Die Prädiktoren wurden mittels „Einschluss“-Variante dem Modell hinzugefügt.

Anhand der bekannten Kriterien wurden die Residuen untersucht, um die Passung zwischen Daten und Modell und das Vorhandensein einflussreicher Fälle zu überprüfen (s. 4.2.2). Die Einflussstatistik nach Cook sowie die DFBeta-Werte für die

Konstante sowie das Modell lagen für alle Fälle jeweils unter 1. Daraus kann geschlossen werden, dass keine Fälle das Modell übermäßig beeinflussten. Der maximale Wert der standardisierten Residuen war mit 3.04 auffällig. Insgesamt vier Fälle überschritten den Wert von 1.96, was einem Anteil von 1.6% entspricht. Laut Field (2013) spricht ein Wert < 5% aber dafür, dass die Passung zwischen Modell und Daten noch angemessen ist. Der durchschnittlich zu erwartende Hebelwert lag bei einem Wert von .023. Insgesamt 26 Fälle überschritten die dreifache Höhe des Durchschnitts. Diese Fälle schienen somit Ausreißer in Bezug auf die Prädiktoren zu sein. Laut Stevens (2002) müssen identifizierte Ausreißer nicht notwendigerweise aus dem Datensatz entfernt werden. Aufgrund der teilweise bereits niedrigen Werte in den einzelnen Zellen wurde auf das Bereinigen der Daten verzichtet, da dies das Problem der fehlenden Werte weiter verstärkt hätte (s. auch 4.4.2). Die Analysen deuteten auf das Vorhandensein von Ausreißern im Regressionsmodell hin, wenngleich eine gewisse Passung zwischen Modell und Daten durchaus angenommen werden konnte.

Fügte man die Prädiktoren dem Modell hinzu, so ließ sich eine signifikante Verbesserung des Modells ablesen (p <.001). Die anfängliche -2-Log-Likelihood reduzierte sich von 347.924 im Null-Modell auf 47.525 im Modell 1. Der Gesamtprozentsatz der Richtigen stieg auf 72.9%. Der Hosmer-Lemeshow-Test war aufgrund der Nicht-Signifikanz (p =.506) ein Hinweis darauf, dass Modell und Daten gut zueinander passten.

Da die Voraussetzungen für das Testverfahren teilweise verletzt waren (s. 3.8.1.4), wurde sich für den Einsatz von Bootstrapping entschieden.[42] Bei den Noviz:innen und auch bei den Expert:innen sank (im Vergleich zu den Laien) die Wahrscheinlichkeit, den Tatverdacht zu bejahen (s. Tabelle 4.16). Beide Ergebnisse erreichten das Signifikanzniveau (ps < .05). Diejenigen mit Zeitdruck entschieden sich laut Regressionskoeffizient mit etwas höherer Wahrscheinlichkeit für die Bejahung des Tatverdachtes als diejenigen ohne Zeitdruck. Mit Blick auf den Delikttyp stieg für die Körperverletzung die Wahrscheinlichkeit der Bejahung des Tatverdachtes (im

[42] Siehe Fußnote 38.

Vergleich zum Diebstahl). Die Prädiktoren „Zeitdruck“ und „Delikt“ verfehlten aber das Signifikanzniveau (*p*s > 0.05).

Tabelle 4.16. Regressionsergebnisse der F6 für die Prädiktoren (1000 Bootstrapping-Stichproben)

Prädiktoren	*b*	SE *b*	*p*	BCa Konfidenzintervall (95%) Unterer Wert	Oberer Wert
Expertise(1)	-.82	.379	.027	-1.449	-.331
Expertise(2)	-1.9	.433	< .001	-2.662	-1.386
Zeitdruck(1)	.075	.315	.811	-.572	.693
Delikt(1)	.277	.311	.346	-.361	.972
niedrig(1)	.647	.452	.131	-.261	1.734
niedrig(2)	.074	1.273	.906	-2.028	2.222
niedrig(3)	.345	.581	.523	-.895	1.691
hoch(1)	-1.273	4.914	.101	-20.642	.276
hoch(2)	-.962	.398	.009	-1.73	-.33
hoch(3)	-1.499	.526	.002	-2.457	-.758

Anmerkungen. R^2 = .23 (Nagelkerke), Expertise(1) = Noviz:innen, Expertise(2) = Expert:innen, Zeitdruck(1) = mit, Delikt(1) = Körperverletzung, niedrig/hoch(1) = Auszug aus dem Bundeszentralregister, niedrig/hoch(2) = Aussage des Zeugen, niedrig/hoch(3) = Aussage des Beschuldigten.

Wurde der Bundeszentralregisterauszug, die Aussage des Zeugen oder des Beschuldigten als Beweismittel mit der niedrigsten Relevanz ausgewählt (im Vergleich zum Augenschein), stieg die Wahrscheinlichkeit der Bejahung des Tatverdachtes. Der Prädiktor „Beweismittel mit niedrigster Relevanz“ war aber in keinem Fall signifikant (*p*s > .05). Wurde der Bundeszentralregisterauszug, die Aussage des Zeugen oder des Beschuldigten als Beweismittel mit der höchsten Relevanz ausgewählt (im Vergleich zum Augenschein), sank die Wahrscheinlichkeit der Bejahung des Tatverdachtes. Im Falle der als relevant eingestuften Aussagen des Zeugen und des Beschuldigten erwiesen sich die Prädiktoren als signifikant (*p*s < .05). Die Effekte waren klein (s. auch Nagelkerkes R^2). Laut den BCa-Konfidenzintervallen (95%) der signifikanten Prädiktoren konnte deren Entscheidungsrichtung eindeutig vorgegeben werden.

Mit Blick auf das Beweismittel mit der *niedrigsten* Relevanz kann die Fragestellung F6 dahingehend beantwortet werden, dass sich, im Vergleich zur Referenzgruppe des Augenscheins, keines der Beweismittel als signifikanter Prädiktor für die Entscheidung über das Vorliegen des Tatverdachtes erwiesen hat. Betrachtet man die *höchste* Relevanz, so ist festzustellen, dass die Wahl der Aussage des Zeugen oder

des Beschuldigten mit signifikant geringerer Wahrscheinlichkeit zur Bejahung des Tatverdachtes führte. Unabhängig von der Fragestellung ergab sich zudem die Expertise als signifikanter Prädiktor, da Noviz:innen und Expert:innen weniger wahrscheinlich den Tatverdacht bejahten als die Referenzgruppe der Laien.

4.12 Exkurs: Explorative Auswertung der qualitativen Daten zu Nachermittlungen und zur Entscheidungsqualität

Die Studienteilnehmenden bekamen an zwei Stellen im Versuchsablauf (s. 3.2) die Möglichkeit, optionale Angaben zu offenen Fragen zu machen. Diese bezogen sich auf eventuelle Nachermittlungen (s. 4.12.1) und auf Kriterien einer schlechten Entscheidung (s. 4.12.2). Beide offenen Fragen standen nicht im Fokus der Auswertungen, weswegen sie an dieser Stelle in einem Exkurs ausgeführt werden.

4.12.1 Nachermittlungen im Strafprozess

Die Auswertungen der Angaben zu den Nachermittlungen erfolgte in Anlehnung an die vier Beweiskategorien (s. 2.1.6). Dies bedeutet, dass die gemachten Angaben von der Autorin dahingehend bewertet wurden, welcher Beweiskategorie sie am ehesten entsprechen würden, sollten sie ermittelt werden. Sofern eine Angabe nicht eindeutig zugeordnet werden konnte, wurde sie als „Sonstiges“ kodiert. Es wurde die Anzahl der Antworten für die jeweilige Kategorie berechnet. In der für die Delikte separaten Auswertung wird zwischen den Expertise- und Zeitdruck-Gruppen unterschieden.[43]

Insgesamt äußerten 50 Teilnehmende Vorschläge zu möglichen Nachermittlungen im Fall „Diebstahl“. Differenziert nach Expertise-Gruppen lässt sich erkennen, dass die meisten davon Expert:innen waren (N = 25), gefolgt von Noviz:innen (N = 14) und Laien (N = 11). Diese 50 Personen machten 83 Angaben zu Nachermittlungen, die den Beweiskategorien zugeordnet wurden. Auch hier lagen die Expert:innen in der Summe der Antworten (52) vor den Noviz:innen (18) und Laien (13). In der

[43] Auch wenn die Frage nach Ermittlungsmethoden nicht unter dem Einfluss der Zeiterfassung erhoben wurde, so war dies für das Lesen der Vignette der Fall, auf deren Inhalt sich die Erhebung dieser qualitativen Daten explizit bezieht. Aus diesem Grunde findet ein differenzierter Blick auf die Zeitdruck-Level statt.

Tabelle 4.17 ist dargestellt, wie sich die Anzahlen der genannten Antworten je Beweiskategorie auf die Zeitdruck- und Expertise-Gruppen verteilen.

Tabelle 4.17. Anzahl der gemachten Angaben je Beweiskategorie zu Nachermittlungen im Fall „Diebstahl“ in Abhängigkeit von Expertise und Zeitdruck

Beweiskategorie	Expertise							
	Laien		Noviz:innen		Expert:innen		Gesamt	
	o.Z.	m.Z.	o.Z.	m.Z.	o.Z.	m.Z.	o.Z.	m.Z.
Augenschein (14)	2	1	1	1	3	6	6	8
Urkunde (32)	1	5	6	3	8	9	15	17
Zeug:innen (20)	1	1	1	2	6	9	8	12
Beschuldigter (14)	1	1	2	1	1	8	4	10
Sonstiges (3)	-	-	1	-	2	-	3	-
Gesamt (83)	5	8	11	7	20	32	36	47

Anmerkungen. o. Z. = ohne Zeitdruck. m. Z. = mit Zeitdruck. Die genannten Anzahlen beziehen sich nicht auf die Anzahl der Personen, sondern auf die der gemachten Angaben. Eine Person kann dabei mehrere Angaben gemacht haben, die wiederum auf einzelne Beweiskategorien aufgeteilt wurden.

Mit Abstand die meisten vorgeschlagenen Ansatzpunkte für Nachermittlungen bezogen sich auf eine Form der Urkunde, gefolgt von Aussagen von Zeug:innen, dem Augenschein und Nachbefragungen des Beschuldigten. Als Urkundenbeweis gewertet wurden beispielsweise die folgenden Inhalte: Seriennummer der Kopfhörer, Fingerabrücke am Auto, Kontoauszüge oder andere Kauf- und Zahlungsnachweise. Aussagen von Zeug:innen bezogen sich in der Regel nicht auf den bereits bekannten Zeugen, sondern auf Personen, die in dem Markt arbeiten, in dem angeblich die Kopfhörer gekauft wurden, oder auf Personen, die die Geldschenkung zum Geburtstag oder den Besitz der Kopfhörer belegen können. Angaben zum Augenschein beinhalteten zumeist das Vorhandensein von Überwachungskameras im Park, im Einkaufsgeschäft (zum Zeitpunkt des Kaufs) und auf dem Parkplatz. Es wurden aber auch Ermittlungen zum Abnutzungszustand der Kopfhörer oder zur räumlichen Entfernung vom Haus des Beschuldigten zum Park vorgeschlagen. Nachbefragungen des Beschuldigten haben laut den Antworten insbesondere den Zweck, nähere Informationen zum Ort des Kaufs zu erhalten (z. B. Name des Marktes). Auch das Tatmotiv wurde als Gegenstand von Nachermittlungen genannt, welches sich über die Befragung des Beschuldigten erschließen lassen würde. Als „Sonstiges“ kodierte Antworten waren zwar inhaltlich durchaus relevant, allerdings ließ sich dabei keine eindeutige Kategorie ableiten. Um näheres über die „Um-

stände im Park (belebt/Menschenleer [*sic*])“ herauszufinden, könnten Überwachungskameras zum Einsatz kommen, aber auch (erneute) Vernehmungen weiterer Zeug:innen oder des Beschuldigten stattfinden. Somit war nicht eindeutig, worauf sich die Antwort der teilnehmenden Personen beziehen sollte.

Im Fall „Körperverletzung“ machten 38 Personen insgesamt 64 Angaben zu Nachermittlungen, die den Beweiskategorien zugeordnet wurden. Ebenso wie im Fall „Diebstahl“ stieg mit dem Expertise-Level auch die Anzahl der Teilnehmenden (Laien: $N = 7$, Noviz:innen: $N = 13$, Expert:innen: $N = 18$). Die Expert:innen lagen mit ihrer Anzahl von Angaben gleichauf (jeweils 24), wohingegen die Laien 16 Ideen äußerten und somit nicht allzu weit hinter den Fachgruppen lagen. Die Tabelle 4.18 führt die Anzahlen der genannten Antworten je Beweiskategorie in Abhängigkeit von Zeitdruck und Expertise auf. Hinsichtlich der Verteilung auf die Beweiskategorien zeigt sich, dass keine der Kategorien über- oder unterrepräsentiert war. Die Anzahlen der gemachten Angaben hatten ein vergleichbares Ausmaß und schwankten zwischen 10 (Augenschein) und 17 (Urkunde) Angaben.

Tabelle 4.18. Anzahl der gemachten Angaben je Beweiskategorie zu Nachermittlungen im Fall „Körperverletzung“ in Abhängigkeit von Expertise und Zeitdruck

Beweiskategorie	Expertise							
	Laien		Noviz:innen		Expert:innen		Gesamt	
	o.Z.	m.Z.	o.Z.	m.Z.	o.Z.	m.Z.	o.Z.	m.Z.
Augenschein (10)	5	1	-	3	-	1	5	5
Urkunde (17)	3	-	2	5	4	3	9	8
Zeug:innen (12)	4	-	1	4	1	2	6	6
Beschuldigter (13)	3	-	-	4	2	4	5	8
Sonstiges (12)	-	-	2	3	3	4	5	7
Gesamt (64)	15	1	5	19	10	14	30	34

Anmerkungen. o. Z. = ohne Zeitdruck. m. Z. = mit Zeitdruck. Die genannten Anzahlen beziehen sich nicht auf die Anzahl der Personen, sondern auf die der gemachten Angaben. Eine Person kann dabei mehrere Angaben gemacht haben, die wiederum auf einzelne Beweiskategorien aufgeteilt wurden.

Der häufigste vorgeschlagene Urkundenbeweis beinhaltete Nachermittlungen zur Zusammensetzung der Asche und zum Vergleich mit den Spuren an den Schuhen des Beschuldigten. Auch die Ortung des Mobiltelefons des Beschuldigten zur Tatzeit wurde genannt. Eine Nachbefragung des Beschuldigten soll insbesondere Auskünfte über alternative Erklärungen zur Herkunft der Aschespuren und über die Trainingsgewohnheiten geben. Auch das Motiv wurde mehrfach als Gegenstand

von Nachermittlungen angegeben. Ähnlich viele Angaben bezogen sich auf die Befragung zusätzlicher Zeug:innen, die das Tatgeschehen beobachtet haben, die Aussagen zu den Trainingsgewohnheiten des Beschuldigten (z. B. regelmäßiges Training auf Ascheplätzen) und zum Alibi machen können (z. B. beim Joggen beobachtet), oder die Auskünfte über das Temperament des Beschuldigten geben können. Unter den Augenschein fielen insbesondere Anschauungsmaterial zum Verschmutzungsgrad der Schuhe oder Angaben zum Vorhandensein von Überwachungskameras. Als „Sonstiges" kodierte Antworten waren, ebenfalls wie beim Fall „Diebstahl", nicht eindeutig zuzuordnen. Häufig wurden Nachermittlungen genannt, die Klarheit über eine mögliche Beziehung zwischen Beschuldigtem und Geschädigtem bringen. Dies könnte über eine Befragung des Beschuldigten, des Geschädigten oder der Zeug:innen geschehen. Des Weiteren galt es laut den Teilnehmenden, das Alibi zu überprüfen, aber es wurden keine eindeutigen Angaben gemacht, auf welche Weise dies zu ermitteln ist.

4.12.2 Kriterien der Entscheidungsqualität

Woran würden die Teilnehmenden erkennen, dass sie keine gute Entscheidung getroffen haben? Dies bezieht sich auf die dichotome Frage nach dem Vorliegen des Tatverdachtes (s. 3.4.2). Für die Auswertung der offenen Frage standen zunächst keine Kategorien zur Verfügung, da ihr keine spezifischen Annahmen oder Hypothesen zugrunde lagen. Nach der Sichtung der Angaben durch die Autorin wurde sich für die folgende Einteilung der Antworten entschieden:

- objektive Falsifizierung der Entscheidung, z. B. das Finden neuer Beweismittel, welche die Schuld oder die Unschuld beweisen,
- subjektives Entscheidungserleben, z. B. Zweifel, Dauer der Entscheidung,
- Feedback, z. B. Entscheidungen anderer Instanzen (z. B. Ablehnung des Antrags durch das Gericht) oder Rückmeldungen von Kolleg:innen,
- Sonstiges, z. B. „Andere überzeugende Argumente zur Einschätzung der Indizien", „Fortsetzung der Verfolgung eines in Wahrheit Unschuldigen", „An einer möglicherweise erneuten Straftat des Beschuldigten".

Es wurde die Anzahl der Antworten für die jeweilige Kategorie berechnet. Die Tabelle 4.19 bezieht sich auf den Fall „Diebstahl". Insgesamt 11 Expert:innen, 3 Noviz:innen und 8 Laien äußerten sich in 22 Angaben. Davon konnten allerdings 5

Antworten nicht näher ausgewertet oder als „Sonstiges“ kodiert werden, da sie sich nicht auf die gestellte Frage bezogen (z. B. „Zeugenaussage zu menschen [*sic*], Unschuldiger im Verdacht, nur, weil er gleiche Kopfhörer hatte“). Die gemachten Angaben erinnerten eher an eine Rechtfertigung der getroffenen Entscheidung. Eine Person (Expert:in) bezieht sich in der Begründung ihrer Antwort auf den Erfahrungsschatz: „Die Erfahrung zeigt, dass Zufälle dieser Art, die es hier gebraucht hätte, so gut wie nie gibt“. Eine andere Person (Expert:in) äußerte sich folgendermaßen: „Wenn ich sie nicht für gut halten würde, hätte ich sie nicht getroffen“. Aufgrund der geringen Anzahl lässt sich keine eindeutige Tendenz für oder gegen ein Kriterium ausmachen, wenngleich das Kriterium „Feedback“ die meisten Angaben erhielt. Das Erleben von Zeitdruck veränderte nicht die Antwortrate der Teilnehmenden.

Tabelle 4.19. Einordnung der Entscheidungsqualität im Fall „Diebstahl“ in Abhängigkeit von Expertise und Zeitdruck

Kriterium	Expertise						Gesamt	
	Laien		Noviz:innen		Expert:innen			
	o.Z.	m.Z.	o.Z.	m.Z.	o.Z.	m.Z.	o.Z.	m.Z.
Objektive Falsifizierung (5)	2	2	-	-	-	1	2	3
Subjektives Erleben (4)	1	1	-	1	-	1	1	3
Feedback (7)	4	-	-	-	1	2	5	2
Sonstiges (4)	-	-	-	1	1	2	1	3
Gesamt (20)	7	3	-	2	2	6	9	11

Anmerkungen. o. Z. = ohne Zeitdruck. m. Z. = mit Zeitdruck. Die genannten Anzahlen beziehen sich nicht auf die Anzahl der Personen, sondern auf die der gemachten Angaben. Eine Person kann dabei mehrere Angaben gemacht haben, die wiederum auf einzelne Qualitätskriterien aufgeteilt wurden.

In Tabelle 4.20 sind die Ergebnisse für den Fall „Körperverletzung“ zusammengefasst. Die 27 Angaben stammten von 26 Personen, genauer gesagt von 10 Expert:innen, 6 Noviz:innen und 10 Laien. Der Rücklauf war (ebenso wie für die andere Vignette) sehr gering.

Tabelle 4.20. Einordnung der Entscheidungsqualität im Fall „Körperverletzung“ in Abhängigkeit von Expertise und Zeitdruck

Kriterium	Expertise							
	Laien		Noviz:innen		Expert:innen		Gesamt	
	o.Z.	m.Z.	o.Z.	m.Z.	o.Z.	m.Z.	o.Z.	m.Z.
Objektive Falsifizierung (9)	3	1	-	2	2	1	5	4
Subjektives Erleben (7)	-	2	1	-	2	2	3	4
Feedback (4)	-	1	1	-	1	1	2	2
Sonstiges (4)	1	-	1	1	1	-	3	1
Gesamt (24)	4	4	3	3	6	4	13	11

Anmerkungen. o. Z. = ohne Zeitdruck. m. Z. = mit Zeitdruck. Die genannten Anzahlen beziehen sich nicht auf die Anzahl der Personen, sondern auf die der gemachten Angaben. Eine Person kann dabei mehrere Angaben gemacht haben, die wiederum auf einzelne Qualitätskriterien aufgeteilt wurden.

Aus den gleichen Gründen wie im Fall „Diebstahl“ konnten 3 von 27 Angaben nicht weiter ausgewertet werden (z. B. „zu wenig Beweise“). Eine Person (Expert:in) gab an, dass insbesondere die eigene Einschätzung des Beschuldigten wichtig sei. Sie fasste es folgendermaßen zusammen: „Das Problem ist, dass der persönliche Eindruck fehlt und die Entscheidung ‚auf dem Papier‘ leicht fällt, gut aber auch durch den persönlichen Eindruck revidiert werden kann“. Im Vergleich zum Fall des Diebstahls sprachen sich die Teilnehmenden vermehrt für das Kriterium des subjektiven Erlebens aus. Eine Fokussierung auf ein bestimmtes Kriterium lässt sich nicht erkennen, wenngleich Angaben zur objektiven Falsifizierung tendenziell häufiger genannt wurden. Ebenso wie beim Delikt des Diebstahls reduzierte das Erleben von Zeitdruck nicht die Antwortrate der Teilnehmenden, da die Anzahlen der Angaben zwischen den Zeitdruck-Gruppen zumindest auf deskriptiver Ebene vergleichbar waren.

5 Diskussion

Zunächst werden die Ziele und die Ergebnisse der Studie zusammengefasst (s. 5.1). Besagte Ergebnisse werden anschließend im Zusammenhang mit dem theoretischen und empirischen Hintergrund diskutiert (s. 5.2). Eine kritische Betrachtung der Methode sowie Angaben zu Limitationen finden sich in Abschnitt 5.3. Im Abschnitt 5.4 wird auf die Implikationen der Studie und Forschungsdesiderate eingegangen, woraufhin ein Fazit erfolgt (s. 5.5).

5.1 Zusammenfassung der Studienziele und Ergebnisse

Im Fokus dieser Studie stand das Ermittlungsverfahren als erste Phase des Strafverfahrens, an dessen Ende die Staatsanwaltschaft anhand der vorliegenden Beweislage (prozessweisende) Entscheidungen über das Vorliegen des Tatverdachtes und über weitere Verfahrensschritte trifft. Ein Ziel war es, prozess- und personenbedingte Faktoren dahingehend zu untersuchen, welchen Einfluss sie auf diese Entscheidungen haben. Die Studienteilnehmenden wurden gemäß ihrer juristischen Vorbildung als Laie, Noviz:in oder Expert:in eingeteilt. Eine Teilstichprobe erhielt eine Fallbeschreibung über einen Diebstahl, die andere Teilstichprobe befasste sich mit dem Delikt der Körperverletzung. Zudem wurde der eine Teil der Proband:innen unter Zeitdruck gesetzt, während der andere Teil zur sorgfältigen Betrachtung der Beweislage aufgefordert wurde. Auf Grundlage der nicht eindeutigen Beweislage, die den Teilnehmenden in Form der Vignetten präsentiert wurde, galt es zu erforschen, inwiefern sich die (Quasi-)Experimentalgruppen in ihren Entscheidungsergebnissen und anderen damit verbundenen Merkmalen voneinander unterscheiden und welche Vorhersagen diesbezüglich getroffen werden können.

Zunächst werden die Ergebnisse der konfirmatorischen Analysen zusammengefasst (H1–H4). Von zentraler Bedeutung für diese Studie war die Frage nach dem Vorliegen des Tatverdachtes. Anhand der Fallbeschreibung sollte sich für oder gegen diesen entschieden werden (H1). Im Einklang mit der Hypothese zeigte sich ein Effekt der Expertise, denn sowohl die Noviz:innen als auch die Expert:innen bejahten mit geringerer Wahrscheinlichkeit den Tatverdacht als Naive – wenngleich der Effekt nur klein war. Entgegen der Hypothese machte das Erleben von Zeitdruck dabei keinen Unterschied, der Inhalt der Vignette dagegen schon: Lag der Fall der

Körperverletzung vor, so war die Wahrscheinlichkeit, den Tatverdacht zu bejahen, insgesamt höher. Das individuelle Beweismaß ist ausschlaggebend dafür, wie überzeugend die vorliegende Beweislage empfunden wird (und ob dementsprechend der Tatverdacht gesehen wird oder nicht). Daher wurde gemäß der H2 untersucht, inwiefern sich die (Quasi-)Experimentalgruppen diesbezüglich ähneln oder unterscheiden. Es konnten entgegen der Erwartungen keine signifikanten Unterschiede für die Ausprägungen der Expertise festgestellt werden, wenngleich sich andeutete, dass die Laien insgesamt überzeugter von der Beweislage waren. Im Einklang mit der Hypothese waren dagegen die fehlenden Effekte des Zeitdrucks und des Delikttyps. Wird der Tatverdacht bejaht, gilt es, einen nächsten Verfahrensschritt zu bestimmen, der in dieser Studie entweder als das Erheben einer Anklage oder als Einstellung des Verfahrens (wegen Geringfügigkeit oder unter Auflagen und Weisungen) definiert wurde. Die sich darauf beziehende H3 wurde nicht bestätigt, da die Beantwortung der Frage nach dem nächsten Verfahrensschritt nicht in signifikanter Abhängigkeit zum Delikttyp, zur Zeitdruck- oder zur Expertise-Gruppe (oder einer Interaktion davon) stand und die Effekte zudem klein waren. Die H4 befasste sich damit, wie der Schweregrad des jeweiligen Delikts eingeschätzt wird. Der Zeitdruck spielte keine Rolle, was im Einklang mit der Hypothese war. Die Expert:innen schätzten die Schwere der Körperverletzung erwartungsgemäß geringer ein als die Naiven, wohingegen sich die Naiven und Noviz:innen wider Erwarten nicht unterschieden. Für den Diebstahl zeigten sich gemäß der Hypothese keine Unterschiede.

In den exploratorischen Analysen (F1–F6) wurde zunächst die Frage untersucht, ob sich die Gruppen in bestimmten Merkmalen unterscheiden, die den Entscheidungsprozess über das Vorliegen des Tatverdachtes betreffen (F1). Es zeigte sich der teils große Effekt der Expertise, demzufolge sich alle drei Gruppen voneinander unterschieden. Expert:innen fiel die Entscheidung leichter und sie fühlten sich sicherer und überzeugter als die Vergleichsgruppen. Auf Noviz:innen traf dies im Vergleich zur Gruppe der Naiven ebenfalls zu. Weder das Delikt noch der Zeitdruck spielten hier eine signifikante Rolle. Als mögliche personenbedingte Einflussfaktoren galten in dieser Studie die kognitive Reflexion sowie das Kognitionsbedürfnis (F2). Es zeigten sich keine signifikanten Unterschiede zwischen den Expertise-Gruppen hin-

sichtlich der Reflexionsfähigkeit. Für Need for Cognition ergab sich ein kleiner Effekt der Expertise. Dies ließ sich allerdings nur auf deskriptiver Ebene weiterführend interpretieren, da die Noviz:innen eine größere Denkmotivation angaben als die Vergleichsgruppen. Fraglich war außerdem, welche personen- und prozessbedingten Faktoren die Entscheidung über den Tatverdacht vorhersagen (F3). Es stellte sich heraus, dass nur die Expertise als relevant galt, wohingegen die kognitive Reflexion, Need for Cognition, der Delikttyp, der Zeitdruck sowie die genutzte Lesezeit keinen signifikanten Einfluss hatten. Im Vergleich zu den Laien sank die Wahrscheinlichkeit, den Tatverdacht zu bejahen, mit der Zugehörigkeit zur Gruppe der Expert:innen – das Ergebnis der Noviz:innen verpasste knapp das Signifikanzniveau. Allerdings war der Effekt als klein einzustufen. Weiterführend befasste sich die F4 mit den gleichen Prädiktoren wie die F3, um zu untersuchen, ob diese die Anzahl der genutzten Beweismittel (Augenschein, Urkunde, Zeugenaussage, Beschuldigtenaussage) vorhersagen können. Die Nutzung vieler Informationselemente wurde von einer hohen Ausprägung der kognitiven Reflexion sowie einer langen Lesezeit signifikant vorhergesagt (kleine Effekte). Welches Beweismittel dabei als bedeutsam oder unbedeutsam eingeschätzt wurde, war nicht von den gleichen Prädiktoren abhängig (F5).[44] Ob ein Beweismittel als unbedeutsam galt, wurde durch den Prädiktor „Expertise“ (Expert:innen) vorhergesagt. Das Delikt verpasste als Prädiktor knapp das Signifikanzniveau. Für eine hohe Bedeutsamkeit waren dagegen der Delikttyp und der Zeitdruck relevant. Mit Blick auf einzelne Parameter zeigte sich, dass beim Delikt „Diebstahl“ im Vergleich zur Referenzkategorie die Wahrscheinlichkeit sank, entweder die Urkunde, die Aussage des Zeugen oder des Beschuldigten als am unbedeutsamsten einzustufen. Bei diesem Delikt stieg wiederum die Wahrscheinlichkeit, entweder die Urkunde, die Aussage des Zeugen oder des Beschuldigten als bedeutsam zu bewerten. Fehlte der Zeitdruck, so sank einerseits die Wahrscheinlichkeit, die Zeugenaussage als unbedeutsam einzustufen. Andererseits stieg die Wahrscheinlichkeit, die Aussage des Zeugen oder des Beschuldigten als am bedeutsamsten einzuschätzen. Im Vergleich mit den Fachpersonen sank für die Laien die Wahrscheinlichkeit, die Urkunde als am wenigsten relevant einzuordnen. Abschließend befasste sich die F6 mit der Frage, ob

[44] Aufgrund der problematischen Zellgrößen sind die inferenzstatistischen Ergebnisse der F5 nur unter Vorbehalt anzunehmen (s. 4.10.3).

das Beweismittel mit der höchsten beziehungsweise niedrigsten Relevanz die Entscheidung über das Vorliegen des Tatverdachtes vorhersagt. Keines der Beweismittel mit der niedrigsten Relevanz erwies sich als signifikanter Prädiktor. Wurde der Aussage des Zeugen oder des Beschuldigten die höchste Bedeutsamkeit zugemessen, sank die Wahrscheinlichkeit zur Bejahung des Tatverdachtes. Ebenfalls sank, unabhängig von der F6, die Wahrscheinlichkeit zur Bejahung mit der Zugehörigkeit zu einer der beiden Gruppen mit juristischer Vorbildung.

5.2 Einordnung der Ergebnisse vor dem theoretischen und empirischen Hintergrund

Die Einordnung der Ergebnisse orientiert sich an der Darstellung des theoretischen und empirischen Hintergrundes, der abschnittsweise die Einflussfaktoren betrachtet (s. 2). Daher wird aufeinanderfolgend auf die jeweiligen Erkenntnisse zu den einzelnen Faktoren eingegangen, hier beginnend mit der Expertise (s. 5.2.1), gefolgt vom Zeitdruck (s. 5.2.2), von kognitiver Reflexion (s. 5.2.3) und dem Kognitionsbedürfnis (s. 5.2.3). Anschließend gilt es, den Delikttyp (s. 5.2.5) sowie die Einordnung der Ergebnisse hinsichtlich des Beweismaßes (s. 5.2.6), des Schweregrades (s. 5.2.7) und der Beweismittel (s. 5.2.8) zu diskutieren. Zum Schluss erfolgen Ausführungen zu den qualitativen Analysen (s. 5.2.9).

5.2.1 Die Bedeutung der Expertise

Das signifikante Ergebnis der Haupthypothese lautet, dass sowohl juristische Noviz:innen als auch Expert:innen mit geringerer Wahrscheinlichkeit den Tatverdacht bejahten als Laien (H1; s. 4.2.2). Insbesondere die beiden „Extremgruppen“ handelten auf deskriptiver Ebene konträr: Laien bejahten, Expert:innen verneinten. Die Gruppe der Noviz:innen wies dagegen keine klare Entscheidungstendenz auf, was sich auch an den ermittelten Konfidenzintervallen ablesen ließ. Innerhalb der Gruppe der Fachpersonen war der Konsens über das Vorliegen des Tatverdachtes dabei eindeutig, da knapp über zwei Drittel dieser Teilstichprobe die gleiche Antwort gaben (Verneinung). Gleichermaßen eindeutig war die Übereinstimmung der Naiven, bei denen prozentual gesehen ähnlich viele Personen die gleiche Antwort gaben (Bejahen). Diese Diskrepanz zwischen den beiden Gruppen weist daraufhin, dass es relevante Merkmale geben muss, die zu diesem konträren Antwortverhalten

führen. Das Ergebnis lässt den Schluss zu, dass die Expertise ein solch relevantes Merkmal darstellen könnte. Zu der Tatsache, dass sich derartige Unterschiede in der Einschätzung des Tatverdachtes in Abhängigkeit von Expertise zeigten, passen die Befunde der Fragestellungen F3 und F6, dass ebenjener Faktor einen signifikanten Prädiktor für diese Einschätzung darstellte: Die Zugehörigkeit zur Gruppe der Expert:innen und auch der Noviz:innen (knapp nicht signifikant und nur minimal auffälliges Konfidenzintervall) reduzierte die Wahrscheinlichkeit, dem Vorliegen des Tatverdachtes zuzustimmen (s. 4.8.2; 4.11.2).

Laut Shanteau (1988) unterscheiden sich Menschen unterschiedlicher Expertise in ihrer Fähigkeit, relevante und irrelevante Informationen zu differenzieren (s. auch Brams et al., 2019). Es ist denkbar, dass sich in den Vignetten, die eine nicht eindeutige Beweislage präsentierten, gewisse Informationselemente als – aus juristischer Fachperspektive – wenig relevant herausgestellt haben (s. auch 4.10.2). Dies könnte zu deren recht eindeutiger Antworttendenz geführt haben, da die Expert:innen möglicherweise auf vergleichbare erfahrungsbasierte Verknüpfungen der Beweismittel zurückgegriffen haben und dementsprechend zu einem ähnlichen Ergebnis kamen. Das erfahrungsbasierte Wissen, welches Naive und Expert:innen unterscheidet (Herbig & Glöckner, 2009; Shanteau, 1992), könnte hier insofern zum Tragen kommen, als dass sich Fachpersonen in ihrer Bewertung der Beweismittel auf ebenjenen Erfahrungs- und Wissensschatz verlassen haben (s. auch Mitchell, 1989). Ebenso ist es möglich, dass die Expert:innen auf ähnliche (heuristische) Entscheidungsstrategien zurückgegriffen haben, aus denen eine solche Übereinstimmung resultierte. Dazu passt das deskriptive Ergebnis, dass den Expert:innen die in den Vignetten behandelten Delikte eher vertraut waren (s. 3.3). Für diese Fähigkeiten von Expert:innen, scheinbar relevante von irrelevanten Informationen zu unterscheiden, gab es Hinweise in dieser Studie, da deren Expertise einen signifikanten Prädiktor für die Wahl des Beweismittels mit der niedrigsten Relevanz darstellte (F5; s. 4.10.3). Für die Laien war es im Vergleich zu den Fachpersonen weniger wahrscheinlich, die Urkunde als am wenigsten bedeutsam einzustufen. Die Laien schienen somit auch anderen Beweismitteln eine ähnliche oder geringere Relevanz zuzuschreiben. Für Expert:innen schien dagegen mit großer Mehrheit eine Übereinstimmung darüber vorzuliegen, dass der Urkundenbeweis die geringste Bedeutung besitzt. Im Umkehrschluss ist allerdings auch zu erwähnen, dass vergleichbare

Mechanismen auf die Laien zutreffen könnten, da diese ebenfalls eine deutliche Entscheidungstendenz innerhalb ihrer Gruppe zeigten – wenngleich deren Strategien zu einem anderen (konträren) Ergebnis führten.

Lediglich die Noviz:innen lassen sich nicht eindeutig zuordnen (Konfidenzintervall), da die Verteilung der Antworten auf gegensätzliche Meinungen innerhalb der Gruppe oder auch auf eine mögliche individuelle Unsicherheit hinweist. Betrachtet man dieses mittlere Expertise-Level als Übergang zwischen den beiden „extremen" Polen – der durch die Ausbildung geebnet wird – so könnte man annehmen, dass die besagte Ausbildung zu einer Art Veränderung oder Entwicklung im Umgang mit Beweismitteln beiträgt. Diese Entwicklung könnte sich vom Loslösen einer „laienhaften" hin zum Äußern einer fachlichen Reaktion bewegen. Diese Vermutung lässt sich in diesem querschnittlichen Design allerdings nicht überprüfen. Die Tatsache, dass die Übereinstimmung zwischen den Expert:innen größer ist als die zwischen den Noviz:innen, spricht für die Annahme von Dickert et al. (2012), dass die Kongruenz zwischen Urteilen bei größer werdenden Überschneidungen der Wissensstrukturen – also mit fortschreitender Ausbildung – steigt (s. auch Mitchell, 1989). Die Unentschiedenheit der Noviz:innen könnte auch darin begründet liegen, dass sie neben dem (einmalig präsentierten) deliktbezogenen Auszug aus dem StGB und der Grafik zu möglichen Entscheidungsoptionen keinerlei Hilfsmittel zur Bearbeitung der Fallbeschreibungen erhielten (s. 3.2).[45] Doch laut Nievelstein et al. (2010) sind solche Hilfsmittel durchaus relevant, insbesondere bei fortgeschrittenen Noviz:innen, wie sie in dieser Studie angesichts der durchschnittlichen Semesterzahl durchaus beschrieben werden können (s. 3.3). Allerdings lässt sich auch in dieser Studie keine klare richtige oder falsche Antwort auf die Frage nach dem Tatverdacht geben. Dickert et al. (2012) argumentieren sogar, dass „congruency with judgments from the German Federal Court of Justice does not necessarily equate to ‚correct' judgments" (S. 231). Es zeigten sich zwar Unterschiede durch die Expertise-Ausprägungen, ob dies aber einer „besseren" Antwort auf Seiten der Fachpersonen entspricht, lässt sich nicht klären, da weder objektive Kriterien noch eine

[45] Es lassen sich keine Hinweise ableiten, dass die Noviz:innen bei der Bearbeitung der Fallbeschreibungen weniger motiviert oder engagiert waren als die Vergleichsgruppen, und dass aus diesem Grund eine unklare Tendenz zustande kam (s. auch Ettenson, 1987).

Grundwahrheit vorliegen (s. auch Shanteau, 2000; Weiss & Shanteau, 2012). Vielmehr geht es an dieser Stelle um die Untersuchung der Prozesse, die sich auf die juristische Entscheidungsfindung auswirken (Dickert et al., 2012).

Die Expertise spielte allerdings nicht nur bei der Einschätzung des Tatverdachtes eine Rolle, sondern es wurde ein großer Effekt auf die Prozessmerkmale der Leichtigkeit, Sicherheit und Überzeugung deutlich (F1; s. 4.6.2). Dieses Ergebnis lässt sich so interpretieren, dass die Expert:innen aufgrund ihrer Praxis bereits einen gewissen Erfahrungsschatz im Umgang mit juristischen Inhalten und Fragestellungen sowie ein ausgebautes mentales Netzwerk besitzen (s. auch Herbig & Glöckner, 2009; Mitchell, 1989). Auch galten die hier behandelten Delikte als vertraut (s. 3.3). Da Expert:innen besonders ausgebildet sind und in ihrer praktischen Arbeit hohe kognitive Leistungen erbringen, ist die Wahrscheinlichkeit groß, eine starke Überzeugung für die eigenen Entscheidungen zu entwickeln (Kahneman, 2011). Noviz:innen besitzen zwar eher theoretisches als praktisches Hintergrundwissen, übersteigen aber auch damit sicherlich den Wissensstand der Naiven. Aus diesem Grund ist es nachvollziehbar, dass mit steigendem Level der Expertise auch das subjektive Ausmaß der Leichtigkeit im Entscheiden sowie die Gefühle von Sicherheit und Überzeugung zunehmen. Inwiefern diese Gefühle der Überzeugung und der Sicherheit tatsächlich eine fehlerhafte Kalibrierung zwischen Richtigkeit und Sicherheit darstellen und eine gewisse Selbstüberschätzung „provozieren“ (Chi, 2006; Cooke, 1991; Schweizer, 2005), kann für die vorliegende Stichprobe nicht eindeutig geklärt werden. Mitunter ist schlicht der Erfahrungsschatz der Expert:innen ausschlaggebend dafür, dass sie sich mit einer solchen Frage (nach dem Tatverdacht) und dem Abwägen von Beweismitteln sicherer fühlen als Personen, für die dies nicht dem Arbeitsalltag entspricht. Doch selbst wenn das Risiko besteht, dass ein Gefühl von Überzeugung in eine Selbstüberschätzung übergeht, so argumentieren Guthrie et al. (2001) folgendermaßen: „On balance, the social benefits of having confident, decisive judges likely outweigh the costs associated with an occasional erroneous decision caused by such self-assurance“ (S. 83). Der besagte Erfahrungsschatz ist nicht nur für die Prozessmerkmale relevant, sondern auch für die Einschätzung des Schweregrades des Deliktes. Schweizer (2015) argumentiert, dass sich Fachpersonen weniger von emotionalen Fallinhalten beeinflussen lassen. Die Ergebnisse dieser Studie bestätigen dies dahingehend, als dass Expert:innen den Schweregrad der

Körperverletzung geringer einstuften als die Naiven, wohingegen dies für den mutmaßlich weniger emotionalen Diebstahl erwartungsgemäß nicht festzustellen war (H4; s. 4.5.2).

Keinen Einfluss zeigte die Expertise dagegen auf die Entscheidung über den nächsten Verfahrensschritt (H3; s. 4.4.2). Auf deskriptiver Ebene wirkten Expert:innen beim Delikt „Diebstahl" aufgrund der gleichen Verteilung auf „Einstellung" und „Anklage" unentschieden; Laien und Noviz:innen tendierten mehrheitlich zur Einstellung. Für das Delikt der Körperverletzung wählten alle drei Gruppen bevorzugt die Einstellung. Dieser deskriptive Trend zeigte sich auch auf deliktübergreifender Ebene.[46] Eine mögliche Begründung dafür, dass sich die Expertise hier nicht auswirkte, ist, dass die Frage nach dem nächsten Verfahrensschritt nicht so uneindeutig zu beantworten ist wie die Frage nach dem Tatverdacht. Gibt es bei letzterer aufgrund der unsicheren Beweislage gewisse Freiheitsgrade in der Beantwortung (je nach Beweiswürdigung), scheinen diese Freiheiten beim nächsten Verfahrensschritt weniger oder nicht vorzuliegen. Die Teilnehmenden entschieden sich mehrheitlich für die Einstellung, deren Wahl auf Grundlage der Fallbeschreibung womöglich „provoziert" wurde.[47] Das Ausmaß der Unsicherheit könnte sich demnach für den nächsten Verfahrensschritt reduziert haben. Allerdings ist es auch möglich, dass das Wählen der Einstellung eine Korrektur der Tatverdacht-Antwort darstellt. So könnten Teilnehmende bei der Überlegung darüber, welcher nächste Schritt geeignet ist, mittlerweile aufkommende Zweifel berücksichtigen und sich folglich für die weniger drastische Option der Einstellung entscheiden. Die kurze Zeitspanne, die zur Entscheidung benötigt wurde, spricht allerdings gegen zeitintensives Abwägen (s. 4.1). Ebenfalls unwahrscheinlich ist, dass die Fachpersonen im Sinne einer arbeitsreduzierenden Einstellungstendenz handelten (s. auch Jahn, 2015; s. 2.1.5), da das Bearbeiten des hypothetischen Falles nicht mit nachfolgenden Aufgaben verbunden war, wie es wiederum in der Praxis üblich wäre. Vermutlich lag für die Teilnehmenden keine Verhältnismäßigkeit zwischen der Tat und einer Anklageer-

[46] Dass auch Naive sich mehrheitlich für die Einstellung aussprachen, spricht dafür, dass keine Verständnisschwierigkeiten beim Bearbeiten der Fallbeschreibungen entstanden waren, da sich nicht „inflationär" für die allgemein bekanntere Option der Anklage entschieden wurde.

[47] Eine ähnliche Begründung diente in der Pilotstudie als Ausschlusskriterium für alternative Fallbeschreibungen, deren Bearbeitung zu einheitlichen Antwortmustern führte (s. 3.6).

hebung vor, weswegen sie sich überwiegend für eine Form der Einstellung entschieden. Das Level der Expertise war außerdem nicht relevant für die Fähigkeit zur kognitiven Reflexion (s. 5.2.3), für das Kognitionsbedürfnis (s. 5.2.4) oder für die Anzahl der genutzten Beweismittel (s. 5.2.8). Auch für das Beweismaß zeigten sich keine signifikanten Unterschiede, wenngleich sich erwartungsgemäß andeutete, dass Laien von der Beweislage insgesamt überzeugter waren (s. 5.2.6).

5.2.2 Die Bedeutung des Zeitdrucks

Die Manipulation von Zeitdruck wurde als Experimentalfaktor eingebaut, um einen gewissen Praxisbezug zu erreichen (externe Validität; Guthrie et al., 2007; Savolainen, 2006; s. 2.2.3.1; 2.3.3.2). Des Weiteren wurde der Faktor aufgrund seiner Nähe zu den Dual-Prozess-Annahmen ausgewählt, weil ebenjene Annahmen eine Begründung für die Beeinflussbarkeit juristischer Entscheidungen darstellten (Guthrie et al., 2007; s. 2.3.2). Entgegen der Erwartungen spielte die Manipulation von Zeitdruck in den meisten Analysen keine signifikante Rolle auf die untersuchten Variablen. Die Teilnehmenden zeigten ähnliche Reaktionen hinsichtlich des Vorliegen des Tatverdachtes (H1 und F3; s. 4.2.2; 4.8.2), des nächsten Verfahrensschrittes (H3; s. 4.4.2), der erfassten Prozessmerkmale (Leichtigkeit, Sicherheit, Überzeugung; F1; s. 4.6.2), der Einschätzung des Beweismaßes (H2; s. 4.3.2), des Schweregrades des Delikts (H4; s. 4.5.2) und der Anzahl der genutzten Beweismittel (F4 und F6; s. 4.9.2; 4.11.2), unabhängig davon, ob die Aufforderung zur Geschwindigkeit oder zur Genauigkeit erfolgt war. Lediglich in der Gruppe der Expert:innen ergaben sich Unterschiede auf deskriptiver Ebene: Diejenigen ohne Zeitdruck wählten vermehrt die Einstellung, wohingegen unter Zeitdruck die Wahl des Verfahrensschritts nahezu gleich verteilt war (H3; 4.4.1). Dieses deskriptive Ergebnis steht im Gegensatz zur subjektiven Einschätzung dieser Fachpersonen, dass Zeitdruck eher einen geringen Einfluss auf juristische Entscheidungen hat (s. 3.3). Unabhängig von der Expertise drehte sich die Einschätzung des Schweregrades des Deliktes je nach Zeitdruckmanipulation um. Diejenigen mit Zeitdruck erachteten die Körperverletzung als schwerer, wohingegen diejenigen ohne Zeitdruck die Schwere des Diebstahls höher ansetzten. Diese Unterschiede waren aber nur deskriptiv und in sehr kleinem Ausmaß zu erkennen (H4; s. 4.5.1).

Inferenzstatistisch wirkte sich der Faktor „Zeitdruck“ nur auf die Beweismittel mit der niedrigsten beziehungsweise höchsten Relevanz aus (F5; s. 4.10.3). Das Fehlen von zeitlichem Stress führte dazu, dass die Wahrscheinlichkeit sank, die Zeugenaussage als unbedeutsam einzuschätzen. Gleichzeitig war dies ein signifikanter Prädiktor für die höchste Relevanz. Genauer gesagt stieg die Wahrscheinlichkeit, die Zeugenaussage sowie die Aussage des Beschuldigten als relevant einzustufen, wenn kein Zeitdruck induziert worden war. Somit spielte (fehlender) Zeitdruck in erster Linie im Umgang mit den Beweismitteln (insbesondere hinsichtlich der Zeugenaussage) eine Rolle, wirkte sich aber nicht signifikant auf die eigentlichen Entscheidungen aus.

Auf deskriptiver Ebene gab es keinen Hinweis darauf, dass sich die Teilnehmenden in Abhängigkeit vom Zeitdruck mehr für den Verfahrensschritt der Anklage aussprachen (H3; s. 4.4.2). Laut Liu et al. (2019) kann aber das Erleben von Zeitdruck zu verstärktem Strafverhalten führen. Deren Ergebnis wird somit nicht durch die Befunde bestätigt, vermutlich, weil das Erleben von Zeitdruck nicht stark genug war, um sich auf das Strafverhalten auszuwirken. Im Rahmen des Manipulationschecks wurde bereits deutlich, dass nur eine der drei Zeitmessungen zwischen Teilnehmenden ohne Aufforderung zur Schnelligkeit und denjenigen mit induziertem Zeitdruck signifikante Unterschiede aufwies (s. 4.1). Laut Rastegary und Landy (1993) ist das Erleben von Zeitdruck sehr individuell. Zudem sei das Gefühl von Dringlichkeit entscheidend. Die Studienteilnehmenden erhielten zwar die Aufforderung zur Geschwindigkeit oder zur Genauigkeit, allerdings wurden diese nicht dazu befragt, ob sie sich wirklich gestresst fühlten und wie sie das Messen der Zeit wahrnahmen (s. auch Alison et al., 2013). Die Effektivität der Beeinflussung wurde in dieser Studie lediglich objektiv anhand der Zeitmessungen ermittelt. Es besteht die Möglichkeit, dass die Manipulation nicht ausreichend war, um Teilnehmende tatsächlich in die intendierte Drucksituation zu bringen (s. auch Zakay, 1993). Die Teilnehmenden kamen daher vermutlich nicht in die Not, sich an das Erleben von Zeitdruck anpassen zu müssen (s. auch Rice & Trafimow, 2012; Rieskamp & Hoffrage, 2008) – zumal eine lange Bearbeitungszeit mit keinerlei Konsequenzen einherging. Die Güte der Entscheidung und die Geschwindigkeit standen somit nicht in Konkurrenz zueinander (*speed-accuracy-trade-off*; E. J. Johnson et al., 1993;

Wickelgren, 1977). Schnelles Denken bedeutet zudem nicht, dass ein Prozess intuitiv abläuft (Croskerry et al., 2014), sondern es kann sich auch um einen Hinweis dafür handeln, dass insgesamt kohärente, zueinander passende Informationen vorliegen (Glöckner & Betsch, 2012). Demzufolge waren die Inhalte der Fallbeschreibung möglicherweise nicht so uneindeutig wie ursprünglich angenommen, sodass sich insbesondere die Messung einer kurzen Zeit (zum Entscheiden oder zum Lesen) nicht zwingend auf die uneindeutige Beweislage zurückführen ließe. Dagegen spricht aber die deskriptive Tendenz zur Mitte bei der Einschätzung des Beweismaßes, welche die intendierte fehlende Eindeutigkeit der Beweislage durchaus unterstützt (s. 5.2.6). Daher sind die Interpretationen der Ergebnisse, die sich auf eine unzureichende Zeitdruckmanipulation beziehen, plausibler.

Zeitdruck kann dazu führen, dass Menschen auf heuristische, intuitive Entscheidungsstrategien zurückgreifen, die wiederum dann bestärkt werden, wenn es sich um bekannte, routinierte Umgebungen oder kognitiv fordernde Aufgaben handelt (Glöckner & Ebert, 2011; Rice & Trafimow, 2012; Schweizer, 2009; s. 2.2.3.1; 2.3.3.2; 2.3.3.3). Es wurde davon ausgegangen, dass die eingesetzten Instruktionen ausreichen, um duale Prozesse zu aktivieren. Da die hier untersuchten (juristischen) Materialien und Fragestellungen für die Naiven (und mitunter auch für noch nicht fortgeschrittene Noviz:innen) mit sehr großer Wahrscheinlichkeit neu und ungewohnt waren, ist es fraglich, inwiefern diese Personen überhaupt auf intuitives Verhalten zurückgreifen konnten. Prinzipiell wäre der Einsatz von Typ-1-Prozessen und von wissensbasierter Intuition somit aufgrund der Natur der untersuchten Aufgaben in erster Linie für die Expert:innen möglich (s. auch J. K. Phillips et al., 2004; Stanovich & West, 2000). Für die Naiven (und gegebenenfalls für die Noviz:innen) würde intuitives Verhalten eher den Zweck erfüllen, die kognitiv anspruchsvolle Aufgabe in ihrer Komplexität zu reduzieren. Die Tatsache, dass es sich um eine reflektierte, zum Denken motivierte Stichprobe handelte (s. 5.2.3; 5.2.4), lässt aber die Vermutung zu, dass dieser Bedarf zur Reduktion von Komplexität nicht übermäßig ausgeprägt war. Möglicherweise war die Motivation zum schnellen Arbeiten für alle Teilnehmenden unter Zeitdruck zunächst gegeben, wurde aber sukzessive weniger beachtet, umso mehr sich mit der (ungewohnten) Aufgabe beschäftigt wurde. Dies könnte auch die Gruppen mit juristischer Vorbildung betreffen, für die

eine vor der Präsentation der Vignette gestellte Aufforderung zur Schnelligkeit womöglich nicht ausreichte, um während der *gesamten* Bearbeitung bedacht zu werden. Auch die Wiederholung der jeweiligen Instruktion am Ende der Fallbeschreibung wäre demnach nicht ausreichend gewesen, zumal sie je nach Bildschirmgröße des genutzten Endgeräts nicht ununterbrochen sichtbar gewesen sein könnte.[48] Durch die Novität, die die Studie insbesondere für die Naiven besaß, reichte die bloße Instruktion zur Schnelligkeit mitunter nicht aus, um Zeitdruck zu induzieren, aufrechtzuerhalten und damit verbundene Prozesse zu aktivieren. Es ist außerdem möglich, dass die Manipulation nicht die Nutzung der Typ-1/Typ-2-Prozesse aktivierte, sondern andere Mechanismen (z. B. die Motivation zum Denken; Smith & Levin, 1996), und dass sowohl heuristische als auch systematische Prozesse gleichzeitig aktiv waren (Henderson & Levett, 2020; s. auch Diederich & Trueblood, 2018).

Fehlte Zeitdruck, so war es signifikant weniger wahrscheinlich, die Zeugenaussage als unbedeutsam einzuschätzen, und so war es wahrscheinlicher, die Zeugenaussage sowie die Aussage des Beschuldigten als relevant einzustufen (F5; s. 4.10.3). Dass sich die Aufforderung zur Genauigkeit nicht auf die eigentlichen Entscheidungen, sondern auf die Handhabung der Beweismittel zu konzentrieren scheint, passt zu der Annahme, dass zeitlicher Stress die Suche nach Informationen beeinflussen kann (Rice & Trafimow, 2012; Rieskamp & Hoffrage, 2008; s. 2.2.3.1). Laut Oh et al. (2016) kann Zeitdruck dazu führen, sich nur auf bestimmte Informationen zu fokussieren und als irrelevant eingestufte Aspekte zu ignorieren. Allerdings war in dieser Studie das *Fehlen* des Zeitdrucks für die Einschätzung der Relevanz bestimmter Beweismitteln entscheidend. Auch hier besteht – wie bereits diskutiert wurde – die Möglichkeit, dass die Manipulation nicht dahingehend erfolgreich war, spürbaren Zeitstress zu induzieren und dass in der Gruppe *Genauigkeit* andere Prozesse aktiviert wurden (Smith & Levin, 1996). Somit war mitunter nicht die Aufforderung zur Genauigkeit, sondern eine anders geartete Motivation (z. B. zum Denken) ausschlaggebend für die Effekte. Es bleibt allerdings festzuhalten, dass die

[48] Knapp die Hälfte der Teilnehmenden nahm über das Smartphone an der Umfrage teil (s. 3.3), was für die Vermutung spricht, dass die Bildschirme zu klein waren, um sowohl die Fallbeschreibung als auch die Instruktion gleichzeitig anzuzeigen.

überwiegende Nichtsignifikanz des Faktors „Zeitdruck“ in dieser Studie nicht zwingend in einer fehlerhaften Annahme von Dual-Prozess-Modellen, sondern vielmehr in dessen wenig erfolgreicher Manipulation begründet liegt (s. 5.3).

5.2.3 Die Bedeutung der kognitiven Reflexion

Die kognitive Reflexion galt im Rahmen dieser Studie als personenbedingter Einflussfaktor auf die Entscheidungsfindung, der aufgrund seiner Nähe zu den Dual-Prozess-Annahmen ausgewählt wurde (s. 2.2.3.3). Dessen signifikanter Einfluss bezog sich allerdings nicht auf die eigentliche Entscheidung über den Tatverdacht (F3; s. 4.8.2), sondern auf die Anzahl der genutzten Beweismittel (F4; s. 4.9.2), wenngleich in kleinem Ausmaß. Menschen, die in der Lage sind, erste intuitive Impulse zu unterdrücken, sind vermutlich generell dazu bereit, sich mit einem Angebot an Informationen auseinanderzusetzen, bevor es zu einer Schlussfolgerung oder einer Reaktion kommt – im Vergleich zu intuitiv denkenden Menschen. Folglich ist es nachvollziehbar, dass diejenigen, die hohe Reflexionsfähigkeiten besaßen, auch das Bedenken einer höheren Anzahl an Beweismitteln leisten konnten.

Bezogen auf die gesamte Leistung im CRT zeigte sich (nur) auf deskriptiver Ebene ein Effekt von Expertise, denn Fachpersonen schnitten besser ab als Noviz:innen, die wiederum eine bessere Leistung erzielten als Naive (F2; s. 4.7.1). Die erzielten Mittelwerte (insbesondere der Fachpersonen) überstiegen die Leistungen der richterlichen Stichprobe von Guthrie et al. (2007). Es ist davon auszugehen, dass es sich auch bei den Naiven aufgrund der Art der Rekrutierung um eine überwiegend akademische Teilstichprobe handelte, unabhängig von der Ausprägung der juristischen Expertise (s. 3.3). Dies wäre eine Erklärung für die insgesamt hohen Werte aller Expertise-Gruppen im CRT, da zwar keine vollständige, aber eine gewisse Überschneidung mit Intelligenz besteht (Stanovich & West, 2008, 2014).[49] Somit handelte es sich laut dem CRT bei den Teilnehmenden um überwiegend reflektierte Entscheider:innen. Auch wenn die Items in dieser Studie randomisiert präsentiert wurden, ergaben sich deskriptive Unterschiede in deren Lösbarkeit. Die Seerosen-Aufgabe fiel den Teilnehmenden am leichtesten, gefolgt von den Aufgaben zur Anzahl von Maschinen und zum Schläger-Ball-Preis. Dies entspricht den Befunden

[49] Im Abschnitt 5.3 werden im Zusammenhang mit dem CRT gewisse Limitationen aufgeführt, die ebenfalls eine Begründung für hohe Werte darstellen.

von Guthrie et al. (2007). Sie interpretierten die sich steigernde Leistung zum Ende hin mit dem reduzierten Verlass auf Intuition (Lerneffekt). Dass in dieser Studie trotz einer Randomisierung ein ähnliches Gefälle beobachtet werden konnte (wenn auch bei durchgehend besserer Leistung als in der Studie von Guthrie et al., 2007), spricht weniger für einen Lerneffekt, sondern vielmehr für eine tatsächliche Diskrepanz in den Schwierigkeitsgraden der einzelnen Aufgaben (Brañas-Garza et al., 2015). Allerdings lässt sich nicht ermitteln, ob die Teilnehmenden bereits zu Beginn die korrekte Lösung wussten oder diese durch Nachdenken fanden (Bago & Neys, 2019; Raoelison et al., 2020).

Es scheint für das Lösen der Aufgaben nicht nur das Unterdrücken einer ersten spontanen Reaktion eine wichtige Rolle zu spielen, sondern auch das Vorhandensein von Wissen. Fehlt dieses Wissen, kann es zu Fehlern in der Schlussfolgerung kommen (Szaszi et al., 2017). Insbesondere die Tatsache, dass die Aufgaben des CRT auf numerischen Inhalten basieren, spricht dafür, dass fehlende Kenntnisse oder Fähigkeiten in dem Bereich auch zu einer schwachen Leistung im CRT beitragen könnten (Stanovich, 2018; Toplak et al., 2014). In dieser Studie schnitten Männer auf deskriptiver Ebene im Mittel besser ab als Frauen oder nichtbinäre Personen (s. 4.7.1). Die bessere Leistung von Männern zeigte sich nicht nur mit Blick auf die mittlere Gesamtpunktzahl, sondern es handelte sich auch um die Gruppe, die am häufigsten alle Items korrekt lösen konnte und die in geringster Anzahl keines der drei Items richtig beantwortete. Diese Befunde ähneln denen von Frederick (2005) und passen in die Diskussion, ob und inwiefern sich die Leistung im CRT mit mathematischen Fähigkeiten begründen lässt (Brañas-Garza et al., 2015; Juanchich et al., 2020). Dass aber auch Frauen oder nichtbinäre Personen einen recht hohen Mittelwert erzielten, spricht gegen die Annahme, dass gute Leistung übermäßig durch mutmaßlich mit dem Geschlecht einhergehende numerische Fähigkeiten erzielt werden kann (s. auch Campitelli & Gerrans, 2014; Liberali et al., 2012). Inwiefern sich diese Fähigkeit zum Unterdrücken intuitiver Reaktionen auch auf den (juristischen) Alltag übertragen lässt, bleibt unklar (Juanchich et al., 2016). Eine Nichtsignifikanz beziehungsweise ein nur kleiner Effekt von kognitiver Reflexion lässt sich damit begründen, dass es sich hinsichtlich dieser Fähigkeit um eine recht homogene Gruppe handelte, und dass Leistungen im CRT nur in sehr

kleinem Maße an einer Varianzaufklärung in lebensechten Entscheidungen mitwirken (Juanchich et al., 2016).

5.2.4 Die Bedeutung des Need for Cognition

Auch dieser personenbedingte Einflussfaktor wurde aufgrund der inhaltlichen Nähe zu Dual-Prozess-Annahmen und der Bereitschaft zum (reflektieren) Denken ausgewählt (s. 2.2.3.4). Auffällig war die sehr schwache, nicht signifikante Interkorrelation des CRT mit der *Need-for-Cognition*-Kurzskala (r = .046; s. 3.4.3). Die Forschungsergebnisse von Frederick (2005) ergaben dagegen einen Wert von r = .22.[50] Dieser Wert kam allerdings nicht mittels der hier genutzten Kurzskala zustande. Daher besteht die Möglichkeit, dass ein solch kleiner, nicht signifikanter Zusammenhang im Aufbau der Kurzskala begründet ist. Pechtl (2009) weist darauf hin, dass das Kognitionsbedürfnis, wie es in den ursprünglichen Skalen erhoben wird, bestimmte Subdimensionen erfasst. Daher könnte es sich bei den vier Items der Kurzskala um solche Subdimensionen handeln, die keinen starken Zusammenhang mit der Fähigkeit zur kognitiven Reflexion aufweisen. Auch wenn kein Zusammenhang zwischen NFC und sozialer Erwünschtheit vorzuliegen scheint (Cacioppo & Petty, 1982), so ist es dennoch möglich, dass die Teilnehmenden ihre Motivation zum Denken „falsch“ einschätzten. In einer Studie von Pennycook et al. (2017) entsprach die Leistung im analytischen Denken nicht der Leistung, die aufgrund der Angabe zum NFC zu erwarten gewesen wäre. Dass dies auch die vorliegende schwache Interkorrelation begründet, ist allerdings unwahrscheinlich, da die Leistung im analytischen Denken (hier auch CRT) ebenso wie die Ausprägung des NFC im Mittel eher als hoch eingeschätzt werden können (s. 4.7.1).

Insgesamt waren die Teilnehmenden nach eigenen Angaben zum Denken motiviert. Das Level der Expertise hatte einen kleinen, aber signifikanten Effekt auf die Ausprägung des Kognitionsbedürfnisses (F2; s. 4.7.2). Anschließende paarweise Vergleiche gaben aber keine zusätzlichen Auskünfte über signifikante Unterschiede. Laut deskriptiver Ebene besaßen die Noviz:innen die stärkste Ausprägung, wobei das Ausmaß für Laien und Expert:innen im Mittel sehr vergleichbar und nur etwas schwächer war als das der Noviz:innen (s. 4.7.1). Auch wenn die Expert:innen (und

[50] Auch diese Tatsache stellt im Nachhinein eine Begründung dafür dar, die kognitive Reflexion und das Kognitionsbedürfnis nicht gemeinsam anhand einer MANOVA zu analysieren (s. 3.8.2).

sicherlich auch die Naiven) in ihrem Berufsalltag mit kognitiven Herausforderungen konfrontiert sind, lässt sich doch vermuten, dass die Teilstichprobe derjenigen, die sich zum Zeitpunkt der Teilnahme im Studium oder im Referendariat befunden hat, eine noch ausgeprägtere Bereitschaft zum Denken aufweist. Andernfalls würde sich sicherlich nicht für ein solch intensives Studium entschieden (s. auch Glöckner et al., 2013). Im Zusammenhang mit kognitiver Reflexion wurde bereits argumentiert, dass hohe Werte mit der Zusammensetzung der Stichprobe begründet werden können (s. 5.2.3). Da auch das Kognitionsbedürfnis eine positive Korrelation mit Intelligenz aufweist, lässt sich die gruppenübergreifend hohe Denkmotivation ebenfalls dadurch erklären (Fleischhauer et al., 2010; Hill et al., 2013).

Das Kognitionsbedürfnis erwies sich nicht als signifikanter Prädiktor für die Entscheidung über das Vorliegen des Tatverdachtes oder die Anzahl der genutzten Beweismittel (F3 und F4; s. 4.8.2; 4.9.2). In der Studie von Henderson und Levett (2020) waren Menschen mit hohem Kognitionsbedürfnis nicht zwingend besser in der Differenzierung unterschiedlicher Beweislagen als diejenigen mit niedrigem Bedürfnis. Hieraus lässt sich schlussfolgern, dass diejenigen mit höheren NFC-Werten in dieser Studie nicht zwangsläufig ein Vorteil in der Bearbeitung der Vignetten, die Fälle mit nicht eindeutiger Beweislage präsentierten, hatten.[51] Dies könnte die Nichtsignifikanz des Prädiktors begründen. Außerdem steigt die Wahrscheinlichkeit für das Auftreten von NFC-Effekten, wenn sich die persönliche Relevanz für die zu bearbeitende Aufgabe als niedrig bis mittel einstufen lässt (Cacioppo et al., 1996; Petty et al., 2009). Da die behandelte juristische Thematik für die Noviz:innen und Expert:innen durchaus in deren Interessensgebiet fällt, war die persönliche Relevanz möglicherweise zu hoch, um das Auftreten der Effekte zu induzieren, da auch diejenigen mit eigentlich niedrigen Ausprägungen zum Denken motiviert wurden. Inwiefern sich die Neigung zum Denken (bei den Fachpersonen) auch auf den beruflichen Kontext übertragen lässt, ist nicht zu klären. Laut Bieneck (2006) ist eine Trennung zwischen einem allgemeinen und einem bereichsspezifischen Kognitionsbedürfnis zwar möglich, bringt aber nur einen geringen Erkennt-

[51] Allerdings wurde nicht untersucht, inwiefern sich die verschiedenen Ausprägungen der Denkmotivation auf die Einschätzung des Beweismaßes auswirkten.

nisgewinn. Mit Blick auf die hier eingesetzten Vignetten spielte das Kognitionsbedürfnis in der Beweiswürdigung nur eine geringe, nicht signifikante Rolle für die Teilnehmenden.

5.2.5 Die Bedeutung des Delikttyps

Aus ursprünglich vier Vignetten wurden im Rahmen einer Pilotstudie zwei ausgesucht, die sich für die Zwecke der Studie als geeignet erwiesen haben (s. 3.6). Beide wurden in dieser Hauptstudie eingesetzt, um sie an einer geeigneten Stichprobe zu testen. Der Delikttyp wurde in den Hypothesen und Fragestellungen als unabhängige Variable oder Prädiktor berücksichtigt, um sicherzustellen, dass eventuelle Effekte in Abhängigkeit zu eigentlich relevanten Variablen auftreten und nicht durch die Fallbeschreibung entstehen (s. 2.5).

Der Delikttyp hatte keinen Einfluss auf die Wahl des nächsten Verfahrensschrittes (H3; s. 4.4.2), auf die Einschätzung des Beweismaßes (H2; s. 4.3.2), auf das Empfinden von Leichtigkeit, Sicherheit und Überzeugung (F1; s. 4.6.2) und auf die Anzahl der genutzten Beweismittel (F4; s. 4.9.2). Allerdings war der Einfluss des Delikttyps signifikant für die Entscheidung über das Vorliegen des Tatverdachtes, da im Fall der Körperverletzung die Wahrscheinlichkeit für das Bejahen stieg (kleine Effektstärke; H1; s. 4.2.2). Das Überschreiten des Wertes von 1 im Konfidenzintervall war nur marginal und ändert nicht die Interpretation. Zudem war der Delikttyp ein signifikanter Prädiktor des Beweismittels mit höchster Relevanz (für das Beweismittel mit niedrigster Relevanz wurde das Signifikanzniveau knapp verpasst; F5; s. 4.10.3). Mit Blick auf einzelne Parameter zeigte sich, dass beim Delikt des Diebstahls (im Vergleich zur Körperverletzung) die Wahrscheinlichkeit sank, entweder den Urkundenbeweis, die Aussagen des Zeugen oder des Beschuldigten zu wählen, wohingegen die Wahrscheinlichkeit stieg, entweder die Urkunde, die Aussagen des Zeugen und des Beschuldigten als relevant einzustufen. Außerdem war der Delikttyp für die Einschätzung der jeweiligen Schwere relevant, da sich erwartungsgemäße Unterschiede zwischen Laien und Fachpersonen hinsichtlich der Körperverletzung zeigten, aber nicht hinsichtlich des Diebstahls (H4; s. 4.5.2). Dadurch wird deutlich, dass das Delikt keinesfalls unberücksichtigt bleiben sollte, da sich der in einer Vignette behandelte Inhalt nachweislich auf abhängige Variablen aus-

wirken kann. McKay et al. (2014) argumentieren, dass Verbrechen gegen die körperliche Unversehrtheit die Moral der Menschen anspricht. Daher ist es nachvollziehbar, dass sich die Teilnehmenden hier vermehrt für eine Bejahung des Tatverdachtes aussprachen als bei einem Eigentumsdelikt. Auch wenn die Unschuldsvermutung eigentlich unabhängig von der Natur des Deliktes zum Tragen kommen sollte (s. 2.3.3.4), könnte es sein, dass es für die Teilnehmenden in einem Fall von Körperverletzung „schlimmer“ war, eine möglicherweise schuldige Person laufen zu lassen, weswegen von besagter Unschuldsvermutung unbewusst Abstand genommen wurde. Auf deskriptiver Ebene traf es zumindest für die Laien durchaus zu, dass die Körperverletzung im Mittel als schwerer galt als der Diebstahl. Laut Bieneck (2006) korreliert das Strafmaß mit der wahrgenommenen Schwere der Tat. Da ein Berücksichtigen der Unschuldsvermutung für Naive ungewohnt ist (und sie in dieser Studie nicht explizit darauf hingewiesen wurden), könnte sich dies im vermehrten Bejahen des Tatverdachts – was dem „höheren“ Strafmaß entspricht – im Zusammenhang mit der als schwerer wahrgenommenen Tat widerspiegeln.

5.2.6 Die Bedeutung des Beweismaßes

Das Strafverfahren dient dazu, eine Straftat rekonstruierbar und begreifbar zu machen (Schroeder & Verrel, 2017). Das Beweismaß beschreibt den Grad der Überzeugung, einen Sachverhalt als wahr anzuerkennen, der im Rahmen der freien Beweiswürdigung nach § 261 StPO erreicht wird (Schweizer, 2015). Eine absolute Sicherheit über die Wahrheit kann dabei nie erreicht werden, zumal die freie Beweiswürdigung nur subjektive, aber keine objektive Gewissheit mit sich bringt (Ackermann et al., 2022; Holländer, 2019; Schweizer, 2015). “The ground truth of cases is never known” (Spellman, 2007, S. 8). Die Teilnehmenden mussten angeben, wie stark sie von der Beweislage in der jeweiligen Fallbeschreibung überzeugt waren (H2; s. 4.3.2). Weder das Delikt noch die Zeitdruckmanipulation wirkte sich signifikant auf die Einschätzungen aus. Auch wenn das Signifikanzniveau für die Expertise nicht erreicht wurde, so deutete sich an, dass Laien ein höheres Beweismaß als die Vergleichsgruppen (insbesondere Expert:innen) angaben. Für die Analyse der Relevanz der Beweismittel zeigte sich, dass die Beweiswürdigung der Expertise-Gruppen – zumindest mit Blick auf nicht relevante Beweismittel – womöglich nicht gleich verlief, da unterschiedliche Gewichtungen vorgenommen wurden

(F5; s. 4.10.3). Somit passt das Ergebnis, dass sich die Gruppen (zumindest deskriptiv) auch im daraus resultierenden Überzeugungsgrad unterschieden: Die ungleiche Würdigung der Beweise könnte als eine Erklärung für ungleiche Ausprägungen des Überzeugungsgrades dienen.

Betrachtet man die Mittelwerte, so lässt sich gruppenübergreifend eine Tendenz zur Mitte ausmachen, die aufgrund der genutzten Skalierung bei $M = 49.5$ lag (s. 4.3.1). Bei der Konstruktion der Vignetten wurde beabsichtigt, eine nicht eindeutige Beweislage darzustellen (s. 3.4.1). Die Tatsache, dass sich die Teilnehmenden im Mittel weder stark in die Richtung der fehlenden Überzeugung (0%) noch in die Richtung der vollen Überzeugung (99%) bewegten, spricht dafür, dass diese Absicht erfolgreich war. Die Tendenz zur Mitte bedeutet allerdings auch, dass das für das Strafrecht notwendige Beweismaß, das bei 90% beginnt, nicht erreicht wurde. Mit einem Wert nahe der 50% entsprach das Beweismaß lediglich der überwiegenden, aber nicht der vollen Überzeugung (s. auch Schweizer, 2016; s. 2.1.6.1).

Betrachtet man nicht nur die Mittelwerte, sondern auch die Standardabweichungen, ist insbesondere für die Expert:innen erkennbar, dass deren Einschätzungen eine recht große Spanne umfassten (s. 4.3.1). Dies könnte derart interpretiert werden, dass die Frage nach dem Beweismaß missverständlich formuliert war. Es könnte allerdings auch sein, dass die Teilnehmenden tatsächlich in ungleichem Maß von der Beweislage überzeugt waren. Ebenso ist es möglich, dass die Teilnehmenden sich in ihrem Verständnis des Beweismaßes unterschieden und somit interindividuelle Maßstäbe angesetzt wurden – unabhängig davon, welches Maß gesetzlich gefordert war. Laut Lavie et al. (2020) sind Noviz:innen und Fachpersonen zwar besser als Naive darin, das juristische Beweismaß mit numerischen Werten zu verbinden, aber auch die Fachpersonen setzen Werte zu hoch an. Schweizer (2016) argumentiert, dass das tatsächlich gelebte und das theoretisch erforderliche Beweismaß nicht immer übereinstimmen. In seiner Studie lag der statistisch ermittelte Überzeugungsgrad von Richter:innen unterhalb der geforderten Grenze. In dieser Studie gab ebenfalls nur ein kleiner Teil der Stichprobe ein Beweismaß von ≥ 90% an. Angesichts der Anzahl derjenigen, die den Tatverdacht bejahten (s. 4.2.1), liegt die Vermutung nahe, dass auch die Fachpersonen vom eigentlich geforderten Überzeugungsgrad abgewichen sind. Womöglich setzen die Expert:innen für das Bestimmen des hinreichenden Tatverdachtes generell einen anderen Maßstab an als bei

der Schuldfrage. Auch Schmittat et al. (2022) argumentieren, dass eine Anklageerhebung nicht mit einem Urteilsspruch gleichzusetzen ist und dass in beiden Prozessen unterschiedliche Strategien oder Kriterien eingesetzt werden. In ihren Studien war die Anzahl der Anklagen geringer als die Anzahl der Urteile. Rund die Hälfte der Fachpersonen der vorliegenden Studie gab das Zivilrecht als derzeitiges Fachgebiet an (45%; s. 3.3). Somit wäre es denkbar, dass die Zivilrechtler:innen gemäß „ihres" Beweismaßstabes und nicht aus strafrechtlicher Perspektive handelten. Inwiefern die Noviz:innen, die im Mittel als fortgeschritten beschrieben werden können (s. 3.3), sich an tatsächlichen Beweismaßen orientiert haben, ist nicht zu erkennen. Aufgrund des fortgeschrittenen Studiums ist es wahrscheinlich, dass die theoretische Bedeutung der Überzeugungsgrade bekannt ist. Auch Laien besitzen kein einheitliches Verständnis für das Beweismaß (Baucum et al., 2018; Daftary-Kapur et al., 2010; Dhami et al., 2015; Mueller-Johnson et al., 2018; D. Simon, 2012; Wright & Hall, 2007). Da den Teilnehmenden in dieser Studie keine näheren Ausführungen zur Interpretation des Überzeugungsgrades vermittelt wurden, lässt sich nicht feststellen, ob sich das Entscheidungsverhalten geändert hätte, wenn quantifizierbare oder juristische Definitionen eingesetzt worden wären (s. auch Kagehiro & Stanton, 1985). Es wurde nicht berechnet, ab welchem Grad der Überzeugung das Bejahen des Tatverdachtes wahrscheinlicher wurde („Kipppunkt"; s. auch Wright & Hall, 2007; s. 5.3). Ein solcher Wert würde zusätzlich eine natürliche Entscheidungsgrenze definieren (s. auch Schweizer, 2016). Dass die Laien eher den Tatverdacht bejahten und auch deren Beweismaß (deskriptiv) höher war, spricht für das allgemeine Verständnis, dass eine höhere, wenn auch nicht den Wert von 90% übersteigende Überzeugung mit dem Bejahen des Tatverdachtes einhergeht. Dies spricht somit auch dafür, dass legale Standards ohne die Vorgabe von Wahrscheinlichkeiten angewendet werden können (Glöckner & Engel, 2013), wenngleich in dieser Studie gewisse Einschränkungen in deren exakter Umsetzung erkennbar wurden.

5.2.7 Die Bedeutung des Schweregrades des Delikts

Für die Einschätzung des Schweregrades ergab sich eine signifikante Interaktion der Expertise mit dem Delikttyp (H4; s. 4.5.2). Für die Laien war, im Vergleich zu den Fachpersonen, das Delikt der Körperverletzung als schwerer einzustufen. Ent-

gegen der Erwartungen ähnelten die Noviz:innen in ihrem deskriptiven Antwortverhalten eher den Naiven. Der Zeitdruck spielte keine bedeutsame Rolle. Der Unterschied zwischen Laien und Fachpersonen lässt sich durch den Erfahrungsschatz besagter Fachpersonen erklären, da diese mit einer größeren Anzahl an sowie einer breiteren Diversität von Delikten in Berührung kommen und somit mehr Vergleichswerte besitzen. Derartige Erfahrungen können dazu führen, dass fehlende Elemente einer Fallbeschreibung ergänzt werden, um eine erzählte Geschichte kohärent werden zu lassen (Pennington & Hastie, 1986, 1991; s. 2.1.6.2). Eventuell fehlten auch in den genutzten Fallbeschreibungen gewisse Inhalte, die die Expert:innen dank ihres Wissens ergänzt haben. Dass Noviz:innen auf deskriptiver Ebene eher den Naiven ähnelten, könnte andeuten, dass kein theoretischer, sondern ein praktischer Erfahrungsschatz notwendig ist, um ausreichende Eindrücke über Schweregrade sammeln zu können (s. auch Bieneck, 2006). Da die Wahrscheinlichkeit, den Tatverdacht zu bejahen, für das Delikt der Körperverletzung höher war (H1; s. 4.2.2), liegt die Vermutung nahe, dass sich auch der damit verbundene Schweregrad auf diese Entscheidung auswirkte, wenngleich dies nicht inferenzstatistisch überprüft wurde.

In der Studie von Bieneck (2006) lag eine positive Korrelation für das Strafmaß und den Schweregrad vor. Laut Lundberg (2016) ist mit steigender Schwere die Wahrscheinlichkeit für einen Schuldspruch niedriger. Demnach gelte es bei schweren Taten einen höheren Grad der Überzeugung zu erreichen, bevor ein Schuldspruch gesprochen werden kann. Das Beweismaß passe sich an die Schwere der Tat an. Hier bestand ein kleiner, positiver und signifikanter Zusammenhang (r = .126, p = .043) zwischen dem Schweregrad und dem Beweismaß (s. 4.5.2): Je höher die Schwere des Delikts, desto höher der Grad der Überzeugung. Da nicht ermittelt wurde, ab welchem Beweismaß die Wahrscheinlichkeit für das Bejahen des Tatverdachtes („Schuldspruch“) stieg (s. 5.2.6), lassen sich die Ergebnisse von Lundberg (2016) nicht direkt replizieren. Zudem lässt der positive Zusammenhang keine Rückschlüsse auf eine Kausalität zu. Daher bleibt es offen, ob ein als schwer eingestuftes Delikt ein hohes Beweismaß bewirkt oder ob ein hoher Grad der Überzeugung dazu führt, dass ein Delikt als schwer angesehen wird – was letztlich beides mit dem Bejahen des Tatverdachtes einhergehen könnte. Unabhängig davon,

was letztlich zutrifft, stützen die Ergebnisse die Implikation, den Delikttyp und auch dessen Schweregrad in zukünftiger Forschung zu berücksichtigen (s. 5.4).

5.2.8 Die Bedeutung der Beweismittel

In dieser Studie wurde zwischen den Beweiskategorien *Augenschein*, *Urkunde*, *Zeug:innenaussage* sowie *Einlassung der beschuldigten Person* unterschieden (s. 2.1.6). Nicht nur die Kategorien (F5 und F6; s. 4.10.3; 4.11.2), sondern auch die Anzahl der Beweismittel (F4; s. 4.9.2), die in die Entscheidungsfindung einbezogen wurden, waren Gegenstand der Untersuchungen.[52] Deskriptiv stellte sich heraus, dass die Teilnehmenden letztlich fast alle Informationselemente berücksichtigten und im Mittel nur eines der Beweismittel außen vor ließen (F4; s. 4.9.1). Dies bestätigt das Ergebnis, dass es sich um eine eher reflektierte Stichprobe mit hoher Motivation zum Denken handelt (s. 5.2.3; 5.2.4) – zumal eine höhere Fähigkeit zur kognitiven Reflexion eine größere Anzahl von genutzten Beweismitteln signifikant vorhersagte. Ungeachtet davon, welches Beweismittel letztlich ausschlaggebend für eine Entscheidungsrichtung war, wurde deutlich, dass ein Großteil der dargebotenen Informationen genutzt und sich nicht auf einige wenige Elemente fixiert wurde. Dies widerspricht allerdings einer anfänglichen Annahme, dass sich aufgrund des Zeitdrucks im Sinne eines *speed-accuracy-trade-offs* noch stärker für das Auslassen von Informationen entschieden werden musste (s. auch Oh et al., 2016; s. 5.2.2). Dass in den Zeitdruck-Gruppen kein Unterschied in der genutzten Anzahl auszumachen war, könnte darin begründet liegen, dass alle Informationen bereits verfügbar waren und nicht noch gesucht oder erarbeitet werden mussten (s. auch Dummel et al., 2016). Auch die Expertise-Gruppen unterschieden sich nicht signifikant in der genutzten Menge. Die Noviz:innen integrierten im Schnitt die meisten Informationen, was zu deren Angabe passt, ein hohes Kognitionsbedürfnis zu besitzen (s. 5.2.4). Dass laut den Mittelwerten gruppenübergreifend nicht alle Kategorien einbezogen wurden, führt nicht zwingend zu einem Verlust in der Entscheidungsqualität – auch weil Unsicherheit in einer Situation sogar dazu beitragen kann, dass auf relevante Informationen verzichtet werden muss (Gigerenzer, 2006). Es ist möglich, dass nicht die vorhandene Menge, sondern die (subjektive) Qualität der Informationen ausschlaggebend für deren Berücksichtigung war (Ettenson et al.,

[52] Siehe Fußnote 44.

1987). Folglich war das individuelle (unbewusste) Nichtberücksichtigen einer Information begründet, auch weil es sich bei der Bearbeitung der Fallbeschreibung gewissermaßen um eine unsichere, größtenteils ungewohnte Situation handelte. Zudem ist die hier ermittelte Anzahl kein Indikator dafür, ob sich auch in den Fällen in der Praxis (nicht) auf alle vorhandenen Beweismittel bezogen wird (s. auch Shanteau, 1992). Obwohl es in der Fragestellung dahingehend formuliert wurde, sich auf die *vier* Elemente zu beziehen, lässt sich nicht ausschließen, dass die Teilnehmenden andere Informationen einrechneten oder gar Informationen für ihre Entscheidungen nutzten und mitzählten, die nicht abgefragt wurden (s. 5.3).

Für die Beweiskategorien wurde deutlich, dass einzelnen Beweismitteln eine niedrigere beziehungsweise höhere Relevanz beigemessen wurde, je nachdem, welche Expertise-, Zeitdruck- oder Deliktgruppen betrachtet wurde (F5 und F6; s. 4.10.3; 4.11.2). Betrachtet man aber die Konfindenzintervalle der signifikanten Prädiktoren, wird deutlich, dass die Entscheidungsrichtung nur für die Beweismittel mit niedrigster, aber nicht für die mit höchster Relevanz eindeutig vorgeben werden konnte. Welche Information wichtig ist (im Vergleich zur Referenzgruppe), lässt sich scheinbar nicht eindeutig durch das Delikt oder den Zeitdruck vorhersagen. Die Tatsache, dass auf deskriptiver Ebene in allen Expertise-Gruppen das Erscheinungsbild des Beschuldigten häufiger eine hohe Relevanz erhielt (Augenschein), wenn Zeitdruck vorlag, könnte so interpretiert werden, dass unter Stress schnell zu verarbeitende Informationen stärker berücksichtigt werden (F5; s. 4.10.2). Das Betrachten eines Fotos lässt sich schneller handhaben als das Lesen und Nachvollziehen (und Hinterfragen) der Aussage des Zeugen oder des Beschuldigten. Stehen Personen nicht unter Zeitdruck, so haben diese mehr Ressourcen, sich einer Aussage zu widmen (Kang et al., 2012). Mit dem direkten Beweis (Augenschein; z. B. Foto) ist nicht der gleiche aufwendige Gedankenprozess verknüpft wie mit einem indirekten Beweis (z. B. Aussage; Schweizer, 2015; s. 2.1.6). Dazu passt, dass diejenigen, die sich um *Genauigkeit* bemühen sollten, sich mit steigender Wahrscheinlichkeit für die Aussage des Zeugen oder des Beschuldigen als relevantes Beweismittel entschieden. Mitunter wurde auch aus dem Grund eine der Aussagen gewählt, weil durch das Nachdenken bereits beträchtliche Ressourcen für Überlegungen aufgewendet wurden und sich somit im Sinne einer Versunkenen-Kosten-Falle dafür entschieden wurde (s. auch Kahneman, 2011). Dass die Aussage des Zeugen

gewählt wurde, bedeutet aber nicht zwingend, dass dies schließlich zur Bejahung des Tatverdachtes führte. Es ist möglich, dass die Aussage deswegen als relevant eingeschätzt wurde, weil sie das Vorliegen des Tatverdachtes nicht eindeutig bestätigen kann und somit entlastend wirkt. Die „Richtung“ der Relevanz lässt sich an den Ergebnissen ebenso wenig ablesen wie die eingeschätzte Glaubwürdigkeit des Zeugen (oder des Beschuldigten; Effer-Uhe & Mohnert, 2019; Niehaus et al., 2009).

Es wurde sich aus methodischen Gründen für das Erfassen der vier Beweiskategorien entschieden, weil ihnen die Informationselemente der Vignetten zugewiesen werden konnten und diese dadurch quantifizierbar wurden. Zudem lassen sich durch eine Manipulation der Beweisvariablen Einflüsse auf das Urteilen differenzieren (Hupfeld-Heinemann & Helversen, 2009). Jedes Beweismittel könnte allerdings wieder in weitere Informationselemente unterteilt werden, zumal nicht alle Details einer Fallbeschreibung als eigenes Beweismittel gewertet wurden (z. B. Verletzung des Geschädigten; s. 5.3). Zudem war die Festlegung des Augenscheins als Referenzkategorie in den Analysen begründet, aber doch willkürlich (s. 4.10.3). Es bleibt unklar, inwiefern sich insbesondere die Expert:innen auf eine solche kategorische Einteilung von Beweismittel bezogen oder ob sie bei der Beweiswürdigung in anderen Mustern oder Geschichten dachten (s. auch Pennington & Hastie, 1991, 1992; s. 2.1.6.2). Dazu passt die Annahme von Mitchell (1989), dass Expert:innen Sachverhalte und deren Verknüpfungen aufgrund ihrer kognitiven Schemata anders „sehen“. Mitunter ist die hier vorgenommene Aufteilung zu abstrakt, sodass sich daran keine signifikante und reale Gewichtung der einzelnen Kategorien festmachen lässt, weil ebenjene Gewichtung auch von der jeweiligen Interpretation abhängt (Sagana & Sauerland, 2020). In der Erklärung des Versuchsablaufs erhielten die Naiven zwar eine Darstellung der möglichen Beweismittel, allerdings ist auch für diese Teilstichprobe unklar, ob sie bei der Bearbeitung der Vignette auf Grundlage dieses Kategoriensystems dachten und handelten. Dass sich Unterschiede in der Einschätzung der Relevanz ergaben, ist aber ein Hinweis darauf, dass interindividuelle Mechanismen dabei eine Rolle spielen, wenngleich in dieser Studie noch nicht die dafür adäquaten Variablen erfasst wurden.

5.2.9 Einordnung der qualitativen Ergebnisse

Für die Auswertung der Angaben zu möglichen Nachermittlungen wurde sich an den vier Beweiskategorien orientiert (s. 2.1.6; 4.12.1). Eine alternative Betrachtungsweise könnte sich damit befassen, ob sich eine Angabe für den Beschuldigten als be- oder entlastend einordnen lässt (Lidén et al., 2019; s. auch Tersago et al., 2020). Doch auch wenn es andere Ansatzpunkte zur Auswertung gibt, galt die Betrachtung von Kategorien aufgrund ihrer Relevanz für die gesamte Studie als inhaltlich begründet und angemessen. In der Summe lieferten die Expert:innen in beiden Delikten die meisten Antworten. Die Vermutung liegt nahe, dass ein höherer Grad der Expertise mit höherer Kreativität und mit größerem Ideenreichtum für Ermittlungsmethoden im Zusammenhang steht – womöglich aufgrund des reichen Erfahrungsschatzes. Dazu passt die Annahme, dass die Thematik für die Expert:innen eine persönliche Bedeutung hat, sodass die Motivation zum (weiterführenden) Denken angeregt wurde und keine NFC-Effekte auftraten (Petty et al., 2009; s. 5.2.4). Auf die Gesamtheit der Daten bezogen, hinderte das Vorhandensein von Zeitdruck beim Lesen der Vignette scheinbar nicht daran, sich über die gelesenen Inhalte im Nachhinein noch Gedanken zu machen, da zumindest die Laien und Expert:innen unter Zeitdruck mehr Angaben lieferten. Dies wäre ein weiterer Hinweis auf die erfolglose Manipulation (s. 5.2.2). Im Fall „Körperverletzung“ ist es insbesondere bei den Laien auffällig, dass nur eine Angabe unter dem Einfluss von Zeitdruck gemacht wurde, wohingegen ohne diesen Stressfaktor 15 Antworten abgegeben wurden. Dies könnte daran liegen, dass die Teilnehmenden bei der schnellen Bearbeitung der Vignetten nicht so intensiv auf Detailinformationen geachtet haben. Somit konnten sie sich in der erst später gestellten Frage zu den Nachermittlungen womöglich nicht mehr an die genauen Gegebenheiten (und noch offene Fragen) erinnern, zumal die Inhalte für Naive eine gewisse Novität mit sich brachten. Eine solche Diskrepanz in der Anzahl der Antworten zeigte sich auch bei den Noviz:innen – allerdings in entgegengesetzter Richtung, da diejenigen unter Zeitdruck 19 Angaben machten (ohne zeitlichen Stress: 5 Angaben). Eine Interpretation für das Ergebnis der Noviz:innen ist, dass sich während der Bearbeitung des Falles unter Stress einige Fragen nicht beantworten ließen, sodass diese dann in Form von Nach-

ermittlungen Raum fanden. Somit wären eine Erfahrung mit der juristischen Thematik sowie die individuelle Merkfähigkeit maßgeblich dafür, ob sich trotz Zeitdruck im Nachgang noch an zu klärende Aspekte erinnert wird.

Unterscheidet man die Beweiskategorien dahingehend, dass der Augenschein sowie Urkunden eher physischer Natur sind und die Befragungen des Beschuldigten oder von Zeug:innen eher auf menschlicher Interaktion beruhen, so zeigte sich anhand der Anzahl der gemachten Angaben im Fall „Diebstahl“, dass sich insgesamt 46 Angaben auf physische, aber 34 Angaben auf verbale Beweise bezogen. Dies könnte dahingehend interpretiert werden, dass sich die Versuchsteilnehmenden insbesondere darum bemühten, eine gewisse Objektivierbarkeit der nicht eindeutigen Beweislage zu erhalten. Physische Beweismittel wirken sich auf die Einschätzung der Schuldfrage aus (Pearson et al., 2018; s. auch Daftary-Kapur et al., 2010). Außerdem können zusätzliche (valide) Beweismittel bei Aussagen von beschuldigten Personen helfen, zwischen Wahrheit und Täuschung zu unterscheiden (Wyler, 2021). Inhaltlich ging es für die Teilnehmenden in den Nachermittlungen überwiegend darum, durch „sachliche“ Hinweise die Angaben des Beschuldigten (und teilweise auch des Zeugen) zu bestätigen oder zu widerlegen. Genau genommen dient auch die Nachbefragung des Beschuldigten dazu, da sich Informationen zum Einkaufsort ergeben, die wiederum zum Überprüfen von Überwachungskameras (Augenschein) führen könnten. Es wurde sich scheinbar auf die wenigen Angaben zum Tathergang fokussiert, um aus diesen weitere Hinweisstränge ableiten zu können. Im Gegensatz zum Fall „Diebstahl“ ist die Verteilung der Beweiskategorien in eher physische und verbale Beweismittel für den Fall „Körperverletzung“ nahezu gleich verteilt (physisch: 27; verbal: 25). In den Nachermittlungen ging es zwar ebenfalls um eine Objektivierung der nicht eindeutigen Beweislage, aber auch um das Erfragen bestimmter oder anderer Perspektiven, um durch die Würdigung der (neuen) Beweise zu einem „guten“ Urteil zu kommen (z. B. Frage nach dem Motiv; soziale Beziehung zwischen beschuldigter und geschädigter Person). Dennoch bleibt unklar, inwiefern die Gleichverteilung der Angaben entweder in der Beschreibung des Falles oder tatsächlich im Delikt der Körperverletzung begründet ist, welches einen signifikanten Einfluss auf die Entscheidung hatte (H1; s. 4.2.2). Eine Körperverletzung lässt sich womöglich besser bewerten oder einschätzen, wenn die Beweismittel Auskünfte über zwischenmenschliche Dynamiken liefern können. Somit wäre

es denkbar, dass sich die Nachermittlungen qualitativ unterscheiden, je nachdem welches Delikt betrachtet wird (s. 5.2.5).

Nach Sichtung der gemachten Angaben zur Entscheidungsqualität wurden diese danach kategorisiert, ob sie sich auf eine objektive Falsifizierung der Entscheidung, auf das subjektive Entscheidungserleben, auf Feedback oder auf Sonstiges beziehen (s. 4.12.2). Die Rücklaufquote für diese optionale Frage war für beide Vignetten sehr gering, weswegen die Übertragung der Ergebnisse nur begrenzt sinnvoll ist. Der fehlende Rücklauf könnte an motivationalen oder zeitlichen Gründen gelegen haben, aufgrund derer man sich nicht mit der Qualität der Entscheidung auseinandersetzen wollte. Es ist auch möglich, dass die Frage an sich unverständlich formuliert oder zu abstrakt gehalten war. Die Tatsache, dass nur wenige Fachpersonen die Güte ihrer Entscheidung in Worte fassen und diese Güte eher an einem Gefühl als an konkreten Kriterien festmachen konnten (s. 4.12.2), deutet an, dass Menschen mit Expertise Schwierigkeiten in der Verbalisierung ihres (kognitiven) Entscheidungsverhaltens haben können (Shanteau, 1988). Laut Bullens et al. (2014) kann das Eintreten eines negativen Ergebnisses ein Hinweis auf eine falsche Entscheidung sein. Ein schlechtes Gefühl könnte ein derartiges negatives Ergebnis darstellen. Das im juristischen Kontext unrealistische Differenzieren zwischen richtigen und falschen Antworten erschwert vermutlich das Abwägen der Entscheidungsqualität, weil es Fachpersonen mutmaßlich vielmehr um das Vermeiden von großen Fehlern, aber weniger um das Finden der einzig wahren und richtigen Lösung geht (s. auch Dickert et al., 2012; Shanteau, 1988, 2000). Allerdings gaben auch die anderen beiden Gruppen insgesamt wenig Rückmeldung, sodass das Expertise-Level nicht die alleinige Erklärung sein kann. Für die Naiven war die Frage nach der Entscheidungsqualität vermutlich zu abstrakt, da bestimmte praxisbezogenen Abläufe (z. B. Feedback durch Revision) nicht bekannt sind. Fasst man Expertise-übergreifend für beide Delikte das Feedback und die objektive Falsifizierung derart zusammen (25 Angaben), dass sie eine externe Quelle der Rückmeldung darstellen (im Gegensatz zum subjektiven Erleben: 11 Angaben), so wird deutlich, dass zumindest diese kleine Teilstichprobe die Güte ihrer Entscheidung insbesondere mithilfe solcher von außen begleiteten Rückmeldungen besser einschätzen könnte. Lehrreiches Feedback ist in der juristischen Praxis allerdings ein Randphänomen (Rachlinski & Wistrich, 2017; Schweizer, 2015; Spellman, 2007; s. 2.3.3.3).

5.3 Kritische Bewertung der Methode und Darstellung von Limitationen

Es gilt, das methodische Vorgehen zu diskutieren. Dazu wird zunächst auf die Konzeption der Studie als fragebogenbasierte Onlinestudie eingegangen. Im Zusammenhang mit der Betrachtung des Fragebogendesigns wird dessen Konstruktion besprochen. Auch auf das Gütekriterium der Validität wird eingegangen. Daran anschließend werden die Vignettenmethode und die Umsetzung der (Quasi-)Experimentalgruppen sowie der personenbedingten Einflussfaktoren kritisch beleuchtet. An einigen Stellen ergeben sich bereits erste Ableitungen zu Forschungsdesideraten. Auf diese wird in Abschnitt 5.4 ausführlich eingegangen.

Der Einsatz von Onlineumfragen hat gewisse Vorteile. Zu diesen Vorteilen zählt, dass Teilnehmende an einem für sie günstigen Zeitpunkt teilnehmen können, dass durch die schnelle Verteilung der Zugangsdaten zur Umfrage ebenjener Zugang zu möglichen Teilnehmenden erleichtert wird, dass sie in der Regel mit geringen Kosten verbunden sind und dass die Daten je nach genutzter Software bereits zur Weiternutzung aufbereitet werden (McCready, 2006). Auf der Onlinemethode basierende Störvariablen, wie Missverständnisse bei der Bearbeitung der Umfrage oder ablenkende Umweltreize am Ort der Teilnahme (z. B. Geräuschkulisse im Raum oder vom Endgerät abhängige Störungen, wie eingehende Nachrichten oder Anrufe), wurden nicht näher kontrolliert. Die Vorteile der Onlinemethode überwogen aber eventuelle Nachteile. Diese Studie basierte auch auf der Fragebogenmethode (s. 3.2). Ein Nachteil von Fragebögen ist, dass die Antworten bei geschlossenen Fragen bereits vorgegeben sind (Goddard & Villanova, 2006). Davon abweichende Antworten lassen sich nicht erfassen. Es wurde darauf verzichtet, bei geschlossenen Fragen eine Antwort zu ermöglichen, die „unentschlossen“, „keine der genannten Antworten“ oder ähnlichem entspricht (Goddard & Villanova, 2006). Dies führte womöglich dazu, dass Teilnehmende sich für eine der angebotenen Antworten entscheiden mussten, obwohl diese nicht ihre Meinung repräsentierten – zumal die Items im Forced-choice-Format gehalten waren. Es wurden zwei optionale, offene Fragen integriert, da im Vorfeld keine ausreichenden Antwortoptionen für das Konstruieren einer geschlossenen Frage herausgearbeitet werden konnten (s. 3.4.2). Für die Studie war es ausreichend, sich auf einzelne deskriptive Angaben stützen zu können (Goddard & Villanova, 2006; s. 4.12).

In Ratingskalen „werden Abschnitte eines Merkmalskontinuums durch verbale Beschreibungen, Beispiele, Zahlen usw. vorgegeben, und der Untersuchungsteilnehmer markiert jene Stufe, die seinem Empfinden entspricht“ (Trimmel, 2009, S. 85). Im vignettenbezogenen Fragebogen kamen Ratingskalen in Form von bipolaren, verbal-numerischen Likert-Skalen zum Einsatz (s. 3.4.2). Dabei wurden nur die Endpunkte (1 und 7) verbal beschrieben, während die einzelnen Antwortoptionen (2 bis 6) mit numerischen Werten versehen wurden. Für diese Likert-Skalen wurde eine Intervallskalierung angenommen. Allerdings kann das Skalenniveau bei solchen Schätzskalen diskutiert werden, da die für eine Intervallskalierung erforderliche Äquidistanz der Kategorien nicht zwingend gegeben ist (Goddard & Villanova, 2006; Pospeschill, 2013; Trimmel, 2009). Es wurde darauf geachtet, die Endpunkte als semantische Gegensatzpaare zu definieren, um somit einen neutralen Mittelpunkt und ähnliche Ausprägungen links und rechts der Mitte zu erreichen (Pospeschill, 2013). Die Tatsache, dass nicht nur einzelne wenige Abstufungen von den Versuchsteilnehmenden ausgewählt wurden, sondern eine Streuung der Abstufungen erkennbar war (s. beispielsweise 4.6.1; 4.7.1), spricht für das Vorliegen einer Intervallskalierung (Leonhart, 2017; Pospeschill, 2013). Ein alternatives Format wäre es, die Frage nach der Leichtigkeit der Entscheidung auf einer Skalierung zu basieren, die je Punktwert eine Intensität ausdrückt (z. B. 1 = *gar nicht*, 2 = *kaum*, 3 = *mittelmäßig*, 4 = *ziemlich* und 5 = *außerordentlich*; s. auch Trimmel, 2009). In diesem Fall wäre eine Abstufung mit fünf Stufen ausreichend, da ansonsten die verbale Zuordnung zu komplex würde. Dies könnte auf die anderen Items übertragen werden (z. B. Sicherheit, Überzeugung). Dadurch ließe sich der Interpretationsspielraum für die Teilnehmenden hinsichtlich der Skalierung begrenzen, da die verbalen Beschreibungen im direkten Zusammenhang mit der Charakterisierung durch Zahlen stünden (Pospeschill, 2013).

Die eingesetzten Likert-Skalen waren außerdem nicht forciert, sodass es eine mittlere Antwortoption gab. Zukünftige Studien können auf forcierte Skalen ausweichen, um eine Antwort in die eine oder andere Richtung zu erzwingen (Pospeschill, 2013). Bei den unipolaren Ratingskalen *Beweismaß* und *Schwere der Tat* wurden die Variablen jeweils auf einem Kontinuum angeordnet (0–99% bzw. 0–100). Die Auswahl dieser Werte war begründet, aber dennoch willkürlich. Anstelle einer solchen Spannweite könnte zur Operationalisierung ebenfalls eine Ratingskala genutzt

werden (z. B. eine mehrstufige Likert-Skala mit den Endpunkten *nicht schwer* und *sehr schwer*). Dies würde die große Spannweite der Antworten stark verkleinern und möglicherweise konkretere Definitionen durch die Versuchsteilnehmenden zulassen. Dies wäre auch vor dem Hintergrund sinnvoll, dass numerische Angaben zum Beweismaß unter Umständen nicht gleichermaßen verstanden und definiert werden (s. 2.3.3.4; s. auch McKay et al., 2014). Wright und Hall (2007) legen dar, wie sich ein Maß berechnen lässt, ab dem sich die Hälfte der Stichprobe für einen Schuldspruch ausspricht. Eine ebensolche Berechnung für das Beweismaß könnte eine Art „Kipppunkt“ für das Vorliegen des Tatverdachtes angeben. Besagter Tatverdacht wurde als Dichotomie erhoben, ähnlich wie dies in der Empirie für die Schuldfrage umgesetzt wird (s. 2.3.3.4). Laut Hope et al. (2008) und Smithson et al. (2007) kann bei der Schuldfrage eine dritte „nicht bewiesen“-Option eingesetzt werden. Auch wenn eine solche Option (hier: „Tatverdacht nicht bewiesen“) mit dem Verneinen gleichzusetzen wäre, so würde dies eine qualitative Erhebung der Frage nach dem „Warum?“ bis zu einem gewissen Grad ersetzen können, da dadurch die Begründung zur Verneinung des Tatverdachtes bei nicht eindeutiger Beweislage abgeleitet werden könnte. Des Weiteren konnte nicht vermieden werden, dass die Richter:innen, deren Berufsgruppe mehrheitlich vertreten war (s. 3.3), ihre gewohnte Rolle einnahmen anstatt in die erforderliche Rolle der Staatsanwaltschaft zu wechseln. Ein Rollen-induzierter Bias kann aber das juristische Verhalten beeinflussen (Egli Anthonioz et al., 2019; Engel & Glöckner, 2013). Somit bleibt die Möglichkeit bestehen, dass sich einige der Richter:innen an der Beantwortung einer „Schuld“-Frage und nicht einer „Tatverdacht“-Frage orientiert haben. Auch Schmittat et al. (2022) argumentieren, dass eine Anklageerhebung nicht mit einem Urteil gleichzusetzen ist. In ihrer Studie stimmten Proband:innen zwar für eine Anklage, sprachen sich weiterführend aber nicht zwingend für einen Schuldspruch (Urteil) aus. Um kontrollieren zu können, worauf sich Teilnehmende in ihren Antworten beziehen, könnte eine genauere Differenzierung in der Operationalisierung erfolgen und beide Maße erhoben werden. Dies würde inhaltlich allerdings einen gedanklichen Perspektivwechsel von der Staatsanwaltschaft zum Gericht mit sich bringen. Mit Blick auf die Entscheidung über den nächsten Verfahrensschritt gilt es anzumerken, dass es in der Praxis weitere Entscheidungs- und Handlungsoptionen gibt, als in dieser Studie berücksichtigt wurden (s. 2.1.2). Durch die möglichst realistische Auswahl an Optionen wurde versucht, einen großen Realitätsbezug zu den

gesetzlichen Rahmenbedingungen erreichen (s. auch Sagana & van Toor, 2020). Allerdings ließen sich nicht alle Entscheidungswege operationalisieren, da dies in der Erfassung und der Auswertung zu komplex werden würde. Es besteht somit weiterhin die Möglichkeit, dass einige der Noviz:innen oder Expert:innen beispielsweise zum beschleunigten Verfahren (§ 417 StPO) tendieren würden, welches in dieser Studie aber nicht zur Auswahl stand.

Das methodische Vorgehen in dieser Studie bringt gewisse Einschränkungen in der Validität mit sich. Auf eine reduzierte interne Validität bezieht sich die Tatsache, dass es sich aufgrund des Faktors „Expertise" um ein quasi-experimentelles Design handelte (Bortz & Schuster, 2010; Trimmel, 2009). Zudem enthielt die Studie einige nicht kontrollierte Variablen, insbesondere hinsichtlich der Teilnahmeumgebung, was aber vorrangig in der gewählten Methode der fragebogenbasierten Onlinestudie begründet ist. Positiv zu bewerten ist allerdings, dass die Proband:innen keine Informationen darüber hatten, dass ein Expertise- oder Zeitdruck-abhängiger Vergleich zwischen Gruppen stattfindet. Auf allgemeine Einschränkungen in der externen Validität, die insbesondere durch die Vignettenmethode hervorgerufen werden (z. B. hypothetische und abstrakte Fallbeschreibungen, fehlendes Konsequenzerleben), wurde bereits eingegangen (s. 2.3.1). Es folgen nun konkrete Bezüge zur vorliegenden Studie. Die externe Validität ist reduziert, wenn die Ergebnisse auf sozial erwünschten Antworten beruhen oder wenn eine Stichprobe nicht repräsentativ ist (s. auch Bortz & Schuster, 2010; Trimmel, 2009). Aufgrund der Anonymität der gesamten Befragung kann sozial erwünschtes Verhalten nicht ausgeschlossen, aber als gering eingestuft werden (Pospeschill, 2013). Solches Verhalten blieb bewusst unkontrolliert. Auch Lidén et al. (2019) fanden in ihrer Studie mit einer Stichprobe aus Vertreter:innen der Staatsanwaltschaft keine Hinweise auf sozial erwünschtes Verhalten. Lediglich die NFC-Kurzskala beruht auf Selbstberichten (s. 3.4.3), was mitunter zu sozial erwünschten Antworten führen könnte (Peters & Dörfler, 2019a). Der Vorteil von Selbstberichten ist allerdings, dass das individuelle Erleben der Teilnehmenden abgebildet werden kann. Die Repräsentativität der Stichprobe ist bereits als hoch einzustufen, könnte aber durch das Befragen insbesondere von Mitarbeitenden der Staatsanwaltschaft weiter erhöht werden.

Die externe Validität ist ebenfalls reduziert, wenn eine Untersuchungsmethode nicht generalisierbar und „unnatürlich" ist. Für diese Studie lassen sich gewisse

Hinweise auf eine reduzierte externe Validität ableiten. Im Sinne der Beweiswürdigungsmodelle werden Fallinformationen miteinander verknüpft und in Beziehung gebracht, sodass sie – kurz gesagt – kohärent werden (s. 2.1.6.2). Durch die Vignetten wurden die Beweismittel bereits aufgearbeitet und in Form einer Geschichte präsentiert. Dadurch erhielten die Teilnehmenden eine bereits festgelegte Menge an Informationen, aus denen es lediglich Schlüsse abzuleiten galt. Das Auswählen von Informationen und die selbstständige Konstruktion von Kohärenz waren somit nicht mehr gefordert (s. auch Ask & Granhag, 2005). Eine praxisnahe *graduelle* Eindrucksbildung und Entscheidungsfindung fanden nicht statt (s. auch Konečni & Ebbesen, 1979, 1982). Die Menge der präsentierten Informationen innerhalb der Fallbeschreibung war überschaubar und der zu bearbeitende Umfang sehr wahrscheinlich geringer, als dies auf tatsächliche Fallakten zutrifft. Insbesondere bei zunehmender Anzahl an Informationen wechseln Expert:innen zu nicht-linearen und nicht-kompensatorischen Entscheidungsstrategien (Einhorn, 1971; Pachur & Marinello, 2013). Daher ist es denkbar, dass die Fallbeschreibungen nicht ausreichend umfassend waren, um den Wechsel von Entscheidungsstrategien oder den Verlass auf Intuition zu initiieren. Des Weiteren enthielten die Fallbeschreibungen Details, die für eine Entscheidung möglicherweise relevant waren, aber nicht abgefragt wurden (z. B. die Beschreibung der Verletzung des Opfers im Fall „Körperverletzung“). Kenntnisse über eine Verletzung oder die Konsequenz eines Verbrechens stellen aber einen relevanten Einflussfaktor dar (Rossi et al., 1985) und sollten nicht unberücksichtigt bleiben. Da die in dieser Studie getroffenen Entscheidungen keinerlei Konsequenzen mit sich brachten, fehlt ein gewisser Praxisbezug (Bieneck, 2009). Allerdings wurde bereits diskutiert, dass fehlendes Feedback, zeitlich verzögerte Rückmeldungen (Konsequenzen) oder gar ein geringes Ausmaß an Verantwortung auch im juristischen Praxisalltag gegeben sind (Guthrie et al., 2007; Rachlinski & Wistrich, 2017; Spellman, 2007). Daher könnte überspitzt argumentiert werden, dass das Fehlen von Konsequenzen im Rahmen dieser Studie dem beruflichen Alltag recht nahe kommt – wenngleich diese theoretischen und praktischen Formen des Konsequenzerlebens selbstverständlich nicht gleichzusetzen sind.

Auch unter dem Gesichtspunkt der (fehlenden) Generalisierbarkeit der Ergebnisse sind Einschränkungen in den externen Validität zu nennen. Zunächst gilt es zu be-

tonen, dass im Rahmen der Studie nicht ausführlich auf alle Inhalte und Besonderheiten des Strafprozessrechts eingegangen werden konnte, sondern dass die zusammenfassende Darstellung nur ausgewählte Inhalte enthielt (s. 2.1.1; 2.1.7). Es gilt außerdem zu beachten, dass in dieser Studie allgemeine Strafsachen gegen Erwachsene im Fokus standen, die erstinstanzlich verhandelt werden. Auf die rechtliche Grundlage bei oder auf eine Übertragbarkeit der Ergebnisse auf die Straftaten Jugendlicher (14–17 Jahre) oder Heranwachsender (18–20 Jahre), auf den Ablauf von Sonderformen des Strafprozesses oder auf Berufungs- und Revisionsverfahren wird nicht näher eingegangen.[53] Auch der direkte Vergleich zwischen Justizsystemen verschiedener Länder – und folglich eine Generalisierung von Ergebnissen – ist nur begrenzt möglich (Dhami & Ayton, 2001). Dickert et al. (2012) argumentieren allerdings, dass die Betrachtung kognitiver und emotionaler Mechanismen von (Laien-)Richter:innen unabhängig vom Rechtssystem erfolgen kann. Demzufolge ist auch eine vom Rechtssystem losgelöste Betrachtung der hier untersuchten (personenbezogenen) Einflussfaktoren begründet. Außerdem besitzt diese Studie einen besonderen Vorteil hinsichtlich der Übertragbarkeit der Ergebnisse: Die Zusammensetzung der Stichprobe mit einem beachtenswerten Anteil an Rechtsexpert:innen (s. 3.3). Erkenntnisse zu juristischer Entscheidungsfindung werden somit nicht anhand einer üblichen nicht-fachlichen, studentischen Stichprobe getroffen, sondern basieren auf einer Gruppe von Fachpersonen mit einem direkten, validen Praxisbezug (s. auch Konečni & Ebbesen, 1979).

Die Kernelemente der Methode waren die neu konstruierten Fallbeispiele (s. 3.4.1). Die Auswahl der Informationen, die in eine Vignette eingearbeitet werden, beruht auf theoretischen Ableitungen, sodass in der Regel nur solche Aspekte ausgewählt werden, denen eine gewisse Relevanz für die Fragestellung zugesprochen werden. Demnach können mit Vignetten nur diese ausgewählten Merkmale untersucht und deren Bedeutung eingeschätzt werden. Folglich kann dieses Vorgehen aber weniger einer „Aufdeckung inhaltlich *erschöpfender* Urteilsregeln“ (Auspurg, Hinz & Liebig, 2009, S. 89) entsprechen, da lediglich die ausgewählten Merkmale fokussiert werden. Es konnte nicht überprüft werden, welches Vorwissen insbesondere bei Expert:innen aktiviert wurde oder inwiefern von der Fallbeschreibung abgewichen

[53] Für weiterführende juristische Informationen wird auf die StPO sowie das StGB verwiesen.

und sich auf eigene Erfahrungen mit vergleichbaren Fällen zurückbezogen wurde (Bieneck, 2009). Des Weiteren wurde nicht überprüft, ob Proband:innen möglicherweise Verständnisschwierigkeiten hatten und sich die hypothetische Situation nicht wie intendiert vorstellen konnten (Groß & Börensen, 2009). Allerdings gaben die Ergebnisse der erfassten Variablen zur Leichtigkeit, zur Sicherheit und zum Realismus der Fälle keinen Hinweis darauf, dass schwerwiegende Probleme auf Seiten der Proband:innen vorlagen (s. 4.6.1; 4.10.2). Für die Eignung der neu konstruierten Fallbeispiele spricht außerdem die Tendenz zur Mitte hinsichtlich des Beweismaßes, welches die beabsichtigte fehlende Eindeutigkeit der Beweislage untermauert (s. 5.2.6). Ebenso spricht für die Eignung der Vignetten die Tatsache, dass beide Fallbeispiele in der Gesamtstichprobe nahezu gleiche Mittelwerte und Standardabweichungen erzielten. Dies stützt die Annahme, dass diese als vergleichbar schwer wahrgenommen wurden. Somit sind die Delikte nicht nur hinsichtlich des durch das StGB festgelegten Strafmaßes vergleichbar, welches das ursprüngliche Kriterium für deren Auswahl war, sondern dies wurde scheinbar auch von den Teilnehmenden derart eingeschätzt (s. 3.4.1; 4.5.1). Es wurde aber nicht kontrolliert, inwiefern sich Merkmale der Studienteilnehmenden, wie Alter oder Bildungshintergrund, auf die Bearbeitung der Vignette ausgewirkt haben könnten. So argumentieren Sauer et al. (2011) zwar, dass sich Vignetten als Methode für die allgemeine Bevölkerung eignen, aber dass bestimmte Merkmale je nach Anzahl der Vignetten und ihrer Dimensionen zu unterschiedlichen Reaktionen führen, die letztlich nicht auf das Fallbeispiel, sondern auf diese Merkmale zurückzuführen sein könnten. Auch wenn in dieser Studie keine Dimensionen manipuliert wurden, so bleibt dennoch die Möglichkeit bestehen, dass die Komplexität je nach Alter oder Bildungshintergrund zu einer gewissen Überforderung geführt haben könnte. Somit wären die erfassten Antworten eher das Ergebnis der Methode und nicht der tatsächlichen Einstellung der Befragten (Auspurg, Hinz, Liebig & Sauer, 2009) – zumal die Bearbeitung abstrakter Fallbeschreibungen wesentlich von einer gewissen Vorstellungskraft der Studienteilnehmenden abhängt.

Auch mit Blick auf die für die Studie zentralen (Quasi-)Experimentalgruppen lassen sich gewisse Limitationen und Anpassungsideen ableiten (Expertise, Zeitdruck, Delikt). Zunächst gilt es festzuhalten, dass die vielen Untergruppen des Designs zu teilweise sehr problematischen Zellgrößen geführt haben. Daran anschließend kam

es zu einer hohen Komplexität in der Auswertung und Interpretation der Ergebnisse (s. auch Peters & Dörfler, 2019a; Trimmel, 2009). Dies könnte durch eine höhere Anzahl an Teilnehmenden oder durch eine inhaltliche Spezifizierung der Fragestellung auf nur eine dieser Gruppen umgangen werden. In dieser Studie galten die Laien in den meisten Analysen als Referenzgruppe für Vergleiche, da sie keinerlei juristische Vorbildung besitzen. Ein Wechsel der Referenzgruppe auf die Expert:innen würde sicherlich vergleichbare Ergebnisse liefern, aber möglicherweise auch qualitative Unterschiede in der Interpretation bieten. Fokussiert man folglich den Faktor „Expertise“, ergibt sich weiterführend die Möglichkeit, diesen unter gewissen Gesichtspunkten zu unterteilen. Eine Unterteilung wäre nicht hinsichtlich der verschiedenen Ausprägungen juristischer Vorbildung, sondern hinsichtlich der Domänenspezifität sinnvoll, da ein Unterschied zwischen domänenspezifischen und allgemeinen Rechtsexpert:innen besteht (Schmittat & Englich, 2016). Demzufolge könnten Rechtsexpert:innen des Strafrechts – welches in dieser Studie Gegenstand war – mit Fachpersonen des Öffentlichen Rechts oder des Zivilrechts verglichen werden, die teilweise mit anderen Beweismaßen arbeiten (s. 2.1.6; 5.2.6). Ebenso sinnvoll wäre es, angesichts der vorherrschenden Rolle, die die Staatsanwaltschaft im Ermittlungsverfahren einnimmt, eine rein staatsanwaltschaftliche Stichprobe zu erheben. Nichtsdestotrotz haben all diese verschiedenen Expertise-Unterteilungen gemeinsam, dass ihnen zugeteilte Fachpersonen eine vergleichbare (Grund-)Ausbildung erfahren haben (s. auch Glöckner et al., 2013). Doch auch wenn sich die akademische Ausbildung ähnelt, stellt sich die Frage: Ab wann gilt man als Expert:in? Auch wenn dieser Status mit der Berufsbezeichnung einhergehen kann (Chi, 2006; Shanteau, 1988), sind Unterschiede im Entscheidungsverhalten zwischen Berufsanfänger:innen und Erfahrenen denkbar und untersuchungswürdig. Innerhalb der Gruppe der Laien wurde nicht weiter nach Vorwissen unterschieden. Es ist allerdings möglich, dass sich ehrenamtliche Richter:innen oder anderweitig am Rechtswesen interessierte Personen unter den Teilnehmenden befanden. Diese konnten auf einen breiteren Wissens- und Erfahrungsschatz zurückgreifen als andere Laien. Da sie aber die Frage nach einer Vorbildung im juristischen Fachbereich explizit verneint hatten (s. 3.2), galten sie weiterhin als Laien.

Die Differenzierung in Deliktgruppen war ursprünglich methodisch und nicht rein inhaltlich begründet, um die Eignung der neu erstellten Vignetten an einer relevanten Stichprobe zu untersuchen. Was aber als eine Art methodische Kontrolle gedacht war, erwies sich an einzelnen Stellen als signifikanter Faktor (s. 5.2.5). Dementsprechend gilt es als zukünftige Fragestellung zu untersuchen, inwiefern dies auf den Inhalt und die Formulierung der jeweiligen Vignette oder auf das darin beschriebene Delikt und dessen Natur zurückzuführen ist (s. auch McKay et al., 2014). Demzufolge sollte nicht nur die Erfahrung (von Expert:innen) mit einem Delikttyp erfasst werden, sondern auch die emotionale Reaktion, die ausgelöst wird (Dickert et al., 2012). Allerdings zeigten Pearson et al. (2018), dass ein deliktbezogener Bias bei Nicht-Fachpersonen stärker ausgeprägt war als bei Fachpersonen, zumindest mit Blick auf eine zu bewertende Schuldfrage. Dies lässt die Vermutung zu, dass Emotionen bei Expert:innen eine eher untergeordnete Rolle spielen könnten (s. auch Gabriel, 2009).

Der Zeitdruck wurde in dieser Studie durch Instruktionen manipuliert (s. 3.4.2.1). Alternativ wäre es möglich, andere Operationalisierungen zu wählen, da die Instruktionen nur für eine der drei Zeitmessungen erfolgreich waren (s. 4.1). Zu den Alternativen zählen das Festlegen einer Frist oder das Ablaufen einer fixen Bearbeitungszeit (Wickelgren, 1977). Für die Zeit, die es zum Lesen einer Fallbeschreibung braucht, sollte die Ausgangslage bestimmt werden. Da Menschen unterschiedlich schnell lesen, konnte nicht eindeutig festgestellt werden, ob eine lange Lesezeit daher zustande kam, dass die Person ohne Zeitdruck agierte oder weil sie langsam las. Die Effektivität der Manipulation wurde in dieser Studie unzureichend überprüft. Unabhängig von der Operationalisierung des Zeitdrucks sollten die Studienteilnehmenden daher befragt werden, ob und in welchem Ausmaß sie sich tatsächlich gestresst fühlen und wie sie die Zeit oder das Gefühl von Dringlichkeit wahrnahmen (Alison et al., 2013; Rastegary & Landy, 1993).

Zu den personenbezogenen Einflussfaktoren zählte die Fähigkeit zur kognitiven Reflexion. Personen, denen die Items des CRT vertraut sind, erzielen mitunter höhere Werte als diejenigen, für die die Aufgaben unbekannt sind (Haigh, 2016). Da es sich hier um keine für psychologische Studien übliche Stichprobe aus überwiegend (Psychologie-)Studierenden handelte, war nicht davon auszugehen, dass die Items des CRT übermäßig bekannt waren (s. 3.4.4). Vorherige Kenntnis der Items

hat außerdem keinen oder nur einen schwachen Einfluss auf das Finden der Lösungen beziehungsweise die Vorhersagekraft des CRT (Bialek & Pennycook, 2018; Brañas-Garza et al., 2015). Allerdings konnte nicht überprüft werden, ob Versuchsteilnehmende – die während der Teilnahme Zugang zum Internet hatten – die Lösungen in Suchmaschinen nachgeschaut haben. Die erreichten Punktwerte lassen sich somit nicht nur mit tatsächlich hohen Fähigkeiten zur kognitiven Reflexion erklären, sondern auch mit der Tatsache, dass die Teilnehmenden unbeobachtet waren. Möglicherweise recherchierten einzelne Personen die Lösungen oder baten im Raum anwesende Personen um Hilfe. Mit Blick auf die Leistungen im CRT lässt sich feststellen, dass Männer in der Regel besser abschneiden als Frauen (Brañas-Garza et al., 2015). Es wurde nicht inferenzstatistisch kontrolliert, inwiefern dieser Geschlechterunterschied in der vorliegenden Studie zum Tragen kommt. Auf deskriptiver Ebene zeigten sich tatsächlich gewisse Leistungsunterschiede in Abhängigkeit des Geschlechts: Männer schnitten besser ab, lösten häufiger alle Items korrekt und erzielten weniger häufig null Punkte im Vergleich zu weiblichen und nichtbinären Personen (s. 4.7.1). Des Weiteren argumentieren Brañas-Garza et al. (2015), dass sich der Zeitpunkt, an dem der CRT innerhalb des Versuchsablaufs zum Einsatz kommt, auf die Leistung auswirken kann: Insbesondere ein Einsatz zum Ende eines Experiments, wie es hier der Fall war (s. Abbildung 3.1), kann zu schlechteren Leistungen führen. Es wurde bereits argumentiert, dass der CRT erst zum Ende der Studie eingesetzt wurde, um einem Dropout bei den Hauptvariablen entgegenzuwirken (s. 3.2). Damit wurden gewisse Leistungseinbußen in Kauf genommen. Möchte zukünftige Forschung den Schwerpunkt zunehmend auf kognitive Reflexion legen, sollte der Versuchsablauf angepasst und auf Erweiterungen des CRT zurückgegriffen werden (s. 2.2.3.3), um solche Quellen der Beeinflussung zu umgehen. Eine Verzerrung der Ergebnisse im CRT könnte außerdem durch das Erheben des Vorwissens über die Items oder der mathematischen Fähigkeiten vermieden werden, oder gar durch eine Kontrolle des IQ-Wertes (s. auch Blacksmith et al., 2019; Liberali et al., 2012).

Zur Erfassung des Need for Cognition als personenbezogenen Einflussfaktor kam insbesondere aus ökonomischen Gründen eine Kurzskala mit vier Items zum Einsatz (s. 3.4.3). In dieser Studie wurde das Kognitionsbedürfnis auf einem Kontinuum zwischen dem Minimal- und Maximalpunktwert verstanden. Zur Einteilung

von Menschen mit hohen und niedrigen Ausprägungen könnte – bei einer höheren Anzahl von Items – ein Mediansplit eingesetzt werden (Pechtl, 2009). Auch eine moderate Ausprägung ließe sich als dritte Gruppe ermitteln. Eine solche zwei- oder dreifache Unterteilung würde es ermöglichen, Unterschiede genauer zu untersuchen, und zwar in der Beantwortung der Frage nach dem Tatverdacht (oder der Verurteilungsrate; Leippe et al., 2004), im Erleben von kognitiver Herausforderung und im Umgang mit Informationen (s. auch Verplanken, 1993). Doch Pechtl (2009) argumentiert auch, dass die in den Skalen verwendeten Items nicht gleichermaßen zur Erfassung des Kognitionsbedürfnisses geeignet sind, da teilweise unterschiedliche Konstrukte oder Subdimensionen abgefragt werden (Inhaltsvalidität). Des Weiteren sei eine situationsspezifische Formulierung sinnvoll, um sich von einer allgemeingültigen Vagheit zu entfernen. Bieneck (2006) konnte zwar zeigen, dass die Erfassung von NFC hinsichtlich allgemeiner und rechtsspezifischer Problemstellungen durchaus erfolgen kann, aber nur wenig Mehrwert bietet. Daher sollte eher auf die ursprünglichen Skalen, aber nicht zwingend auf eine bereichsspezifische Messung, zurückgegriffen werden. Nichtsdestotrotz leistet diese Studie durch die Erhebung des Need for Cognition – und auch der kognitiven Reflexion – in einer derart spezifischen Stichprobe von (angehenden) juristischen Fachpersonen einen wichtigen Forschungsbeitrag (s. auch Carnevale et al., 2011). Diese Studie reflektiert zudem auch die Bereitschaft von Rechtsexpert:innen, sich an empirischer Forschung zu beteiligen, welche in einer solchen Fachgruppe eher Skepsis auslösen kann (Engel & Gigerenzer, 2006; Sagana & van Toor, 2020).

5.4 Implikationen der Studie und Forschungsdesiderate

Im Rahmen der kritischen Bewertung der Methode sowie der Darstellung von Limitationen wurden bereits erste Forschungsansätze angerissen (s. 5.3), welche im Folgenden teilweise weiter ausgeführt und durch die Ableitung zusätzlicher Forschungsdesiderate ergänzt werden. Um angesichts der Skepsis von Rechtsexpert:innen den Transfer wissenschaftlicher Erkenntnisse in die Praxis zu erleichtern, sollte der Fokus auch darauf liegen zu erklären, warum und wie solche Erkenntnisse bereichernd sein können (Redding & Reppucci, 1999). Im Zusammenhang mit diesem Transfer ist erneut die Zusammensetzung der Stichprobe als besondere Stärke dieser Studie zu betonen, da die juristischen Expert:innen etwa ein Drittel ausmachten

(s. 3.3). Dadurch lassen sich die Erkenntnisse und Implikationen verlässlicher auf die Praxis übertragen, als dies bei den üblichen studentischen (nicht-fachlichen) Stichproben möglich ist (s. 5.3). Die Tatsache, dass die Anzahl der Expert:innen in Relation zu den beiden anderen Expertise-Gruppen angemessen hoch war, kann dahingehend gewertet werden, dass diese Personen durchaus eine Bereitschaft zur Mitarbeit und ein Interesse an empirischer Forschung besitzen – zumal der als Entschädigung angebotene Wert des Büchergutscheins vermutlich nicht der ausschlaggebende Anreiz für eine Teilnahme war (s. 3.3).

Ein zentrales Ergebnis der Studie ist, dass die „extremen" Expertise-Gruppen bei gleicher Beweislage zu konträren Entscheidungen kommen, zumindest hinsichtlich des Tatverdachtes: Laien bejahten diesen mehrheitlich, wohingegen Fachpersonen verneinten (H1; s. 4.2.2). Die Frage nach dem Tatverdacht stellt sich an einem für das Strafverfahren hochrelevanten Zeitpunkt, nämlich am Ende des Ermittlungsverfahrens, das wegweisend für den weiteren Prozess gilt (Ellison & Brennan, 2016; Morgan et al., 2018; Schmittat et al., 2022; s. 2.1.7; 2.3.5). Die Diskrepanz in den Entscheidungen weist darauf hin, dass in der Gruppe der Fachpersonen durchaus Prozesse und Mechanismen wirksam werden, die sich entweder von den Prozessen und Mechanismen der Laien qualitativ unterscheiden oder die vergleichbar sind, aber andere Ergebnisse mit sich bringen. Demzufolge begründet auch diese Studie die allgemeinen Bestrebungen, juristische Entscheidungsfindung zu untersuchen, um ein tiefergehendes Verständnis dessen zu erhalten (s. auch Danziger et al., 2011; Ellison & Brennan, 2016). Die im Fokus stehende Entscheidung über das Vorliegen des Tatverdachtes entsteht aus der Würdigung der Beweismittel. Die besagte Beweiswürdigung besitzt gemäß der gesetzlichen Vorgaben gewisse Freiheitsgrade (s. 2.1.6). Dass sich Fachpersonen mehrheitlich für die gleiche Annahme (gegen den Tatverdacht) aussprachen, ist ein Hinweis darauf, dass innerhalb der Gruppe auf vergleichbare Weise mit diesen Freiheiten umgegangen wurde. Dennoch gelangte ein Teil der Expert:innen zu einem anderen Schluss, sprach sich für das Vorliegen des Tatverdachtes aus und fühlte sich vermutlich ebenfalls sicher und überzeugt angesichts der eigenen Entscheidung.[54] In der Rechtsprechung ist

[54] Es wurde nicht analysiert, ob sich die Expert:innen in den Prozessmerkmalen der Leichtigkeit, Sicherheit und Überzeugung unterschieden, je nachdem, ob sie den Tatverdacht zuvor bejaht oder verneint hatten. Möglicherweise schätzte die Minderheit, die sich gegen das Vorliegen entschieden hatte, den Entscheidungsprozess anders ein.

das Auftreten von Ungleichheiten laut Maguire (2010) zwar zu erwarten, aber die Prädiktoren für solche Disparitäten gilt es weiter zu untersuchen. Die Notwendigkeit der Fortsetzung der Forschung ist ausgehend von den Ergebnissen dieser Studie auch deswegen begründet, weil die hier signifikanten Prädiktoren „Expertise“ und „Delikt“ lediglich kleine Effekte ausübten und zumindest in Form des Delikttyps derart nicht erwartet wurden. Letzteres spricht dafür, den Einfluss verschiedener Delikttypen auf die entscheidende Fachperson zu untersuchen. Ebenso spricht jenes Ergebnis dafür, dass sich die Art des Delikts signifikant auf die Bewertung des Schweregrades (in Interaktion mit Expertise; H4; s. 5.2.7) und auf die Wahl des Beweismittels mit höchster Bedeutsamkeit (F5; s. 5.2.8) – und damit auf die Beweiswürdigung – auswirkte. Welche Eigenschaften von Delikten lassen sich ausmachen, die die Entscheidung über den Tatverdacht in die eine oder andere Richtung lenken? Entspricht die subjektiv wahrgenommene Schwere einer Tat dem im StGB angegebenen Strafrahmen? Werden Beweise unterschiedlich interpretiert, je nach Natur des Falles? Hier wurde vermutet, dass sich ein Eigentumsdelikt von einer Körperverletzung dahingehend unterscheidet, als dass bei letzterer eine Person tatsächlich physisch angegangen und in ihrer Unversehrtheit geschädigt wird. Möglicherweise wirkt sich das empathische Nachvollziehen von Schmerzen oder Mitleid belastend beziehungsweise „strafschärfend“ aus (s. auch McKay et al., 2014). Nichtsdestotrotz gibt es auch Hinweise darauf, dass Jurist:innen weniger emotional auf Fallinhalte reagieren (Schweizer, 2015). Ein Within-Design könnte Erkenntnisse bieten, inwiefern intra- oder doch interindividuelle Faktoren (im Mixed-Design) wirksam sind.

Diese Studie beschrieb die Expert:innen anhand der Ergebnisse des CRT und der NFC-Kurzskala zwar als eine eher reflektierte, zum Denken motivierte Teilstichprobe (s. 5.2.3; 5.2.4), allerdings weist die empirische Forschung auch auf intuitives Verhalten hin (s. 2.3.4.1). Glöckner und Ebert (2011) betonen, dass es nicht darum gehen darf, Intuition vollends zu verbieten, aber es soll in der Praxis ein Rückbezug zu alternativen Interpretationen stattfinden. Laut Guthrie et al. (2001) kann das Berücksichtigen verschiedener Perspektiven (für Richter:innen) wichtig sein, um eine gute Entscheidung zu treffen. Schmittat et al. (2022) zeigten, dass sich die Wahrscheinlichkeit für eine Verurteilung reduzierte, wenn die Verteidigung im Ermitt-

lungsverfahren eine schriftliche alternative Interpretation der Beweislage präsentierte. Disparitäten innerhalb der Gruppe der Expert:innen, wie sie in gewissem Maße auch in dieser Studie aufgezeigt wurden (Zwei-Drittel-Mehrheit), könnten dadurch reduziert werden, dass sich über die Fälle ausgetauscht und somit (fachliche) Intuition in einem angemessenen Rahmen gehalten wird (Guthrie et al., 2007). Durch die Auswertung der qualitativen Angaben zur Entscheidungsqualität wurde deutlich, dass insbesondere Feedback dafür wichtig ist, das eigene Handeln einordnen zu können (s. 5.2.9). Eine weitere Erkenntnis der Studie war, dass sich Expert:innen signifikant sicherer und überzeugter in ihren Entscheidungen fühlten als Laien und Noviz:innen (F1; s. 4.6.2). Es ist aber denkbar, dass eine solche Überzeugung in der Praxis deswegen kommt, eben weil das nötige Feedback und insbesondere auch ein gewisses Konsequenzerleben ausbleiben. Guthrie et al. (2007) drücken es folgendermaßen aus: „In addition, errors seldom have direct adverse consequences for judges – when the judge slips, the litigant falls“ (S. 34). Die Rahmenbedingungen in der praktischen Arbeit lassen sich zwar nur begrenzt beeinflussen – insbesondere mit Blick auf die Statistiken der Staatsanwaltschaft und der Strafgerichte für das zu leistende Pensum (s. 2.1.5) – allerdings können Veränderungen das Justizsystem zu einer weniger bösartigen Lernumgebung machen (Gigerenzer, 2006; Guthrie et al., 2007; Hupfeld-Heinemann & Helversen, 2009; Schweizer, 2009; s. 2.3.3.3). Sporer und Goodman-Delahunty (2009) argumentieren, dass das Übertragen der Verantwortung von einzelnen Richter:innen auf Gruppen oder Tribunale dazu beitragen kann, Ungleichheiten in der Rechtsprechung zu reduzieren (wie es bei bestimmten Fällen in Deutschland bereits gehandhabt wird; s. auch Machura, 2016). Bisher wurde sich aber vorrangig auf das Hauptverfahren und die Arbeit von Richter:innen bezogen (s auch Guthrie et al., 2007). Aufgrund des Zusammenhangs zwischen den Entscheidungen zu einzelnen Zeitpunkten im Strafverfahren lassen sich Erkenntnisse zum Ermittlungsverfahren auch auf das Hauptverfahren beziehen und umgekehrt (B. D. Johnson & Stewart, 2016). Da hier die Perspektive der Staatsanwaltschaft einbezogen wurde, lassen sich die Annahmen folglich darauf übertragen. So könnte die Verteilung der Verantwortung auf

mehrere Fachpersonen (hier: Mitglieder der Staatsanwaltschaft) am Ende des Ermittlungsverfahrens begründet sein.[55] Es ist untersuchenswert, ob Rechtsexpert:innen, die den Raum bekommen, sich über eine nicht eindeutige Beweislage in einem Delikt mit geringer Schwere auszutauschen, weiterhin Disparitäten in ihren individuellen Entscheidungen (zur Tatverdachtsfrage) aufweisen und ob sich Ähnlichkeiten oder Unterschiede in der Beweiswürdigung zeigen.

Im Rahmen der Hauptverhandlung gilt es zwar, das Urteil zu begründen, allerdings könnte eine solche Erklärung – oder gar das Verfassen einer alternativen Interpretation der Beweislage (s. auch Schmittat et al., 2022) – bereits am Ende des Ermittlungsverfahrens gefordert werden. Eine Erklärung am Ende des Hauptverfahrens trägt zu einer tiefergehenden Elaboration der Beweislage bei (Maegherman, 2021), was sich ebenfalls auf das Ermittlungsverfahren übertragen ließe. Dies würde zu einer strukturierteren (und möglicherweise intensiveren) Auseinandersetzung mit dem Beweismaß in einem Fall führen, welches laut dieser Studie vermutlich auch von den Fachpersonen nicht angemessen umgesetzt wurde (s. 5.2.6). Eine höhere Begründungsdichte und eingeschränktes Ermessen gehen zudem mit reflektierten Entscheidungen einher (Schweizer, 2009). Laut Shanteau und Stewart (1992) gibt es oft den Raum für Verbesserung bei Entscheidungen von Expert:innen. Gleichermaßen könnte ein Austausch mit anderen Rechtsexpert:innen zu einem gesteigerten Verantwortungsgefühl oder zu extrinsischer Motivation führen, da es sich vor relevanten und professionellen Personen zu rechtfertigen gilt (Lerner & Tetlock, 1999; D. Simon, 1998). Zudem kann Verantwortung ein protektiver Faktor gegen Urteilsverzerrungen sein (Schmittat & Englich, 2016). Eine praktische Umsetzung der Maßnahmen zur Verbesserung der Feedbackkultur im Justizsystem bedarf sicherlich politischer Unterstützung: "Still, gains in accuracy, and therefore justice, may be worth the costs of reform“ (Guthrie et al., 2007, S. 43).[56] Doch auch wenn es im Ablauf des Strafprozesses nicht explizit vorgesehen ist, so kann nicht ausgeschlossen werden, dass ein informeller Austausch und ein Feedbackgeben unter Rechtsexpert:innen bereits stattfindet. Eine Verankerung in prozessualen Vorgaben (zum Ermittlungsverfahren) wäre allerdings wünschenswert.

[55] Aus methodischer Sicht kann in Anlehnung an B. D. Johnson und Stewart (2016) allerdings argumentiert werden, dass eine erhöhte Anzahl an Menschen im Strafverfahren zu kaum messbaren Dynamiken führen würde.

[56] Für weitere Ausführungen zu Reformideen s. Combé (2007) und Guthrie et al. (2001, 2007).

Die Noviz:innen wirkten in ihrer Entscheidung über den Tatverdacht unentschlossen, da keine gruppenbezogene Einheitlichkeit zu erkennen war (H1; s. 4.2.1). Dies könnte damit zusammenhängen, dass die Teilnehmenden sich nicht zu den Fallinhalten austauschen konnten, kein Hilfsmaterial zur Hand hatten oder insgesamt noch nicht den Erfahrungsschatz der Expert:innen besitzen, der wiederum als mögliche Erklärung für deren eindeutigere Tendenz dient (s. 5.2.1). Die Disparität war in dieser mittleren Expertise-Gruppe, im Vergleich zu den anderen Ausprägungen, am deutlichsten zu beobachten. Studierende erarbeiten sich in Übungsklausuren Kenntnisse und Fähigkeiten in der (analytischen) Fallbearbeitung (s. auch Bieneck, 2006; Glöckner et al., 2013). Gemäß dem Juristenausbildungsgesetz Nordrhein-Westfalen werden diejenigen zur staatlichen Pflichtfachprüfung zugelassen, die „ferner an Lehrveranstaltungen für Juristinnen und Juristen über die Grundlagen und die Erkenntnismöglichkeiten der politischen Wissenschaft, der Sozialwissenschaft und der Psychologie teilgenommen haben“ (§ 7 Abs. 2). Die Psychologie spielt für die Ausbildung somit nur eine kleine Rolle, aber gerade diesen Kompetenzbereich gilt es im Hinblick auf das Lernen von Wissen über (extra-)legale Wirkfaktoren zu bestärken. Insbesondere die Bedeutsamkeit von Feedback sollte bereits auf universitärer Ebene vermittelt werden, da das Studium und auch das Referendariat einen starken berufsvorbereitenden Charakter haben. Somit stellen sie geeignete Zeitfenster dar, um durch die Aneignung professionellen Handelns auf die bösartige Lernumgebung des Justizsystems adäquat reagieren zu können.

Die Laien waren (deskriptiv) überzeugter von der Beweislage als die Expert:innen und bejahten den Tatverdacht (H1 und H2; s. 4.2.2; 4.3.1). Allerdings lag das Maß der Überzeugung unterhalb der für das Strafrecht angenommenen 90%-Grenze (Schweizer, 2015; s. 5.2.6). Die Naiven nahmen nicht nur die Beweismittel dahingehend anders war, dass deren Würdigung zu einer konträren Entscheidung führte, sondern sie handelten dabei – ebenso wie die Vergleichsgruppen – nicht gemäß den Vorgaben. Naive haben vermutlich ein anderes Verständnis des Beweismaßes. Hinweise darauf, dass dies zumindest im Hauptverfahren vorkommt, bieten die Arbeiten von Baucum et al. (2018), Daftary-Kapur et al. (2010), Dhami et al. (2015) und Mueller-Johnson et al. (2018). Da Laien an der Rechtsprechung mitwirken, implizieren und bestärken die hier erzielten Ergebnisse die Notwendigkeit eines Austauschs zwischen Berufs- und Laienrichter:innen (s. auch Machura, 2016), auch

wenn sich Laien vielmehr mit der Schuldfrage als mit der Frage nach dem Tatverdacht befassen. Die Sinnhaftigkeit des Feedbacks und des Austausches lässt sich vom hier untersuchten Ermittlungsverfahren auf diese kooperative Entscheidungssituation zweier Expertise-Gruppen im Hauptverfahren übertragen. Zum Mehrwert vielseitiger Perspektiven passt die Annahme von Guthrie et al. (2001), dass diese zu guten Entscheidungen von Richter:innen beitragen können. Da Laien im Strafverfahren erst spät zum Einsatz kommen, kann sich deren korrigierende Funktion aber nicht direkt auf den ersten Verfahrensschritt der Ermittlungen auswirken, dessen Bedeutsamkeit wiederholt hervorgehoben wurde (s. 2.1.7; 2.3.5).

Es wurde zuvor mehrfach auf die Disparität in den Entscheidungen innerhalb der Gruppe der Expert:innen eingegangen.[57] Da genaue (qualitative) Begründungen für oder gegen das Vorliegen des Tatverdachtes nicht erhoben wurden, lässt sich nicht eindeutig zuordnen, ob diese Diskrepanz begründet (Inkonsistenz) oder unbegründet (Ungleichheit) zustande kam (s. auch Maguire, 2010). Die Expertise sowie der Delikttyp gelten laut dieser Studie als signifikante Einflussfaktoren (H1; s. 4.2.2). Allerdings waren die Effekte klein, weswegen es darüber hinaus Einflussfaktoren geben muss, die die Entscheidung erklären. Diese Studie konnte zeigen, dass die personenbedingten Faktoren „kognitive Reflexion“ und „Need for Cognition“ nahezu keinen bedeutsamen Beitrag zu dieser Erklärung leisten konnten, wobei sich zumindest die Reflexionsfähigkeit mit kleinem Effekt auf die Beweiswürdigung auswirkte (F4; s. 4.9.2). Ähnliches gilt größtenteils – aus den diskutierten Gründen (s. 5.2.2) – für den prozessbedingten Zeitdruck. Demzufolge gilt es, zusätzliche Perspektiven zu betrachten, um den Erkenntnisgewinn über juristische Entscheidungsfindung zu vergrößern.“When trying to understand the decision made by the court, we ideally want to reconstruct the decision-making process“ (Maegherman, 2021, S. 53). Doch laut Visher (1987) machen extralegale Faktoren nur einen kleinen Teil der Varianz in juristischen Entscheidungen aus (< 10%; s. auch Sommers et al., 2014). Dies wäre ein weiterer Grund, sich zunehmend mit der Beweiswürdigung und mit fallbezogenen Variablen auseinandersetzen, insbesondere zum Zeitpunkt des Ermittlungsverfahrens. Die hier erlangten Erkenntnisse deuten größten-

[57] Eine Interrater-Reliabilität als Maß für die Übereinstimmung zwischen Expert:innen wurde nicht berechnet (s. auch Shanteau, 2000).

teils eher auf eine zufällige und nicht auf eine systematische Variabilität des (Entscheidungs-)Verhaltens hin (s. auch Sporer & Goodman-Delahunty, 2009). Dazu kommen die überwiegend kleinen Effektstärken. Hinsichtlich des nächsten Verfahrensschrittes wären aber dennoch bestimmte personenbedingte Merkmale zu berücksichtigen – auch weil bereits argumentiert wurde, dass in dieser Entscheidungssituation vermutlich weniger fallbedingte Freiheitsgrade vorliegen als bei der Frage nach dem Tatverdacht. Somit könnten personenbedingte Einflüsse in solch „unklaren", freien Situationen stärker zum Tragen kommen als in Situationen, in denen die Beweislage eindeutiger ist. Zu den Merkmalen, die einen größeren Einfluss haben könnten als die hier untersuchten personenbedingten Faktoren, zählen die Strafeinstellungen sowie das Strafbedürfnis der Befragten (Rachlinski & Wistrich, 2017; Sporer & Goodman-Delahunty, 2009; Suhling et al., 2005).

Neben den ausgeführten Folgerungen und Forschungsansätzen gibt es daran ansetzend und weiterführend methodisch-inhaltliche Implikationen und Forschungsdesiderate, die über die in Abschnitt 5.3 genannten Limitationen und Verbesserungsideen hinausgehen. Die eingesetzten Vignetten dienten zur Präsentation der Beweislage eines hypothetischen Falles. Vignetten eignen sich aber auch für die Untersuchung von Urteilen, in denen der Einfluss von Dimensionen genauer betrachtet wird (Beck & Opp, 2001). In dieser Studie blieben diese Dimensionen unverändert. Für nachfolgende Studien wäre es demnach nicht nur aus methodischer, sondern insbesondere aus inhaltlicher Sicht hilfreich, das systematische Variieren von Dimensionen und ihrer Ausprägungen innerhalb der beiden Fallbeschreibungen einzubauen (Alexander & Becker, 1978; Auspurg, Hinz & Liebig, 2009; Beck & Opp, 2001). So zeigt sich, dass Merkmale der beschuldigten Person (z. B. Berufsstatus), der geschädigten Person (z. B. Geschlecht) oder des Verbrechens (z. B. Schwere der Tat) zu unterschiedlichen Reaktionen führen (Durham, 1986; Rossi et al., 1985). Demnach könnte die systematische Randomisierung dieser Merkmale weitere Erkenntnisse darüber liefern, welche dieser Aspekte bei Fällen mit nicht eindeutiger Beweislage zu welchen Entscheidungen führen – zumal die Bedeutung fallbezogener Variablen (z. B. Delikttyp) bereits angedeutet wurde. Zukünftige Forschung könnte sich dabei an den Kategorien der Beweismittel orientieren und je Kategorie verschiedene Ausprägungen einbauen (z. B. der Beschuldigte „macht eine Aussage" oder „macht keine Aussage"). Eher grundlegende Erkenntnisse gilt es aber

zunächst darüber einzuholen, ob Rechtsexpert:innen tatsächlich in den auszumachenden Beweiskategorien denken und welche alternative Operationalisierung sich anbietet (s. 5.2.8). Zudem wäre die Variation von Merkmalen der beschuldigten Person sinnvoll, um deren Einfluss auf eine nicht eindeutige Beweislage zu ermitteln (z. B. Vorstrafen, Geschlecht), da zumindest diese Fakten in der Regel unstrittig sind und gewissermaßen die einzigen *eindeutigen* Informationen in der Fallbeschreibung darstellen. Auch die Randomisierung der Beweismittel (Maegherman, 2021) oder das schrittweise Präsentation der Inhalte könnten sich auf die Würdigung auswirken, weil ein schrittweise Integrieren von Informationen dazu führen kann, einzelne Aspekte zu ignorieren (im Vergleich zur vollständigen Darbietung; Dummel et al., 2016).

Auch wenn die Effekte der Vorstrafen auf die Schuldfrage eher als klein zu bewerten sind (Schmittat, 2022), können diese insbesondere bei nicht eindeutiger Beweislage ausschlaggebend für eine Verurteilung sein (Pearson et al., 2018). Die Vorstrafen des Beschuldigten wurden in dieser Studie als Urkundenbeweis geliefert, da es sich beim Bundeszentralregisterauszug um eine solche Urkunde handelt. Da keine qualitative Befragung der Teilnehmenden über das „Warum?“ erfolgte, lässt sich nicht eindeutig feststellen, ob dieser Auszug hinsichtlich seiner Natur als Urkundenbeweis oder hinsichtlich des darin aufgeführten Inhalts („keine Vorstrafen“; Schmittat, 2022) gewertet wurde. Lässt sich die Art beziehungsweise die Kategorie des Beweismittels überhaupt von der Aussagekraft des Inhalts trennen? Weitere methodische Abwandlungen wären der vermehrte Einsatz von offenen Fragen, um genauere Rückmeldungen hinsichtlich der Fragestellungen zu erhalten (Bieneck, 2009; J. Finch, 1987). Ähnliches gilt für die Einschätzung anderer Beweismittel als (nicht) relevant, aus der bisher nicht abgeleitet werden kann, mit welcher Begründung diese zur Bejahung oder Verneinung des Tatverdachtes führte. Demnach könnte die Methode des lauten Denkens für die Bearbeitung der Fallbeispiele zu einem Erkenntnisgewinn führen. Auch das Ausmaß fachlicher Intuition (und somit der Einsatz von Dual-Prozessen) ließe sich mit dieser Methode ablesen (s. auch Mishra et al., 2015; Thompson, 2009). Eine qualitative Erfassung würde auch weiterhelfen, einschätzen zu können, inwiefern sich die eingesetzten Vignetten noch verbessern lassen. Die hier erfassten Angaben wiesen auf vergleichbare Reaktionen der Proband:innen hin, z. B. dass im „Diebstahl“-Fall zusätzliche Aussagen zu einer

Seriennummer gewünscht werden (s. 4.12.1). Demzufolge könnten die Vignetten inhaltlich angepasst werden, solange die nicht eindeutige Beweislage, die in beiden Fallbeispielen gelungen war (s. 5.2.6), aufrecht bleibt.

5.5 Fazit

Die Teilnehmenden trafen angesichts einer nicht eindeutigen Beweislage in einem Kriminalfall eine Entscheidung über das Vorliegen des Tatverdachts. Zudem wurden Variablen erhoben, die Aussagen über das Erleben des Entscheidungsprozesses und über den Umgang mit Beweismitteln zulassen. Das Hauptergebnis lautete, dass das Level der juristischen Expertise erwartungsgemäß zu Unterschieden (zwischen Laien und Expert:innen) führte. Die untersuchten prozess- und personenbedingten Faktoren bewirkten vor allem kleine und teilweise nicht bedeutsame Effekte, welche vor dem Hintergrund der empirischen Forschungslage diskutiert wurden. Mit Blick auf die juristische Ausbildung sowie die berufliche Praxis wurden Handlungsimpulse abgeleitet, die die Bösartigkeit des Justizsystems als Lernumgebung reduzieren könnten. Darüber hinaus wurden die Bereitschaft der fachlichen Teilstichproben zur Teilnahme an empirischen Studien sowie die Eignung neu konstruierter Fallvignetten deutlich.

Diese Studie begann einleitend mit der Frage, ob Fachpersonen besser als Laien dazu geeignet sind, juristische Urteile zu finden und Entscheidungen zu treffen. Es lässt sich schlussfolgern, dass Expert:innen nicht zwingend *besser* im Entscheiden sind – zumal im juristischen Kontext oft nicht die eine richtige Lösung vorliegt. Aber angesichts ungleicher Entscheidungen zwischen den (Quasi-)Experimentalgruppen wurden doch unterschiedliche Vorgehensweisen (insbesondere in der Beweiswürdigung) und das signifikante Wirken bestimmter Einflussfaktoren deutlich. Expert:innen, und teilweise auch Noviz:innen, entscheiden scheinbar *anders* als Laien. Obwohl sich Disparitäten sogar im Verhalten der Expert:innen zeigten, lässt sich eine wichtige Erkenntnis der Studie folgendermaßen zusammenfassen: „Some disparities in sentencing are the inevitable consequence of the fact that the decisions are performed by humans“ (Sporer & Goodman-Delahunty, 2009, S. 397).

Literatur

Ackermann, R., Clages, H. & Roll, H. (2022). *Handbuch der Kriminalistik: Kriminaltaktik für Praxis und Ausbildung* (6., aktualisierte Aufl.). Richard Boorberg Verlag. https://doi.org/10.5771/9783415069923

Agresti. (2007). *An introduction to categorical data analysis* (2. Aufl.). Wiley.

Alderden, M. A. & Ullman, S. E. (2012). Creating a more complete and current picture: examining police and prosecutor decision-making when processing sexual assault cases. *Violence against women*, *18*(5), 525–551. https://doi.org/10.1177/1077801212453867

Alexander, C. S. & Becker, H. J. (1978). The Use of Vignettes in Survey Research. *Public Opinion Quarterly*, *42*(1), 93–104. https://doi.org/10.1086/268432

Alison, L., Doran, B., Long, M. L., Power, N. & Humphrey, A. (2013). The effects of subjective time pressure and individual differences on hypotheses generation and action prioritization in police investigations. *Journal of Experimental Psychology: Applied*, *19*(1), 83–93. https://doi.org/10.1037/a0032148

Allison, M. & Brimacombe, C. A. E. (2010). Alibi Believability: The Effect of Prior Convictions and Judicial Instructions. *Journal of Applied Social Psychology*, *40*(5), 1054–1084. https://doi.org/10.1111/j.1559-1816.2010.00610.x

Althammer, C. & Tolani, M. (2019). Proof of Causation in German Tort Law. In L. Tichý (Hrsg.), *Veröffentlichungen zum Verfahrensrecht: Bd. 158. Standard of proof in Europe* (1. Aufl., Bd. 158, S. 109–121). Mohr Siebeck.

Anandarajan, A., Kleinman, G. & Palmon, D. (2008). Novice and expert judgment in the presence of going concern uncertainty. *Managerial Auditing Journal*, *23*(4), 345–366. https://doi.org/10.1108/02686900810864309

Anderson, J. M., Kling, J. R. & Stith, K. (1999). Measuring Interjudge Sentencing Disparity: Before and After the Federal Sentencing Guidelines. *The Journal of Law and Economics*, *42*(S1), 271–308. https://doi.org/10.1086/467426

Appelt, K. C., Milch, K. F., Handgraaf, M. J. J. & Weber, E. U. (2011). The Decision Making Individual Differences Inventory and guidelines for the study of individual differences in judgment and decision-making research. *Judgment and Decision Making*, *6*(3), 252–262.

Ask, K. & Granhag, P. A. (2005). Motivational sources of confirmation bias in criminal investigations: the need for cognitive closure. *Journal of Investigative Psychology and Offender Profiling*, *2*(1), 43–63. https://doi.org/10.1002/jip.19

Auspurg, K., Hinz, T. & Liebig, S. (2009). Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey. *Methoden - Daten - Analysen*, *3*(1), 59–96.

Auspurg, K., Hinz, T., Liebig, S. & Sauer, C. (2009). Auf das Design kommt es an. Experimentelle Befunde zu komplexen Settings in Faktoriellen Surveys. *Sozialwissenschaftlicher Fachinformationsdienst soFid*(2009/2), 23–39. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-205089

Austin, W. & Williams, T. A. (1977). A Survey of Judges' Responses to Simulated Legal Cases: Research Note on Sentencing Disparity. *The Journal of Criminal Law and Criminology*, *68*(2), 306–310.

Baber, C. & Butler, M. (2012). Expertise in crime scene examination: comparing search strategies of expert and novice crime scene examiners in simulated crime scenes. *Human factors*, *54*(3), 413–424. https://doi.org/10.1177/0018720812440577

Bago, B. & Neys, W. de (2017). Fast logic? Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B. & Neys, W. de (2019). The Smart System 1: evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Baucum, M., Scurich, N. & John, R. S. (2018). Lay judgements of the probable cause standard. *Law, Probability and Risk*, *17*(3), 225–242. https://doi.org/10.1093/lpr/mgy010

Beck, M. & Opp, K.-D. (2001). Der faktorielle Survey und die Messung von Normen. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *53*(2), 283–306. https://doi.org/10.1007/s11577-001-0040-3

Beilock, S. L., Bertenthal, B. I., Hoerger, M. & Carr, T. H. (2008). When does haste make waste? Speed-accuracy tradeoff, skill level, and the tools of the trade. *Journal of Experimental Psychology: Applied*, *14*(4), 340–352. https://doi.org/10.1037/a0012859

Beilock, S. L., Bertenthal, B. I., McCoy, A. M. & Carr, T. H. (2004). Haste does not always make waste: expertise, direction of attention, and speed versus accuracy in performing sensorimotor skills. *Psychonomic Bulletin & Review*, *11*(2), 373–379. https://doi.org/10.3758/BF03196585

Beißert, H., Köhler, M., Rempel, M. & Beierlein, C. (2014). *Deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition NFC-K: Die Need for Cognition Kurzskala (NFC-K)* (GESIS – Working Papers Nr. 32). Mannheim. https://www.gesis.org/fileadmin/kurzskalen/working_papers/WorkingPapers_2014-32.pdf

Betsch, T., Brinkmann, B. J., Fiedler, K. & Breining, K. (1999). When prior knowledge overrules new evidence: Adaptive use of decision strategies and the role of behavioral routines. *Swiss Journal of Psychology*, *58*(3), 151–160. https://doi.org/10.1024//1421-0185.58.3.151

Betsch, T., Funke, J. & Plessner, H. (2011). *Allgemeine Psychologie für Bachelor: Denken - Urteilen, Entscheiden, Problemlösen.* (1. Aufl.). Springer.

Bialek, M. & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, *50*(5), 1953–1959. https://doi.org/10.3758/s13428-017-0963-x

Bieneck, S. (2006). *Soziale Informationsverarbeitung in der juristischen Urteilsfindung: Experimentelle Untersuchungen zur Ankerheuristik* [Dissertation]. Universität Potsdam, Potsdam.

Bieneck, S. (2009). How adequate is the vignette technique as a research tool for psycho-legal research? In M. E. Oswald, S. Bieneck & J. Hupfeld-Heinemann (Hrsg.), *Social Psychology of Punishment of Crime* (S. 255–271). Wiley.

Blacksmith, N., Yang, Y., Behrend, T. S. & Ruark, G. A. (2019). Assessing the validity of inferences from scores on the cognitive reflection test. *Journal of Behavioral Decision Making*, *32*(5), 599–612. https://doi.org/10.1002/bdm.2133

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F. & Schwarz, N. (1994). Need for Cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben. *Zeitschrift für Sozialpsychologie*(25), 147–154.

Bobadilla-Suarez, S. & Love, B. C. (2018). Fast or frugal, but not both: Decision heuristics under time pressure. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(1), 24–33. https://doi.org/10.1037/xlm0000419

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7., vollständig überarbeitete und erweiterte Aufl.). Springer.

Bouckenooghe, D., Vanderheyden, K., Mestdagh, S. & van Laethem, S. (2007). Cognitive motivation correlates of coping style in decisional conflict. *The Journal of Psychology*, *141*(6), 605–625. https://doi.org/10.3200/JRLP.141.6.605-626

Brams, S., Ziv, G., Levin, O., Spitz, J., Wagemans, J., Williams, A. M. & Helsen, W. F. (2019). The relationship between gaze behavior, expertise, and performance: A systematic review. *Psychological Bulletin*, *145*(10), 980–1027. https://doi.org/10.1037/bul0000207

Brañas-Garza, P., Kujal, P. & Lenkei, B. (2015). *Cognitive Reflection Test: Whom, How, When* (ESI Working Papers 15-25). https://digitalcommons.chapman.edu/esi_working_papers/174/

Büchner, S. (2022). Die Rolle der Staatsanwaltschaft im Ermittlungsverfahren und Strafprozess. *Richter ohne Robe*, *34*(3), 88–89.

Buelow, M. T., Jungers, M. K. & Chadwick, K. R. (2019). Manipulating the decision making process: Influencing a "gut" reaction. *Journal of Clinical and Experimental Neuropsychology*, *41*(10), 1097–1113. https://doi.org/10.1080/13803395.2019.1662374

Bullens, L., van Harreveld, F., Förster, J. & Higgins, T. E. (2014). How decision reversibility affects motivation. *Journal of Experimental Psychology: General*, *143*(2), 835–849. https://doi.org/10.1037/a0033581

Byrne, A. (2013). Mental workload as a key factor in clinical decision making. *Advances in Health Sciences Education: Theory and Practice*, *18*(3), 537–545. https://doi.org/10.1007/s10459-012-9360-5

Cacioppo, J. T. & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., Feinstein, J. A. & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253. https://doi.org/10.1037/0033-2909.119.2.197

Campitelli, G. & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, *42*(3), 434–447. https://doi.org/10.3758/s13421-013-0367-9

Carnevale, J. J., Inbar, Y. & Lerner, J. S. (2011). Individual differences in need for cognition and decision-making competence among leaders. *Personality and Individual Differences*, *51*(3), 274–278. https://doi.org/10.1016/j.paid.2010.07.002

Čavojová, V. & Hanák, R. (2014). How much information do you need? Interaction of intuitive processing with expertise. *Studia Psychologica*, *56*(2), 83–97.

Chambers, K. L. & Zaragoza, M. S. (2001). Intended and unintended effects of explicit warnings on eyewitness suggestibility: evidence from source identification tests. *Memory & Cognition*, *29*(8), 1120–1129. https://doi.org/10.3758/BF03206381

Chi, M. T. H. (2006). Two Approaches to the Study of Experts' Characteristics. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman (Hrsg.), *The Cambridge handbook of expertise and expert performance* (S. 21–30). Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Lawrence Erlbaum Associates.

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Collins, P. M., Manning, K. & Carp, R. (2010). Gender, Critical Mass, and Judicial Decision Making. *Law & Policy*, *32*(2), 260–281. https://doi.org/10.1111/j.1467-9930.2010.00317.x

Combé, D. (2007). *Stellung und Objektivität der Staatsanwaltschaft im Ermittlungsverfahren* (1st ed.). *Reihen des Cuvillier-Verlages - Rechtswissenschaften: v. 5*. Cuvillier Verlag.

Conklin, M. (2020). Reasonable Doubt Ratcheting: How Jurors Adjust the Standard of Proof to Reach a Desired Result. *North Dakota Law Review*, *95*(2), 281–290.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. *Environmental ethics and science policy series*. Oxford University Press.

Coutinho, M. V. C., Thomas, J., Alsuwaidi, A. S. M. & Couchman, J. J. (2021). Dunning-Kruger Effect: Intuitive Errors Predict Overconfidence on the Cognitive Reflection Test. *Frontiers in Psychology*, *12.* https://doi.org/10.3389/fpsyg.2021.603225

Croskerry, P., Petrie, D. A., Reilly, J. B. & Tait, G. (2014). Deciding about fast and slow decisions. *Academic Medicine: Journal of the Association of*

American Medical Colleges, *89*(2), 197–200. https://doi.org/10.1097/ACM.0000000000000121

Curley, L. J., Murray, J., MacLean, R., Munro, J., Lages, M., Frumkin, L. A., Laybourn, P. & Brown, D. (2022). Verdict spotting: investigating the effects of juror bias, evidence anchors and verdict system in jurors. *Psychiatry, Psychology, and Law: An Interdisciplinary Journal of the Australian and New Zealand Association of Psychiatry, Psychology and Law*, *29*(3), 323–344. https://doi.org/10.1080/13218719.2021.1904450

Curşeu, P. L. (2006). Need for Cognition and Rationality in Decision-Making. *Studia Psychologica*, *48*(2), 141–156.

Daftary-Kapur, T., Dumas, R. & Penrod, S. D. (2010). Jury decision-making biases and methods to counter them. *Legal and Criminological Psychology*, *15*(1), 133–154. https://doi.org/10.1348/135532509X465624

Dambacher, M. & Hübner, R. (2015). Time pressure affects the efficiency of perceptual processing in decisions under conflict. *Psychological Research*, *79*(1), 83–94. https://doi.org/10.1007/s00426-014-0542-z

Dando, C. J. & Ormerod, T. C. (2017). Analyzing Decision Logs to Understand Decision Making in Serious Crime Investigations. *Human factors*, *59*(8), 1188–1203. https://doi.org/10.1177/0018720817727899

Danziger, S., Levav, J. & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(17), 6889–6892. https://doi.org/10.1073/pnas.1018033108

Davies, M. (2004). Sentencing Burglars and Explaining the Differences Between Jurisdictions: Implications for Convergence. *British Journal of Criminology*, *44*(5), 741–758. https://doi.org/10.1093/bjc/azh035

Dhami, M. K. (2003). Psychological models of professional decision making. *Psychological Science*, *14*(2), 175–180. https://doi.org/10.1111/1467-9280.01438

Dhami, M. K. & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making*, *14*(2), 141–168. https://doi.org/10.1002/bdm.371

Dhami, M. K., Lundrigan, S. & Mueller-Johnson, K. (2015). Instructions on reasonable doubt: Defining the standard of proof and the juror's task. *Psychology, Public Policy, and Law*, *21*(2), 169–178. https://doi.org/10.1037/law0000038

Dickert, S., Herbig, B., Glöckner, A., Gansen, C. & Portack, R. (2012). The More the Better? Effects of Training, Experience and Information Amount in Legal Judgments. *Applied Cognitive Psychology*, *26*(2), 223–233. https://doi.org/10.1002/acp.1813

Diederich, A. & Trueblood, J. S. (2018). A dynamic dual process model of risky decision making. *Psychological Review*, *125*(2), 270–292. https://doi.org/10.1037/rev0000087

Donkin, C., Little, D. R. & Houpt, J. W. (2014). Assessing the speed-accuracy trade-off effect on the capacity of information processing. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(3), 1183–1202. https://doi.org/10.1037/a0035947

Dror, I. E. (2011). The paradox of human expertise: why experts get it wrong. In N. Kapur, A. Pascual-Leone, V. Ramachandran, J. Cole, S. Della Sala, T. Manly, A. Mayes & O. Sacks (Hrsg.), *The Paradoxical Brain* (S. 177–188). Cambridge University Press.

Dror, I. E. (2020). Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias. *Analytical Chemistry*, *92*(12), 7998–8004. https://doi.org/10.1021/acs.analchem.0c00704

Dummel, S., Rummel, J. & Voss, A. (2016). Additional information is not ignored: New evidence for information integration and inhibition in take-the-best decisions. *Acta psychologica*, *163*, 167–184. https://doi.org/10.1016/j.actpsy.2015.12.001

Durham, A. M., III (1986). The Use of Factorial Survey Design in Assessments of Public Judgments of Appropriate Punishment for Crime. *Journal of Quantitative Criminology*, *2*(2), 181–190.

Ebbesen, E. B. & Konečni, V. J. (1982). Social Psychology and the Law: A Decision-Making Approach to the Criminal Justice System. In V. J. Konečni & E. B. Ebbesen (Hrsg.), *The Criminal Justice System: A Social-Psychological Analysis* (S. 3–23). W. H. Freeman and Company.

Edland, A. & Svenson, O. (1993). Judgment and Decision Making Under Time Pressure. In A. J. Maule & O. Svenson (Hrsg.), *Time pressure and stress in human decision making* (S. 27–40). Plenum.

Effer-Uhe, D. & Mohnert, A. (2019). *Psychologie für Juristen* (1. Aufl.). Nomos.

Egli Anthonioz, N., Schweizer, M., Vuille, J. & Kuhn, A. (2019). Role-induced bias in criminal prosecutions. *European Journal of Criminology*, *16*(4), 452–465. https://doi.org/10.1177/1477370818772772

Ehrlinger, J., Gilovich, T. & Ross, L. (2005). Peering into the bias blind spot: people's assessments of bias in themselves and others. *Personality & Social Psychology Bulletin*, *31*(5), 680–692. https://doi.org/10.1177/0146167204271570

Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5., korrigierte Aufl.). Beltz.

Eifler, S. & Bentrup, C. (2003). *Zur Validität von Selbstberichten abweichenden und hilfreichen Verhaltens mit der Vignettenanalyse* (Bielefelder Arbeiten zur Sozialpsychologie Nr. 208). Bielefeld. https://pub.uni-bielefeld.de/download/1862507/2603130/208_Eifler_Validitat.pdf

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance*, *6*(1), 1–27. https://doi.org/10.1016/0030-5073(71)90002-X

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Ellison, J. M. & Brennan, P. K. (2016). Sentencing Outcomes and Disparity. In B. M. Huebner & T. S. Bynum (Hrsg.), *Handbook of Measurement Issues in Criminology and Criminal Justice* (S. 328–349). Wiley-Blackwell.

Engel, C. (2017). *Empirical Methods for the Law* (2017/07). Bonn. Max Planck Institute for Research on Collective Goods. https://doi.org/10.2139/ssrn.2966095

Engel, C. (2020). Uncertain Judges. *Journal of Institutional and Theoretical Economics*, *176*(1), 44. https://doi.org/10.1628/jite-2020-0007

Engel, C. & Gigerenzer, G. (2006). Law and Heuristics: An Interdisciplinary Venture. In C. Engel & G. Gigerenzer (Hrsg.), *Dahlem workshop reports. Heuristics and the law* (S. 1–16). MIT Press in cooperation with Dahlem University Press.

Engel, C. & Glöckner, A. (2013). Role-Induced Bias in Court: An Experimental Analysis. *Journal of Behavioral Decision Making*, *26*(3), 272–284. https://doi.org/10.1002/bdm.1761

Engel, C., Timme, S. & Glöckner, A. (2020). Coherence-based reasoning and order effects in legal judgments. *Psychology, Public Policy, and Law*, *26*(3), 333–352. https://doi.org/10.1037/law0000257

Englich, B. (2009). Heuristic strategies and persistent biases in sentencing decisions. In M. E. Oswald, S. Bieneck & J. Hupfeld-Heinemann (Hrsg.), *Social Psychology of Punishment of Crime* (S. 295–314). Wiley.

Englich, B., Mussweiler, T. & Strack, F. (2006). Playing dice with criminal sentences: the influence of irrelevant anchors on experts' judicial decision making. *Personality & Social Psychology Bulletin*, *32*(2), 188–200. https://doi.org/10.1177/0146167205282152

Ettenson, R., Shanteau, J. & Krogstad, J. (1987). Expert Judgment: Is More Information Better? *Psychological Reports*, *60*(1), 227–238. https://doi.org/10.2466/pr0.1987.60.1.227

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T. (2010). Intuition and Reasoning: A Dual-Process Perspective. *Psychological Inquiry*, *21*(4), 313–326. https://doi.org/10.1080/1047840X.2010.521057

Festinger, L. (1957). *A theory of cognitive dissonance*. Row Peterson.

Field, A. (2009). *Discovering statistics using SPSS: (and sex and drugs and rock n' roll)* (3. Aufl.). Sage Publications.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4. Aufl.). Sage Publications.

Finch, H. (2005). Comparison of the Performance of Nonparametric and Parametric MANOVA Test Statistics when Assumptions Are Violated. *Methodology*, *1*(1), 27–38. https://doi.org/10.1027/1614-1881.1.1.27

Finch, J. (1987). The Vignette Technique in Survey Research. *Sociology*, *21*(1), 105–114. https://doi.org/10.1177/0038038587021001008

Findley, K. & Scott, M. S. (2006). The Multiple Dimensions of Tunnel Vision in Criminal Cases. *Wisconsin Law Review*(2), 291–397.

Fleischhauer, M., Enge, S., Brocke, B., Ullrich, J., Strobel, A [Alexander] & Strobel, A [Anja] (2010). Same or different? Clarifying the relationship of need for cognition to personality and intelligence. *Personality & Social Psychology Bulletin*, *36*(1), 82–96. https://doi.org/10.1177/0146167209351886

Fleming, M. A., Wegener, D. T. & Petty, R. E. (1999). Procedural and Legal Motivations to Correct for Perceived Judicial Biases. *Journal of Experimental Social Psychology*, *35*(2), 186–203. https://doi.org/10.1006/jesp.1998.1375

Forsterlee, R., Forsterlee, L., Horowitz, I. A. & King, E. (2006). The effects of defendant race, victim race, and juror gender on evidence processing in a murder trial. *Behavioral Sciences and the Law*, *24*(2), 179–198. https://doi.org/10.1002/bsl.675

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Friedel, E., Sebold, M., Kuitunen-Paul, S., Nebe, S., Veer, I. M., Zimmermann, U. S., Schlagenhauf, F., Smolka, M. N., Rapp, M., Walter, H. & Heinz, A. (2017). How Accumulated Real Life Stress Experience and Cognitive Speed Interact on Decision-Making Processes. *Frontiers in Human Neuroscience*, *11*, 302. https://doi.org/10.3389/fnhum.2017.00302

Fritz, C. O., Morris, P. E. & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18. https://doi.org/10.1037/a0024338

Gabriel, U. (2009). Emotions and legal judgments: Normative issues vs. empirical findings. In M. E. Oswald, S. Bieneck & J. Hupfeld-Heinemann (Hrsg.), *Social Psychology of Punishment of Crime* (S. 157–172). Wiley.

Gaissmaier, W., Fifić, M. & Rieskamp, J. (2011). Analyzing response times to understand decision processes. In M. Schulte-Mecklenbeck, A. Kuehberger & J. G. Johnson (Hrsg.), *A Handbook of Process Tracing Methods for Decision Research* (S. 141–162). Psychology Press.

Garcia-Retamero, R. & Dhami, M. K. (2009). Take-the-best in expert-novice decision strategies for residential burglary. *Psychonomic Bulletin & Review*, *16*(1), 163–169. https://doi.org/10.3758/PBR.16.1.163

Gardner, B. O., Kelley, S., Murrie, D. C. & Dror, I. E. (2019). What do forensic analysts consider relevant to their decision making? *Science & Justice: Journal of the Forensic Science Society*, *59*(5), 516–523. https://doi.org/10.1016/j.scijus.2019.04.005

Gigerenzer, G. (2006). Heuristics. In G. Gigerenzer & C. Engel (Hrsg.), *Heuristics and the law* (S. 17–44). MIT Press in cooperation with Dahlem University Press.

Gigerenzer, G., Todd, P. M. & The ABC Research Group (Hrsg.). (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press.

Glöckner, A. (2008a). *Neurorecht ohne Psychologie? Die Rolle verhaltenswissenschaftlicher Betrachtungsebenen bei der Ableitung rechtspolitischer Empfehlungen* (2008, 18). Bonn. Max Planck Institute for Research on Collective Goods. https://www.econstor.eu/handle/10419/26955

Glöckner, A. (2008b). Zur Rolle intuitiver und bewusster Prozesse bei rechtlichen Entscheidungen. In Max-Planck-Gesellschaft (Hrsg.), *Max-Planck-Jahrbuch.* Max-Planck-Gesellschaft.

Glöckner, A. & Betsch, T. (2008a). *Modeling Option and Strategy Choices with Connectionist Networks: Towards an Integrative Model of Automatic and Deliberate Decision Making* (2008, 2). Bonn. Max Planck Institute for Research on Collective Goods. https://www.econstor.eu/handle/10419/26939

Glöckner, A. & Betsch, T. (2008b). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(5), 1055–1075. https://doi.org/10.1037/0278-7393.34.5.1055

Glöckner, A. & Betsch, T. (2012). Decisions beyond boundaries: when more information is processed faster than less. *Acta psychologica*, *139*(3), 532–542. https://doi.org/10.1016/j.actpsy.2012.01.009

Glöckner, A., Betsch, T. & Schindler, N. (2010). Coherence shifts in probabilistic inference tasks. *Journal of Behavioral Decision Making*, *23*(5), 439–462. https://doi.org/10.1002/bdm.668

Glöckner, A. & Ebert, I. D. (2011). Legal Intuition and Expertise. In M. Sinclair (Hrsg.), *Handbook of Intuition Research* (S. 157–167). Edward Elgar Publishing.

Glöckner, A. & Engel, C. (2013). Can We Trust Intuitive Jurors? Standards of Proof and the Probative Value of Evidence in Coherence-Based Reasoning. *Journal of Empirical Legal Studies*, *10*(2), 230–252.

Glöckner, A., Towfigh, E. & Traxler, C. (2013). Development of legal expertise. *Instructional Science*, *41*(6), 989–1007. https://doi.org/10.1007/s11251-013-9266-5

Glöckner, A. & Witteman, C. (2010). Beyond dual-process models: A categorisation of processes underlying intuitive judgement and decision making. *Thinking & Reasoning*, *16*(1), 1–25. https://doi.org/10.1080/13546780903395748

Goddard, R. D. & Villanova, P. (2006). Designing Surveys and Questionnaires for Research. In F. T. L. Leong & J. T. Austin (Hrsg.), *The Psychology Research Handbook: A Guide for Graduate Students and Research Assistants* (S. 114–124). Sage Publications. https://doi.org/10.4135/9781412976626.n8

Gonzalez, C. (2004). Learning to make decisions in dynamic environments: effects of time constraints and cognitive abilities. *Human factors*, *46*(3), 449–460. https://doi.org/10.1518/hfes.46.3.449.50395

Greene, E. & Dodge, M. (1995). The Influence of Prior Record Juror Decision Making. *Law and Human Behavior*(19), 67–78.

Groß, J. & Börensen, C. (2009). Wie valide sind Verhaltensmessungen mittels Vignetten? Ein methodischer Vergleich von faktoriellem Survey und Verhaltensbeobachtung. In P. Kriwy & C. Gross (Hrsg.), *Klein aber fein!* (S. 149–178). VS Verlag für Sozialwissenschaften.

Guo, L., Trueblood, J. S. & Diederich, A. (2017). Thinking Fast Increases Framing Effects in Risky Decision Making. *Psychological Science*, *28*(4), 530–543. https://doi.org/10.1177/0956797616689092

Gusy, C. (2015). Zukunft der Richtervorbehalte. In S. Barton, R. Köbel & M. Lindemann (Hrsg.), *Interdisziplinäre Studien zu Recht und Staat: Bd. 54. Wider die wildwüchsige Entwicklung des Ermittlungsverfahrens* (1. Aufl., S. 195–217). Nomos.

Guthrie, C. P., Rachlinski, J. J. & Wistrich, A. J. (2001). Inside the Judicial Mind. *Cornell Law Review*, *86*, 777–830. https://doi.org/10.2139/ssrn.257634

Guthrie, C. P., Rachlinski, J. J. & Wistrich, A. J. (2007). Blinking on the Bench: How Judges Decide Cases. *Cornell Law Review*, *93*(1), 1–43.

Haigh, M. (2016). Has the Standard Cognitive Reflection Test Become a Victim of Its Own Success? *Advances in cognitive psychology*, *12*(3), 145–149. https://doi.org/10.5709/acp-0193-5

Harman, J. L. (2011). Individual differences in need for cognition and decision making in the Iowa Gambling Task. *Personality and Individual Differences*, *51*(2), 112–116. https://doi.org/10.1016/j.paid.2011.03.021

Henderson, K. S. & Levett, L. M. (2020). The effects of variations in confession evidence and need for cognition on jurors' decisions. *Psychology, Public Policy, and Law*, *26*(3), 245–260. https://doi.org/10.1037/law0000233

Herbig, B. & Glöckner, A. (2009). *Experts and Decision Making: First Steps towards a Unifying Theory of Decision Making in Novices, Intermediates and Experts* (2009, 2). Bonn. Max Planck Institute for Research on Collective Goods. https://doi.org/10.2139/ssrn.1337449

Hermann, D. (2009). Soziologie und Psychologie des Strafverfahrens. In H.-L. Kröber (Hrsg.), *Handbuch der forensischen Psychiatrie: Kriminologie und forensische Psychiatrie* (S. 645–688). Steinkopff.

Hertwig, R. & Todd, P. M. (2003). More Is Not Always Better: The Benefits of Cognitive Limits. In D. Hardman & L. Macchi (Hrsg.), *Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making* (S. 213–231). John Wiley & Sons, Ltd.

Hick, W. E. (1952). On the Rate of Gain of Information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26. https://doi.org/10.1080/17470215208416600

Hilbig, B. E., Erdfelder, E. & Pohl, R. F. (2012). A matter of time: antecedents of one-reason decision making based on recognition. *Acta psychologica*, *141*(1), 9–16. https://doi.org/10.1016/j.actpsy.2012.05.006

Hill, B. D., Foster, J. D., Elliott, E. M., Shelton, J. T., McCain, J. & Gouvier, W. D. (2013). Need for cognition is related to higher general intelligence, fluid intelligence, and crystallized intelligence, but not working memory. *Journal of Research in Personality*, *47*(1), 22–25. https://doi.org/10.1016/j.jrp.2012.11.001

Hoffmann, J. A., Gaissmaier, W. & Helversen, B. von (2017). Justifying the judgment process affects neither judgment accuracy, nor strategy use. *Judgment and Decision Making*, *12*(6), 627–641.

Hogarth, R. M. (2010). Intuition: A Challenge for Psychological Research on Decision Making. *Psychological Inquiry*, *21*(4), 338–353. https://doi.org/10.1080/1047840X.2010.520260

Holländer, P. (2019). Proof and Changing Idea of Truth in Legal Thinking: Reflection on Postmodernism. In L. Tichý (Hrsg.), *Veröffentlichungen zum Verfahrensrecht: Bd. 158. Standard of proof in Europe* (1. Aufl., Bd. 158, S. 3–18). Mohr Siebeck.

Holleran, D., Beichner, D. & Spohn, C. (2010). Examining Charging Agreement Between Police and Prosecutors in Rape Cases. *Crime & Delinquency*, *56*(3), 385–413. https://doi.org/10.1177/0011128707308977

Hope, L., Greene, E., Memon, A., Gavisk, M. & Houston, K. (2008). A third verdict option: exploring the impact of the not proven verdict on mock juror decision making. *Law and Human Behavior*, *32*(3), 241–252. https://doi.org/10.1007/s10979-007-9106-8

Huber, O. & Seiser, G. (2001). Accounting and convincing: the effect of two types of justification on the decision process. *Journal of Behavioral Decision Making*, *14*(1), 69–85. https://doi.org/10.1002/1099-0771(200101)14:1<69::AID-BDM366>3.0.CO;2-T

Hupfeld-Heinemann, J. & Helversen, B. von. (2009). *Models of decision-making on guilt and sanctions*. Wiley-Blackwell.

Hutton, R. J. B. & Klein, G. (1999). Expert decision making. *Systems Engineering*, *2*(1), 32–45. https://doi.org/10.1002/(SICI)1520-6858(1999)2:1<32::AID-SYS3>3.0.CO;2-P

Jahn, M. (2015). Das heutige strafprozessuale Ermittlungsverfahren aus Sicht von Wissenschaft und Justiz: Die Entwicklung in den letzten drei Jahrzehnten und die rechtspolitischen Baustellen. In S. Barton, R. Köbel & M. Lindemann (Hrsg.), *Interdisziplinäre Studien zu Recht und Staat: Bd. 54. Wider die wildwüchsige Entwicklung des Ermittlungsverfahrens* (1. Aufl., Bd. 54, S. 35–91). Nomos.

Jehle, J.-M. (2019). *Strafrechtspflege in Deutschland: Fakten und Zahlen*. Mönchengladbach. Bundesministerium der Justiz und für Verbraucherschutz. https://www.bmj.de/SharedDocs/Downloads/DE/Service/Fachpublikationen/Strafrechtspflege_Deutschland.pdf?__blob=publicationFile&v=15

Johnson, B. D. & Stewart, C. D. (2016). Measurement Issues in Criminal Case-Processing and Court Decision-Making Research. In B. M. Huebner & T. S. Bynum (Hrsg.), *Handbook of Measurement Issues in Criminology and Criminal Justice* (S. 303–327). Wiley-Blackwell.

Johnson, E. J., Payne, J. W., Bettman & James R. (1993). Adapting to Time Constraints. In A. J. Maule & O. Svenson (Hrsg.), *Time pressure and stress in human decision making* (S. 103–116). Plenum.

Joslyn, S. & Hunt, E. (1998). Evaluating individual differences in response to time-pressure situations. *Journal of Experimental Psychology: Applied*, *4*(1), 16–43. https://doi.org/10.1037/1076-898X.4.1.16

Juanchich, M., Dewberry, C., Sirota, M. & Narendran, S. (2016). Cognitive Reflection Predicts Real-Life Decision Outcomes, but Not Over and Above Personality and Decision-Making Styles. *Journal of Behavioral Decision Making*, *29*(1), 52–59. https://doi.org/10.1002/bdm.1875

Juanchich, M., Sirota, M. & Bonnefon, J.-F. (2020). Anxiety-induced miscalculations, more than differential inhibition of intuition, explain the gender gap in cognitive reflection. *Journal of Behavioral Decision Making*, *33*(4), 427–443. https://doi.org/10.1002/bdm.2165

Kagehiro, D. K. & Stanton, W. C. (1985). Legal vs. Quantified Standards of Proof. *Law and Human Behavior*, *9*(2), 159–178.

Kahan, D., Hoffman, D., Evans, D., Devins, N., Lucci, E. & Cheng, K. (2016). "Ideology" or "Situation Sense"? An Experimental Investigation of Motivated Reasoning and Professional Judgment. *University of Pennsylvania Law Review*, *164*(2), 349–439.

Kahneman, D. (2011). *Schnelles Denken, langsames Denken*. Siedler.

Kahneman, D. & Frederick, S. (2005). A Model of Heuristic Judgment. In R. G. Morrison & K. J. Holyoak (Hrsg.), *The Cambridge handbook of thinking and reasoning* (1. Aufl., S. 267–293). Cambridge University Press.

Kahneman, D. & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *The American psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kang, J., Bennett, M., Carbado, D. & Casey, P. (2012). Implicit Bias in the Courtroom. *UCLA law review*, *59*(5), 1124–1187.

Kassin, S. M., Dror, I. E. & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, *2*(1), 42–52. https://doi.org/10.1016/j.jarmac.2013.01.001

Kautt, P. (2009). Heuristic influences over offense seriousness calculations. *Punishment & Society*, *11*(2), 191–218. https://doi.org/10.1177/1462474508101492

Kern, C. (2019). Probability as an element of the standard of proof. In L. Tichý (Hrsg.), *Veröffentlichungen zum Verfahrensrecht: Bd. 158. Standard of proof in Europe* (1. Aufl., Bd. 158, S. 51–64). Mohr Siebeck.

Klein, G., Shneiderman, B., Hoffman, R. R. & Ford, K. M. (2017). Why Expertise Matters: A Response to the Challenges. *IEEE Intelligent Systems*, *32*(6), 67–73. https://doi.org/10.1109/MIS.2017.4531230

Koch, C. M. & Devine, D. J. (1999). Effects of Reasonable Doubt Definition and Inclusion of a Lesser Charge on Jury Verdicts. *Law and Human Behavior*, *23*(6), 653–674.

Konečni, V. J. & Ebbesen, E. B. (1979). External validity of research in legal psychology. *Law and Human Behavior*, *3*(1-2), 39–70. https://doi.org/10.1007/BF01039148

Konečni, V. J. & Ebbesen, E. B. (1982). Social Psychology and the Law: The Choice of Research Problems, Settings, and Methodology. In V. J. Konečni & E. B. Ebbesen (Hrsg.), *The Criminal Justice System: A Social-Psychological Analysis* (S. 27–44). W. H. Freeman and Company.

Kramer, J. H. & Ulmer, J. T. (1996). Sentencing disparity and departures from guidelines. *Justice Quarterly*, *13*(1), 81–106. https://doi.org/10.1080/07418829600092831

Krey, V. & Heinrich, M. (2018). *Deutsches Strafverfahrensrecht: Studienbuch in systematisch-induktiver Darstellung* (2. Aufl.). Kohlhammer Verlag.

Lakens, D. & Caldwell, A. R. (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920951503

Landsman, S. & Rakos, R. F. (1994). A Preliminary Inquiry into the Effect of Potentially Biasing Information on Judges and Jurors in Civil Litigation. *Behavioral Sciences and the Law*(12), 113–126.

Langeheine, R., Pannekoek, J. & van de Pol, F. (1996). Bootstrapping Goodness-of-Fit Measures in Categorical Data Analysis. *Sociological Methods & Research*, *24*(4), 492–516. https://doi.org/10.1177/0049124196024004004

Lavie, S., Ganor, T. & Feldman, Y. (2020). Adjusting legal standards. *European Journal of Law and Economics*, *49*(1), 33–53. https://doi.org/10.1007/s10657-018-9597-4

Leippe, M. R., Eisenstadt, D., Rauch, S. M. & Seib, H. M. (2004). Timing of eyewitness expert testimony, jurors' need for cognition, and case strength as determinants of trial verdicts. *The Journal of Applied Psychology*, *89*(3), 524–541. https://doi.org/10.1037/0021-9010.89.3.524

Lenhard, W. & Lenhard, A. (2017). *Computation of Effect Sizes.* https://doi.org/10.13140/RG.2.2.17823.92329

Leonhart, R. (2017). *Lehrbuch Statistik: Einstieg und Vertiefung* (4., überarbeitete und erweiterte Aufl.). Hogrefe.

Lerner, J. S. & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255

Levine, T. R. (2014). Truth-Default Theory (TDT). *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M. & Pardo, S. T. (2012). Individual Differences in Numeracy and Cognitive Reflection, with Implications for Biases and Fallacies in Probability Judgment. *Journal of Behavioral Decision Making*, *25*(4), 361–381. https://doi.org/10.1002/bdm.752

Lidén, M., Gräns, M. & Juslin, P. (2019). From devil's advocate to crime fighter: confirmation bias and debiasing techniques in prosecutorial decision-making. *Psychology, Crime & Law*, *25*(5), 494–526. https://doi.org/10.1080/1068316X.2018.1538417

Lieber, H. & Sens, U. (2019a). *Fit fürs Schöffenamt. Handbuch für ehrenamtliche Richterinnen und Richter in der Strafgerichtsbarkeit: Band 2: Das Strafverfahren – Grundlagen, Beweisaufnahme, Strafen* (2. Aufl.). Berliner Wissenschafts-Verlag.

Lieber, H. & Sens, U. (2019b). *Fit fürs Schöffenamt: Handbuch für ehrenamtliche Richterinnen und Richter in der Strafgerichtsbarkeit: Band 1: Rechte, Pflichten und Gestaltungsmöglichkeiten im Schöffenamt* (2. Aufl.). Berliner Wissenschafts-Verlag.

Lieberman, J. D. & Arndt, J. (2000). Understanding the limits of limiting instructions: Social psychological explanations for the failures of instructions to disregard pretrial publicity and other inadmissible evidence. *Psychology, Public Policy, and Law*, *6*(3), 677–711. https://doi.org/10.1037/1076-8971.6.3.677

Lin, J.-J., Chang, C.-H. & Pal, N. (2015). A revisit to contingency table and tests of independence: bootstrap is preferred to Chi-square approximations as well as Fisher's exact test. *Journal of biopharmaceutical statistics*, *25*(3), 438–458. https://doi.org/10.1080/10543406.2014.920851

Lindemann, M. (2015). Sind Fehler im Ermittlungsverfahren im weiteren Verfahren korrigierbar? Rechtssoziologische Kritik der Beweisverbotslehre. In S. Barton, R. Köbel & M. Lindemann (Hrsg.), *Interdisziplinäre Studien zu Recht und Staat: Bd. 54. Wider die wildwüchsige Entwicklung des Ermittlungsverfahrens* (1. Aufl., S. 127–149). Nomos.

Liu, Y., Wang, H., Li, L., Wang, Y [Yawei], Peng, J. & Di Baxter, F. (2019). Judgments in a hurry: Time pressure affects how judges assess unfairly shared losses and unfairly shared gains. *Scandinavian journal of psychology*, *60*(3), 203–212. https://doi.org/10.1111/sjop.12532

Lundberg, A. (2016). *Do Judges Vary Their Burden of Proof? Evidence from Federal Bench Trials.* https://doi.org/10.2139/ssrn.2816569

Macdonald, S., Erickson, P. & Allen, B. (1999). Judicial attitudes in assault cases involving alcohol or other drugs. *Journal of Criminal Justice*, *27*(3), 275–286. https://doi.org/10.1016/S0047-2352(98)00065-8

Machura, S. (2016). Understanding the German Mixed Tribunal. *Zeitschrift für Rechtssoziologie*, *36*(2), 273–302. https://doi.org/10.1515/zfrs-2016-0022

Maegherman, E. F. L. (2021). *Facilitating falsification in legal decision making: Problems in practice and potential solutions* [Dissertation]. Universität Maastricht, Maastricht. https://doi.org/10.26481/dis.20210114em

Maguire, N. (2010). Consistency in Sentencing. *Judicial Studies Institute Journal*, 14–54.

Maguire, N., Beyens, K., Boone, M., Laurinavicius, A. & Persson, A. (2015). Using vignette methodology to research the process of breach comparatively. *European Journal of Probation*, *7*(3), 241–259. https://doi.org/10.1177/2066220315617271

Mair, P. & Wilcox, R. (2020). Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods*, *52*, 464–488.

Maule, A. J. & Hockey, G. R. J. (1993). State, Stress, and Time Pressure. In A. J. Maule & O. Svenson (Hrsg.), *Time pressure and stress in human decision making* (S. 83–101). Plenum.

McCready, W. C. (2006). Applying Sampling Procedures. In F. T. L. Leong & J. T. Austin (Hrsg.), *The Psychology Research Handbook: A Guide for Graduate Students and Research Assistants* (S. 147–160). Sage Publications. https://doi.org/10.4135/9781412976626.n10

McElroy, T., Dickinson, D. L. & Levin, I. P. (2020). Thinking about decisions: An integrative approach of person and task factors. *Journal of Behavioral Decision Making*, *33*(4), 538–555. https://doi.org/10.1002/bdm.2175

McKay, C., Nolan, M. & Smithson, M. (2014). Effectiveness of Question Trails as Jury Decision Aids: the Jury's Still Out. *Psychiatry, Psychology, and*

Law: An Interdisciplinary Journal of the Australian and New Zealand Association of Psychiatry, Psychology and Law, *21*(4), 492–510. https://doi.org/10.1080/13218719.2013.839929

Meterko, V. & Cooper, G. (2022). Cognitive Biases in Criminal Case Evaluation: A Review of the Research. *Journal of Police and Criminal Psychology*, *37*(1), 101–122. https://doi.org/10.1007/s11896-020-09425-8

Miletić, S. & van Maanen, L. (2019). Caution in decision-making under time pressure is mediated by timing ability. *Cognitive psychology*, *110*, 16–29. https://doi.org/10.1016/j.cogpsych.2019.01.002

Miller, A. L. (2019). Expertise Fails to Attenuate Gendered Biases in Judicial Decision-Making. *Social Psychological and Personality Science*, *10*(2), 227–234. https://doi.org/10.1177/1948550617741181

Mishra, J., Allen, D. & Pearman, A. (2015). Information seeking, use, and decision making. *Journal of the Association for Information Science and Technology*, *66*(4), 662–673. https://doi.org/10.1002/asi.23204

Mitchell, J. B. (1989). Current Theories on Expert and Novice Thinking: A Full Faculty Considers the Implications for Legal Education. *Journal of Legal Education*, *39*(2), 275–297.

Morgan, R. M., Earwaker, H., Nakhaeizadeh, S., Harris, A. J. L., Rando, C. & Dror, I. E. (2018). Interpretation of forensic science evidence at every step of the forensic science process: Decision-making under uncertainty. In R. Wortley, A. Sidebottom, N. Tilley & G. Laycock (Hrsg.), *Routledge Handbook of Crime Science* (S. 408–420). Routledge.

Morsanyi, K., Busdraghi, C. & Primi, C. (2014). Mathematical anxiety is linked to reduced cognitive reflection: a potential road from discomfort in the mathematics classroom to susceptibility to biases. *Behavioral and Brain Functions*, *10*(1), 31. https://doi.org/10.1186/1744-9081-10-31

Mueller-Johnson, K., Dhami, M. K. & Lundrigan, S. (2018). Effects of judicial instructions and juror characteristics on interpretations of beyond reasonable doubt. *Psychology, Crime & Law*, *24*(2), 117–133. https://doi.org/10.1080/1068316X.2017.1394461

Nett, T., Dorrough, A., Jekel, M. & Glöckner, A. (2020). Perceived Biological and Social Characteristics of a Representative Set of German First Names. *Social Psychology*, *51*(1), 17–34. https://doi.org/10.1027/1864-9335/a000383

Nguyen, D., Kim, E., Wang, Y [Yan], Pham, T. V., Chen, Y.-H. & Kromrey, J. D. (2019). Empirical Comparison of Tests for One-Factor ANOVA Under Heterogeneity and Non-Normality: A Monte Carlo Study. *Journal of Modern Applied Statistical Methods*, *18*(2), 2–30. https://doi.org/10.22237/jmasm/1604190000

Nickolaus, C. (2018). *Ankereffekte im Strafprozess*. *Schriften zur Rechtspsychologie: Bd. 2*. Nomos. https://doi.org/10.5771/9783845294421

Niehaus, S., Englich, B. & Volbert, R. (2009). Psychologie des Strafverfahrens. In H.-L. Kröber (Hrsg.), *Handbuch der forensischen Psychiatrie: Kriminologie und forensische Psychiatrie* (S. 662–688). Steinkopff.

Nievelstein, F., van Gog, T., Boshuizen, H. P. A. & Prins, F. J. (2010). Effects of conceptual knowledge and availability of information sources on law students' legal reasoning. *Instructional Science*, *38*(1), 23–35. https://doi.org/10.1007/s11251-008-9076-3

Oechssler, J., Roider, A. & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, *72*(1), 147–152. https://doi.org/10.1016/j.jebo.2009.04.018

Oh, H., Beck, J. M., Zhu, P., Sommer, M. A., Ferrari, S. & Egner, T. (2016). Satisficing in split-second decision making is characterized by strategic cue discounting. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(12), 1937–1956. https://doi.org/10.1037/xlm0000284

Oh-Descher, H., Beck, J. M., Ferrari, S., Sommer, M. A. & Egner, T. (2017). Probabilistic inference under time pressure leads to a cortical-to-subcortical shift in decision evidence integration. *NeuroImage*, *162*, 138–150. https://doi.org/10.1016/j.neuroimage.2017.08.069

Ordóñez, L. & Benson, L. (1997). Decisions under Time Pressure: How Time Constraint Affects Risky Decision Making. *Organizational Behavior and*

Human Decision Processes, *71*(2), 121–140. https://doi.org/10.1006/obhd.1997.2717

Ormston, R., Chalmers, J., Leverick, F., Munro, V. & Murray, L. (2019). *Scottish jury research: Findings from a large scale mock jury study* (Social Research series). Edinburgh. Scottish Government. https://www.gov.scot/binaries/content/documents/govscot/publications/research-and-analysis/2019/10/scottish-jury-research-fingings-large-mock-jury-study-2/documents/scottish-jury-research-findings-large-scale-mock-jury-study/scottish-jury-research-findings-large-scale-mock-jury-study/govscot%3Adocument/scottish-jury-research-findings-large-scale-mock-jury-study.pdf

Orr, D. & Guthrie, C. (2006). Anchoring, Information, Expertise, and Negotiation: New Insights from Meta-Analysis. *Ohio State Journal on Dispute Resolution*, *21*, 597–628.

Oswald, M. E. (2009). How knowledge about the defendant's previous convictions influences judgments of guilt. In M. E. Oswald, S. Bieneck & J. Hupfeld-Heinemann (Hrsg.), *Social Psychology of Punishment of Crime* (S. 357–377). Wiley.

Oswald, M. E., Bieneck, S. & Hupfeld-Heinemann, J. (Hrsg.). (2009). *Social Psychology of Punishment of Crime*. Wiley.

Oswald, M. E. & Wyler, H. (2018). Fallstricke auf dem Weg zur »richtigen« Entscheidung im Strafrecht: Eine Analyse aus psychologischer Sicht. In S. Barton, M. Dubelaar, R. Kölbel & M. Lindemann (Hrsg.), *Vom hochgemuten, voreiligen Griff nach der Wahrheit* (S. 103–132). Nomos.

Oyeyemi, G. M. & Mbaeyi, G. C. (2018). On the estimation of empty cell probabilities in a contigency table. *Annals. Computer Science Series*, *16*(1), 22–26.

Pachur, T. & Marinello, G. (2013). Expert intuitions: how to model the decision strategies of airport customs officers? *Acta psychologica*, *144*(1), 97–103. https://doi.org/10.1016/j.actpsy.2013.05.003

Pantazi, M., Klein, O. & Kissine, M. (2020). Is justice blind or myopic? An examination of the effects of metacognitive myopia and truth bias on mock jurors and judges. *Judgment and Decision Making*, *15*(2), 214–229.

Pearson, J. M., Law, J. R., Skene, J. A. G., Beskind, D. H., Vidmar, N., Ball, D. A., Malekpour, A., Carter, R. M. & Skene, J. H. P. (2018). Modelling the effects of crime type and evidence on judgments about guilt. *Nature Human Behaviour*, *2*(11), 856–866. https://doi.org/10.1038/s41562-018-0451-z

Pechtl, H. (2009). *Anmerkungen zur Operationalisierung und Messung des Konstrukts 'need for cognition'* (Wirtschaftswissenschaftliche Diskussionspapiere 05/2009). Universität Greifswald. https://www.econstor.eu/handle/10419/41073

Pennington, N. & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, *51*(2), 242–258. https://doi.org/10.1037/0022-3514.51.2.242

Pennington, N. & Hastie, R. (1991). A Cognitive Theory of Juror Decision Making: The Story Model. *Cardozo Law Review*, *13*, 519–557.

Pennington, N. & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology*, *62*(2), 189–206. https://doi.org/10.1037/0022-3514.62.2.189

Pennycook, G., Fugelsang, J. A. & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive psychology*, *80*, 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., Ross, R. M., Koehler, D. J. & Fugelsang, J. A. (2017). Dunning-Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review*, *24*(6), 1774–1784. https://doi.org/10.3758/s13423-017-1242-7

Peters, J. H. & Dörfler, T. (2019a). *Planen, Durchführen und Auswerten von Abschlussarbeiten in der Psychologie und Sozialwissenschaften* (2., aktualisierte und erweiterte Aufl.). Pearson.

Peters, J. H. & Dörfler, T. (2019b). *Schreiben und Gestalten von Abschlussarbeiten in der Psychologie und den Sozialwissenschaften* (2., aktualisierte Aufl.). Pearson.

Petty, R. E., Briñol, P., Loersch, C. & McCaslin, M. J. (2009). The Need for Cognition. In M. R. Leary & R. H. Hoyle (Hrsg.), *Handbook of individual differences in social behavior* (S. 318–329). The Guilford Press.

Pfister, H.-R., Jungermann, H. & Fischer, K. (2017). *Die Psychologie der Entscheidung: Eine Einführung* (4. Aufl.). Springer.

Phillips, J. K., Klein, G. & Sieck, W. R. (2004). Expertise in Judgment and Decision Making: A Case for Training Intuitive Decision Skills. In D. J. Koehler & N. Harvey (Hrsg.), *Blackwell Handbook of Judgment and Decision Making* (S. 297–315). Blackwell Publishing Ltd.

Phillips, S. W. (2009). Using a Vignette Research Design to Examine Traffic Stop Decision Making of Police Officers. *Criminal Justice Policy Review*, *20*(4), 495–506. https://doi.org/10.1177/0887403409333070

Phillips, S. W. & Sobol, J. J. (2010). Twenty Years of Mandatory Arrest. *Criminal Justice Policy Review*, *21*(1), 98–118. https://doi.org/10.1177/0887403408322962

Phillips, W. J., Fletcher, J. M., Marks, A. D. G. & Hine, D. W. (2016). Thinking styles and decision making: A meta-analysis. *Psychological Bulletin*, *142*(3), 260–290. https://doi.org/10.1037/bul0000027

Pickel, K. L., Karam, T. J. & Warner, T. C. (2009). Jurors' Responses To Unusual Inadmissible Evidence. *Criminal Justice and Behavior*, *36*(5), 466–480. https://doi.org/10.1177/0093854809332364

Pospeschill, M. (2013). *Empirische Methoden in der Psychologie*. Reinhardt.

Pronin, E., Gilovich, T. & Ross, L. (2004). Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. *Psychological Review*, *111*(3), 781–799. https://doi.org/10.1037/0033-295X.111.3.781

Pronin, E., Lin, D. Y. & Ross, L. (2002). The Bias Blind Spot: Perceptions of Bias in Self Versus Others. *Personality & Social Psychology Bulletin*, *28*(3), 369–381. https://doi.org/10.1177/0146167202286008

Rachlinski, J. J. (2000). Heuristics and Biases in the Court: Ignorance or Adaptation? *Oregon Law Review*, *79*(1), 61–102.

Rachlinski, J. J. & Wistrich, A. J. (2017). Judging the Judiciary by the Numbers: Empirical Research on Judges. *Annual Review of Law and Social Science*, *13*(1), 203–229. https://doi.org/10.1146/annurev-lawsocsci-110615-085032

Rae, B., Heathcote, A., Donkin, C., Averell, L. & Brown, S. (2014). The hare and the tortoise: emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(5), 1226–1243. https://doi.org/10.1037/a0036801

Raoelison, M., Thompson, V. A. & Neys, W. de (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Rassin, E. (2020). Context effect and confirmation bias in criminal fact finding. *Legal and Criminological Psychology*, *25*(2), 80–89. https://doi.org/10.1111/lcrp.12172

Rastegary, H. & Landy, F. J. (1993). The Interactions among Time Urgency, Uncertainty, and Time Pressure. In A. J. Maule & O. Svenson (Hrsg.), *Time pressure and stress in human decision making* (S. 217–239). Plenum.

Read, S. J., Vanman, E. J. & Miller, L. C. (1997). Connectionism, parallel constraint satisfaction processes, and gestalt principles: (re) introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review*, *1*(1), 26–53. https://doi.org/10.1207/s15327957pspr0101_3

Rebehn, S. (2021). *Strafjustiz am Limit*. https://www.drb.de/newsroom/presse-mediencenter/nachrichten-auf-einen-blick/nachricht/news/strafjustiz-am-limit-1

Redding, R. E. & Reppucci, N. D. (1999). Effects of lawyers' socio-political attitudes on their judgments of social science in legal decision making. *Law and Human Behavior*, *23*(1), 31–54. https://doi.org/10.1023/A:1022322706533

Rice, S. & Trafimow, D. (2012). Time Pressure Heuristics Can Improve Performance Due to Increased Consistency. *The Journal of General Psychology*, *139*(4), 273–288.

Rieskamp, J. & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In G. Gigerenzer, P. M. Todd & The ABC Research Group (Hrsg.), *Simple Heuristics That Make Us Smart* (S. 141–167). Oxford University Press.

Rieskamp, J. & Hoffrage, U. (2008). Inferences under time pressure: how opportunity costs affect strategy selection. *Acta psychologica*, *127*(2), 258–276. https://doi.org/10.1016/j.actpsy.2007.05.004

Rossi, P. H., Simpson, J. E. & Miller, J. L. (1985). Beyond Crime Seriousness: Fitting the Punishment to the Crime. *Journal of Quantitative Criminology*, *1*(1), 59–90.

Rossmo, K. & Pollock, J. (2018). *Case Deconstruction of Criminal Investigative Failures: Final Summary Overview.* U. S. Department of Justice. https://www.ojp.gov/pdffiles1/nij/grants/254340.pdf

Rudolph, U., Böhm, R. & Lummer, M. (2007). Ein Vorname sagt mehr als 1000 Worte. *Zeitschrift für Sozialpsychologie*, *38*(1), 17–31. https://doi.org/10.1024/0044-3514.38.1.17

Ruppenthal, M. (2022, 8. Januar). *Evaluation of Evidence and Decision-Making Processes in the Investigatory Process: A Comparison of Laypeople, Legal Novices and Experts*. osf.io/56bpf

Ruscio, J. (2000). The role of complex thought in clinical prediction: social accountability and the need for cognition. *Journal of Consulting and Clinical Psychology*, *68*(1), 145–154. https://doi.org/10.1037/0022-006X.68.1.145

Sagana, A. (2018). The downward spiral of biases in criminal investigations: From eyewitnesses to forensic experts and judges. In S. Barton, M. Dubelaar, R. Kölbel & M. Lindemann (Hrsg.), *Vom hochgemuten, voreiligen Griff nach der Wahrheit* (S. 133–146). Nomos.

Sagana, A. & Sauerland, M. (2020). The Psychology of Forensic Evidence. *Zeitschrift für Psychologie*, *228*(3), 145–148. https://doi.org/10.1027/2151-2604/a000418

Sagana, A. & van Toor, D. A. G. (2020). The Judge as a Procedural Decision-Maker: Addressing the Disconnect Between Legal Psychology and Legal Practice. *Zeitschrift für Psychologie*, *228*(3), 226–228. https://doi.org/10.1027/2151-2604/a000417

Satzger, H. (2010). Die Rolle des Richters im Ermittlungsverfahren in Deutschland und Frankreich. In H. Jung, J. Leblois-Happe & C. Witz (Hrsg.), *Saarbrücker Studien zum Internationalen Recht: Band 44. 200 Jahre Code d'instruction criminelle: Le Bicentenaire du Code d'instruction criminelle* (S. 93–107). Nomos.

Sauer, C., Auspurg, K., Hinz, T. & Liebig, S. (2011). The Application of Factorial Surveys in General Population Samples: The Effects of Respondent Age and Education on Response Times and Response Consistency. *Survey Research Methods*, *5*(3), 89–102.

Savolainen, R. (2006). Time as a context of information seeking. *Library & Information Science Research*, *28*(1), 110–127. https://doi.org/10.1016/j.lisr.2005.11.001

Schmittat, S. M. (2017). Psychologische Grundlagen der Beweisführung. *Journal für Strafrecht*(5), 444–452.

Schmittat, S. M. (2022). Prior Conviction Evidence: Harmful or Irrelevant? A Literature Review. *Journal of Police and Criminal Psychology.* Vorab-Onlinepublikation. https://doi.org/10.1007/s11896-022-09557-z

Schmittat, S. M. & Burgmer, P. (2020). Lay beliefs in moral expertise. *Philosophical Psychology*, *33*(2), 283–308. https://doi.org/10.1080/09515089.2020.1719053

Schmittat, S. M. & Englich, B. (2016). If you judge, investigate! Responsibility reduces confirmatory information processing in legal experts. *Psychology, Public Policy, and Law*, *22*(4), 386–400. https://doi.org/10.1037/law0000097

Schmittat, S. M., Englich, B., Sautner, L. & Velten, P. (2022). Alternative stories and the decision to prosecute: an applied approach against confirmation bias in criminal prosecution. *Psychology, Crime & Law*, *28*(6), 608–635. https://doi.org/10.1080/1068316X.2021.1941013

Schnurr, S. (2003). Vignetten in quantitativen und qualitativen Forschungsdesigns. In H.-U. Otto (Hrsg.), *Empirische Forschung und soziale Arbeit: Ein Lehr- und Arbeitsbuch* (S. 393–400). Luchterhand.

Schroeder, F.-C. & Verrel, T. (2017). *Strafprozessrecht* (7., neu bearbeitete Auflage). *Grundrisse des Rechts*. C.H. Beck.

Schweizer, M. (2005). *Kognitive Täuschungen vor Gericht: Eine empirische Studie* [Dissertation]. Universität Zürich, Zürich.

Schweizer, M. (2007). Bestätigungsfehler – oder wir hören nur, was wir hören wollen. *Justice - Justiz - Giustizia*(2007/3).

Schweizer, M. (2009). Urteilen zwischen Intuition und Reflexion. *Justice - Justiz - Giustizia*(2009/4).

Schweizer, M. (2013). Comparing holistic and atomistic evaluation of evidence. *Law, Probability and Risk*, *13*(1), 65–89. https://doi.org/10.1093/lpr/mgt013

Schweizer, M. (2015). *Beweiswürdigung und Beweismaß: Rationalität und Intuition*. *Jus Privatum: v. 189*. Mohr Siebeck.

Schweizer, M. (2016). The civil standard of proof—what is it, actually? *The International Journal of Evidence & Proof*, *20*(3), 217–234. https://doi.org/10.1177/1365712716645227

Schweizer, M. (2019). Standard of Proof as Decision Threshold. In L. Tichý (Hrsg.), *Veröffentlichungen zum Verfahrensrecht: Bd. 158. Standard of proof in Europe* (1. Aufl., Bd. 158, S. 19–50). Mohr Siebeck.

Scopelliti, I., Morewedge, C. K., McCormick, E., Min, H. L., Lebrecht, S. & Kassam, K. S. (2015). Bias Blind Spot: Structure, Measurement, and Consequences. *Management Science*, *61*(10), 2468–2486. https://doi.org/10.1287/mnsc.2014.2096

Scurich, N. & John, R. S. (2017). Jurors' Presumption of Innocence. *The Journal of Legal Studies*, *46*(1), 187–206. https://doi.org/10.1086/690450

Scurich, N., Nguyen, K. D. & John, R. S. (2016). Quantifying the presumption of innocence. *Law, Probability and Risk*, *15*(1), 71–86. https://doi.org/10.1093/lpr/mgv016

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta psychologica*, *68*(1-3), 203–215. https://doi.org/10.1016/0001-6918(88)90056-X

Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta psychologica*, *81*(1), 75–86. https://doi.org/10.1016/0001-6918(92)90012-3

Shanteau, J. (2000). Why Do Experts Disagree? In B. Green, R. Cressy, F. Delmar, T. Eisenberg, B. Howcroft, M. Lewis, D. Schoenmaker, J. Shanteau & R. Vivian (Hrsg.), *Risk Behaviour and Risk Management in Business Life* (S. 186–196). Springer.

Shanteau, J. & Stewart, T. R. (1992). Why study expert decision making? Some historical perspectives and comments. *Organizational Behavior and Human Decision Processes*, *53*(2), 95–106. https://doi.org/10.1016/0749-5978(92)90057-E

Shestowsky, D. & Horowitz, L. M. (2004). How the Need for Cognition Scale Predicts Behavior in Mock Jury Deliberations. *Law and Human Behavior*, *28*(3), 305–337.

Shestowsky, D., Wegener, D. T. & Fabrigar, L. R. (1998). Need for cognition and interpersonal influence: Individual differences in impact on dyadic decisions. *Journal of Personality and Social Psychology*, *74*(5), 1317–1328. https://doi.org/10.1037/0022-3514.74.5.1317

Simon, D. (1998). A Psychological Model of Judicial Decision Making. *Rutgers Law Journal*, *30*, 1–142.

Simon, D. (2004). A Third View of the Black Box: Cognitive Coherence in Legal Decision Making. *The University of Chicago Law Review*, *71*(2), 511–586.

Simon, D. (2011). The Limited Diagnosticity of Criminal Trials. *Vanderbilt Law Review*, *64*, 143–223.

Simon, D. (2012). More Problems with Criminal Trials: The Limited Effectiveness of Legal Mechanisms. *Law and Contemporary Problems*, *75*(2), 167–209. https://doi.org/10.2139/ssrn.2106934

Simon, D. (2019). On Juror Decision Making: An Empathic Inquiry. *Annual Review of Law and Social Science*, *15*(1), 415–435. https://doi.org/10.1146/annurev-lawsocsci-101518-042658

Simon, D., Pham, L. B., Le, Q. A. & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *27*(5), 1250–1260. https://doi.org/10.1037/0278-7393.27.5.1250

Simon, D. & Scurich, N. (2011). Lay Judgments of Judicial Decision Making. *Journal of Empirical Legal Studies*, *8*(4), 709–727.

Simon, D. & Scurich, N. (2013). The Effect of Legal Expert Commentary on Lay Judgments of Judicial Decision Making. *Journal of Empirical Legal Studies*, *10*(4), 797–814.

Simon, D., Snow, C. J. & Read, S. J. (2004b). The redux of cognitive consistency theories: evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology*, *86*(6), 814–837. https://doi.org/10.1037/0022-3514.86.6.814

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Sirota, M., Dewberry, C., Juanchich, M., Valuš, L. & Marshall, A. C. (2021). Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *Journal of Behavioral Decision Making*, *34*(3), 322–343. https://doi.org/10.1002/bdm.2213

Smith, S. M. & Levin, I. P. (1996). Need for Cognition and Choice Framing Effects. *Journal of Behavioral Decision Making*(9), 283–290.

Smithson, M., Deady, S. & Gracik, L. (2007). Guilty, not guilty, or …? Multiple options in jury verdict choices. *Journal of Behavioral Decision Making*, *20*(5), 481–498. https://doi.org/10.1002/bdm.572

Sommers, I., Goldstein, J. & Baskin, D. (2014). The Intersection of Victims' and Offenders' Sex and Race/Ethnicity on Prosecutorial Decisions for Violent Crimes. *The Justice System Journal*, *35*(2), 178–204.

Sonnentag, S. (2006). *Abschlussarbeiten und Dissertationen in der angewandten psychologischen Forschung*. Hogrefe.

Spellman, B. A. (2007). On the Supposed Expertise of Judges in Evaluating Evidence. *University of Pennsylvania Law Review*(157), 1–9.

Sporer, S. L. & Goodman-Delahunty, J. (2009). Disparities in sentencing decisions. In M. E. Oswald, S. Bieneck & J. Hupfeld-Heinemann (Hrsg.), *Social Psychology of Punishment of Crime* (S. 379–401). Wiley.

Šrol, J. (2018). These Problems Sound Familiar to Me: Previous Exposure, Cognitive Reflection Test, and the Moderating Role of Analytic Thinking. *Studia Psychologica*, *60*(3), 195–208. https://doi.org/10.21909/sp.2018.03.762

Stanovich, K. E. (2016). The Comprehensive Assessment of Rational Thinking. *Educational Psychologist*, *51*(1), 23–34. https://doi.org/10.1080/00461520.2015.1125787

Stanovich, K. E. (2018). Miserliness in human cognition: the interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E. & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161–188. https://doi.org/10.1037/0096-3445.127.2.161

Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645-65. https://doi.org/10.1017/s0140525x00003435

Stanovich, K. E. & West, R. F. (2008). On the failure of cognitive ability to predict myside and one-sided thinking biases. *Thinking & Reasoning*, *14*(2), 129–167. https://doi.org/10.1080/13546780701679764

Stanovich, K. E. & West, R. F. (2014). The Assessment of Rational Thinking. *Teaching of Psychology*, *41*(3), 265–271. https://doi.org/10.1177/0098628314537988

Stanovich, K. E., West, R. F. & Toplak, M. E. (2014). Rationality, intelligence, and the defining features of Type 1 and Type 2 processing. In J. W. Sherman, B. Gawronski & Y. Trope (Hrsg.), *Dual-process theories of the social mind* (S. 80–91). The Guilford Press.

Statistisches Bundesamt. (2020a). *Rechtspflege: Staatsanwaltschaften* (Fachserie 10 Reihe 2.6). https://www.statistischebibliothek.de/mir/receive/DEHeft_mods_00134509

Statistisches Bundesamt. (2020b). *Rechtspflege: Strafgerichte* (Fachserie 10 Reihe 2.3). https://www.destatis.de/DE/Themen/Staat/Justiz-Rechtspflege/Publikationen/Downloads-Gerichte/strafgerichte-2100230207004.pdf?__blob=publicationFile

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4. Aufl.). Erlbaum.

Stiensmeier-Pelster, J. & Schürmann, M. (1993). Information Processing in Decision Making under Time Pressure. In A. J. Maule & O. Svenson (Hrsg.), *Time pressure and stress in human decision making* (S. 241–253). Plenum.

Suhling, S., Löbmann, R. & Greve, W. (2005). Zur Messung von Strafeinstellungen. *Zeitschrift für Sozialpsychologie*, *36*(4), 203–213. https://doi.org/10.1024/0044-3514.36.4.203

Szaszi, B., Szollosi, A., Palfi, B. & Aczel, B. (2017). The cognitive reflection test revisited: exploring the ways individuals solve the test. *Thinking & Reasoning*, *23*(3), 207–234. https://doi.org/10.1080/13546783.2017.1292954

Tabachnick, B. G. & Fidell, L. S. (2013). *Using Multivariate Statistics* (6. Aufl.). Pearson Education Limited.

Tersago, P., Vanderhallen, M., Rozie, J. & McIntyre, S.-J. (2020). From Suspect Statement to Legal Decision Making. *Zeitschrift für Psychologie*, *228*(3), 175–187. https://doi.org/10.1027/2151-2604/a000412

Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. S. B. T. Evans & K. Frankish (Hrsg.), *In two minds: Dual processes and beyond* (S. 171–196). Oxford University Press.

Thompson, V. A., Prowse Turner, J. A. & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Thomson, K. S. & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. https://doi.org/10.1017/S1930297500007622

Tichý, L. (Hrsg.). (2019). *Veröffentlichungen zum Verfahrensrecht: Bd. 158*. *Standard of proof in Europe* (1 Aufl.). Mohr Siebeck.

Toplak, M. E., West, R. F. & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F. & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Towfigh, E. V. & Glöckner, A. (2015). Entscheidungen zwischen "Intuition" und "Rationalität". *Deutsche Richterzeitung*, 270–273.

Travers, E., Rolison, J. J. & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition*, *150*, 109–118. https://doi.org/10.1016/j.cognition.2016.01.015

Trimmel, M. (2009). *Wissenschaftliches Arbeiten in Psychologie und Medizin* (1. Aufl.). UTB GmbH.

Trueblood, J. S., Holmes, W. R., Seegmiller, A. C., Douds, J., Compton, M., Szentirmai, E., Woodruff, M., Huang, W., Stratton, C. & Eichbaum, Q.

(2018). The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making. *Cognitive Research: Principles and Implications*, *3*(1), 1–14. https://doi.org/10.1186/s41235-018-0119-2

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

van der Vegt, A., Zuccon, G., Koopman, B. & Deacon, A. (2020). How searching under time pressure impacts clinical decision making. *Journal of the Medical Library Association:JMLA*, *108*(4), 564–573. https://doi.org/10.5195/jmla.2020.915

van Maanen, L., Fontanesi, L., Hawkins, G. E. & Forstmann, B. U. (2016). Striatal activation reflects urgency in perceptual decision making. *NeuroImage*, *139*, 294–303. https://doi.org/10.1016/j.neuroimage.2016.06.045

Verplanken, B. (1993). Need for Cognition and External Information Search: Responses to Time Pressure during Decision-Making. *Journal of Research in Personality*, *27*(3), 238–252. https://doi.org/10.1006/jrpe.1993.1017

Visher, C. A. (1987). Juror Decision Making: The Importance of Evidence. *Law and Human Behavior*, *11*(1), 1–17.

Weiss, D. J. & Shanteau, J. (2012). Decloaking the privileged expert. *Journal of Management & Organization*, *18*(3), 300–310. https://doi.org/10.5172/jmo.2012.18.3.300

West, R. F., Meserve, R. J. & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*, *103*(3), 506–519. https://doi.org/10.1037/a0028857

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67–85. https://doi.org/10.1016/0001-6918(77)90012-9

Wistrich, A. J., Guthrie, C. P. & Rachlinski, J. J. (2005). Can Judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding. *University of Pennsylvania Law Review*, *153*(4), 1251–1345.

Wistrich, A. J., Rachlinski, J. J. & Guthrie, C. (2015). Heart versus Head: Do Judges Follow the Law or Follow Their Feelings? *Texas Law Review*, *93*, 855–923.

Wixted, J. T., Mickes, L. & Fisher, R. P. (2018). Rethinking the Reliability of Eyewitness Memory. *Perspectives on Psychological Science*, *13*(3), 324–335. https://doi.org/10.1177/1745691617734878

Wood, S. M., DeVault, A., Miller, M. K., Kemmelmeier, M. & Summers, A. D. (2019). Decision-making in civil litigation: Effects of attorney credibility, evidence strength, and juror cognitive processing. *Journal of Applied Social Psychology*, *49*(8), 498–518. https://doi.org/10.1111/jasp.12600

Wright, D. B. & Hall, M. (2007). How a "Reasonable Doubt" Instruction Affects Decisions of Guilt. *Basic and Applied Social Psychology*, *29*(1), 91–98. https://doi.org/10.1080/01973530701331254

Wyler, H. (2021). Befragungsmethoden und Erkenntnisgewinn im Strafverfahren aus psychologischer Sicht. In R. Deckers & G. Köhnken (Hrsg.), *Die Erhebung und Bewertung von Zeugenaussagen im Strafprozess: Juristische, aussagepsychologische und psychiatrische Aspekte* (S. 110–162). BWV Berliner Wissenschafts-Verlag GmbH.

Zakay, D. (1993). The Impact of Time Perception Processes on Decision Making under Time Stress. In A. J. Maule & O. Svenson (Hrsg.), *Time pressure and stress in human decision making* (S. 59–72). Plenum.

Zapf, P. A., Kukucka, J., Kassin, S. M. & Dror, I. E. (2018). Cognitive bias in forensic mental health assessment: Evaluator beliefs about its nature and scope. *Psychology, Public Policy, and Law*, *24*(1), 1–10. https://doi.org/10.1037/law0000153

Anhang: Untersuchungsmaterialien

A-1 Grafik “Entscheidungsoptionen

A-2 Vignette “Diebstahl”

A-3 Vignette “Körperverletzung”

A-1 Grafik "Entscheidungsoptionen"

1. Erhebung einer Anklage
• **Hinreichender Tatverdacht** gegen die beschuldigte Person liegt vor → Eröffnung einer **Hauptverhandlung vor Gericht** zur **Klärung** der **Schuldfrage** und des **Strafmaßes**
2. Einstellung unter Auflagen
• **Hinreichender Tatverdacht** gegen die beschuldigte Person liegt vor • **Jedoch: Absehen** von einer **Anklage**, da **stattdessen Auflagen und Weisungen** erteilt werden, z.B. gemeinnützige Leistung → **Verfahrensbeendigung** ohne Hauptverhandlung vor Gericht
3. Einstellung wegen Geringfügigkeit
• **Hinreichender Tatverdacht** gegen die beschuldigte Person liegt vor • **Jedoch: Schuld** der Tatperson wird als **gering** angesehen **und** es besteht **kein öffentliches Interesse** an einer Verfolgung → **Verfahrensbeendigung** ohne Hauptverhandlung vor Gericht

A-2 Vignette “Diebstahl”

In diesem Fall geht es um Diebstahl.

§242 StGB

(1) Wer eine fremde bewegliche Sache einem anderen in der Absicht wegnimmt, die Sache sich oder einem Dritten rechtswidrig zuzueignen, wird mit Freiheitsstrafe bis zu fünf Jahren oder mit Geldstrafe bezahlt.

(2) Der Versuch ist strafbar.

Michael B. (23 Jahre, deutscher Staatsbürger) wird vorgeworfen, einer ihm unbekannten Person On-Ear-Kopfhörer (neuwertig, schwarz, seit wenigen Wochen erhältlich) im Wert von 169,00€ aus dem Auto gestohlen zu haben. Die Tatzeit war der 16.12.2019 gegen 20:15 Uhr.

Der Geschädigte hatte sein Auto auf einem Parkplatz am Rande der örtlichen Einkaufsstraße geparkt. Er gab an, dass er seine Einkäufe im Kofferraum verstaut habe, als ihm eingefallen sei, dass er sein Parkticket noch nicht entwertet hatte. Der Geschädigte sei zum Parkautomaten gegangen und habe dabei vergessen, sein Auto zu verriegeln. Da er damit beschäftigt gewesen sei, das Parkticket zu entwerten, habe er den Tathergang nicht mitbekommen. Erst durch die lauten Rufe des Zeugen sei der Geschädigte auf den Diebstahl aufmerksam geworden.

Ein Zeuge, der in einem in der Nähe geparkten Auto saß, beobachtete den Diebstahl. Er sagte aus, dass eine unbekannte Person über den Parkplatz gegangen sei und sich zielstrebig dem Auto des Geschädigten genähert habe. Die Person habe aus dem Kofferraum die Tüte eines Elektromarktes herausgenommen und den Parkplatz in Richtung eines angrenzenden Parks verlassen. Da der Parkplatz nur spärlich beleuchtet wird, habe der Zeuge die Tatperson nicht gut erkennen können. Der Zeuge gab an, dass es sich um einen Mann zwischen 1,70m und 1,85m in dunkler Winterkleidung gehandelt habe. Andere Personen seien zur Tatzeit nicht in der Nähe des Tatortes gewesen.

Die Beschreibung zu Größe und Kleidung der Tatperson wurde an eine sich zu dem Zeitpunkt in der Nähe befindende Polizeistreife übermittelt. Diese wurde in dem oben genannten Park auf Michael B. aufmerksam. Michael B. trug eine schwarze

Winterjacke, Mütze und Handschuhe. Da der Zeuge die Tatperson nicht gut gesehen hatte, war eine spätere Gegenüberstellung nicht aussagekräftig.

Die Tüte des Elektromarktes oder der leere Karton des Artikels wurden nicht bei Michael B. gefunden. Allerdings trug er das Kopfhörermodell, welches der Geschädigte als gestohlen beschrieben hatte, Marke und Farbe stimmten überein. Aufgrund der Neuwertigkeit des Produkts gab es keine besonderen Erkennungsmerkmale, an denen der Besitz des Geschädigten hätte festgestellt werden können.

Bei der polizeilichen Vernehmung bestritt Michael B. die Tat und sagte aus, einen Abendspaziergang durch den Park gemacht zu haben. Er könne aber keine Person angeben, die ihn begleitet hätte und dies bestätigen könnte. Michael B. erklärte, dass er die Kopfhörer vor knapp zwei Wochen in einem Elektrofachmarkt von seinem Geburtstagsgeld in bar bezahlt habe. Die Quittung habe er aus Gewohnheit direkt nach dem Kauf weggeworfen.

Michael B. ist laut dem Bundeszentralregisterauszug nicht vorbestraft.

A-3 Vignette “Körperverletzung”

In diesem Fall geht es um Körperverletzung.

§223 StGB

(1) Wer eine andere Person körperlich misshandelt oder an der Gesundheit schädigt, wird mit Freiheitsstrafe bis zu fünf Jahren oder mit Geldstrafe bestraft.

(2) Der Versuch ist strafbar.

Matthias P. (24 Jahre, deutscher Staatsbürger) wird vorgeworfen, eine ihm unbekannte Person von hinten geschubst zu haben. Der Geschädigte verstauchte sich bei dem Versuch, den Sturz abzufangen, das rechte Handgelenk. Außerdem erlitt er Schürfwunden am rechten Arm sowie Knie. Die Tatzeit war der 08.01.2019 gegen 20:30 Uhr.

Der Geschädigte gab an, dass er nach dem Training an einem öffentlichen Sportplatz gerade dabei gewesen sei, sein Fahrrad aus der Halterung eines Fahrradständers zu nehmen, als er plötzlich eine Hand am Schulterblatt gespürt habe. Durch einen kräftigen Stoß sei er auf den roten Aschebelag neben dem Fahrradständer gestürzt. Dabei sei seine Brille zu Boden gefallen und zerbrochen, sodass er keine Angaben zu Beobachtungen nach dem Sturz machen konnte. In dem Fahrradständer befanden sich zu dem Zeitpunkt weitere Fahrräder, die alle dicht beieinanderstanden.

Ein Zeuge, der gerade vom Sportplatz ging, sei durch den Schrei des Geschädigten aufmerksam geworden. Er sagte aus, den Geschädigten erblickt zu haben, als dieser bereits auf dem Boden lag. Der Zeuge habe bemerkt, wie eine unbekannte Person eines der Fahrräder in dem Fahrradständer zügig aufgeschlossen habe und in schnellem Tempo in Richtung Hauptstraße davongefahren sei. Da der Zeuge noch einige Meter entfernt gewesen sei und der Fahrradständer nur spärlich beleuchtet wird, habe er die Person nicht gut erkennen können. Der Zeuge gab an, dass es sich um einen Mann in dunkler Sportkleidung gehandelt habe, der auf einem schwarzen Fahrrad ohne Gepäckträger davongefahren sei. Andere Personen seien zur Tatzeit nicht in der Nähe des Tatortes gewesen.

Die Beschreibung zu Kleidung und Fahrrad des Täters wurde an eine sich zu dem Zeitpunkt in der Nähe befindende Polizeistreife übermittelt. Diese wurde an der

oben genannten Hauptstraße auf Matthias P. aufmerksam. Matthias P. schob ein schwarzes Mountainbike neben sich her und überquerte gerade eine Kreuzung, an die eine Grünanlage grenzt. Er trug eine schwarze Trainingshose, einen dunklen Kapuzenpullover und mit roter Asche verschmutzte Laufschuhe. Da der Zeuge die Tatperson nicht gut gesehen hatte, war eine spätere Gegenüberstellung nicht aussagekräftig.

Bei der polizeilichen Vernehmung bestritt Matthias P. die Tat und sagte aus, dass er in der unmittelbar an die Kreuzung grenzenden Grünanlage joggen gewesen und nun mit dem zuvor in der Nähe abgestellten Fahrrad auf dem Heimweg sei. Er könne aber keine Person angeben, die ihn begleitet hätte und dies bestätigen könnte. Matthias P. erklärte, dass er mit seinen Laufschuhen auch auf frei zugänglichen Sportplätzen trainiere, weshalb diese mit Asche verschmutzt waren.

Matthias P. ist laut dem Bundeszentralregisterauszug nicht vorbestraft.