

# **Risk-Assessment bei Gewalt- und Sexualdelinquenz**

– Standardisierte Risk-Assessment Instrumente auf dem Prüfstand –

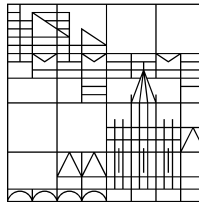
## **Dissertation**

zur Erlangung des akademischen Grades des Doktors der

Naturwissenschaften (Dr. rer. nat.)

an der

Universität  
Konstanz



Mathematisch-Naturwissenschaftliche Sektion  
Fachbereich Psychologie

vorgelegt von

**Juliane Gerth**

Tag der mündlichen Prüfung: 04. Februar 2015

Referent: Prof. Dr. Jérôme Endrass

Referent: Prof. Dr. Thomas Elbert

---

## **Danksagung**

Vielen Dank an all jene, die mich in den letzten Jahren bei den kleinen und großen Schritten bis hin zur Abgabe dieser Dissertation inhaltlich und persönlich unterstützt haben.

*Dank an meine Kollegen und ehemaligen Kollegen im Psychiatrisch-Psychologischen Dienst*

insbesondere an Jérôme Endrass und Astrid Rossegger  
sowie Elisabeth Bauch, Bernd Borchard, Cornel Gmür,  
Catherine Graber, Bettina Kuhn, Katharina Seewald,  
Jay P. Singh, Frank Urbaniok und Thomas Villmar

*Dank an die MitarbeiterInnen des Teams ‚Bedrohungsmanagement‘ der Stadtpolizei Zürich*

insbesondere an Peter Rügger, Sonja Müller und Severine Moor.

*Dank für den wissenschaftlichen Austausch insbesondere auch an*

N. Zoe Hilton,  
Jens Hoffmann und Marnie E. Rice

*Und natürlich Dank in die Runde meines Familien- und Freundeskreises*

insbesondere an meine Eltern und meinen Bruder  
sowie André, Anina, Jana, Jewgenij und Lea

---

## Inhaltsverzeichnis

<b>1. Einleitung.....</b>	<b>15</b>
1.1. Einführung ins Risk-Assessment bei Gewalt- und Sexualstraftätern .....	15
1.1.1. Die unstrukturiert-klinische Methode.....	16
1.1.2. Die mechanische Methode .....	19
1.1.3. Die strukturiert-klinische Methode.....	23
1.2. Validierung von Risk-Assessment Instrumenten .....	24
1.2.1. Mindestanforderungen an Replikationsstudien .....	26
1.2.2. Methodische Aspekte zur Untersuchung der Kriteriumsvalidität.....	30
1.3. Validität von Risk-Assessment Instrumenten bei Intimpartnergewalt.....	34
1.3.1. Übersetzung des Ontario Domestic Assault Risk Assessment (ODARA) .....	36
1.3.2. Kriteriumsvalidität des Ontario Domestic Assault Risk Assessment (ODARA) .....	37
1.3.3. Trennschärfe und Spezifität des Dynamischen Risiko-Analyse-Systems (DyRiAS) .....	40
1.4. Assessment-Strategien bei Drohungen .....	44
1.5. Ausblick .....	49
<b>2. Eigene Arbeiten.....</b>	<b>53</b>
2.1. Current obstacles in replicating risk assessment findings: A systematic review of commonly used actuarial instruments .....	53
2.1.1. Abstract .....	53
2.1.2. Introduction .....	54
2.1.3. Methods .....	56
2.1.4. Results .....	61
2.1.5. Discussion .....	65
2.1.6. Conclusion.....	69
2.2. Examining the predictive validity of the SORAG in Switzerland .....	70
2.2.1. Abstract .....	70
2.2.2. Introduction .....	71
2.2.3. Material and Methods.....	75
2.2.4. Results .....	78
2.2.5. Discussion .....	83
2.2.6. Limitations .....	85

---

2.3. Das Ontario Domestic Assault Risk Assessment (ODARA) – Validität und autorisierte deutsche Übersetzung eines Screening-Instruments für Risikobeurteilungen bei Intimpartnergewalt .....	86
2.3.1. Zusammenfassung .....	86
2.3.2. Prävalenz von Intimpartnergewalt .....	86
2.3.3. Risk-Assessment bei Intimpartnergewalt .....	88
2.3.4. Das Ontario Domestic Assault Risk Assessment – ODARA .....	90
2.3.5. Validität des ODARA .....	97
2.3.6. Zusammenfassung und Schlussfolgerung für die Praxis .....	102
2.3.7. Take Home Message .....	103
2.3.8. Anhang: Deutsche Übersetzung des ODARA .....	103
2.4. Assessing the discrimination and calibration of the Ontario Domestic Assault Risk Assessment in Switzerland.....	119
2.4.1. Abstract .....	119
2.4.2. Introduction .....	119
2.4.3. Method .....	127
2.4.4. Results .....	133
2.4.5. Discussion .....	137
2.4.6. Limitations .....	140
2.4.7. Conclusion.....	140
2.5. Assessing the Risk of Severe Intimate Partner Violence: Validating the DyRiAS in Switzerland.....	142
2.5.1. Abstract .....	142
2.5.2. Introduction .....	142
2.5.3. Methods.....	145
2.5.4. Results .....	147
2.5.5. Discussion .....	151
2.5.6. Limitations .....	152
2.5.7. Conclusion.....	153
2.6. Identifikation von Hoch-Risiko-Drohungen .....	154
2.6.1. Zusammenfassung .....	154
2.6.2. Drohungen.....	154
2.6.3. Drohungen und schwere Gewaltdelikte.....	157

---

2.6.4.	Hoch-Risiko-Drohungen .....	159
2.6.5.	Ausblick .....	165
<b>3.</b>	<b>Literaturverzeichnis .....</b>	<b>167</b>

---

## Tabellenverzeichnis

<b>Table 1.</b> Characteristics of three commonly used actuarial risk assessment instruments .....	58
<b>Table 2.</b> Handling attrition .....	62
<b>Table 3.</b> Correspondence between development and replication studies of three commonly used actuarial risk assessment instruments.....	63
<b>Table 4.</b> Characteristics of previous studies reporting risk bin outcome information for the SORAG .....	73
<b>Table 5.</b> Normative and observed risk bin distribution and recidivism rates for the SORAG.....	79
<b>Tabelle 6:</b> Normwerte des ODARA .....	93
<b>Tabelle 7:</b> Korrigierte Summenwerte des ODARA bei fehlender Information.....	94
<b>Tabelle 8:</b> Validierungsstudien zum ODARA (Stand März 2014).....	98
<b>Table 9:</b> Previous validation studies investigating the discrimination of the ODARA (June 2014) .....	126
<b>Table 10.</b> Correlation between ODARA items and police-registered IPV recidivism in the Zurich sample and internal consistency analysis of the ODARA scale (n = 185).....	131
<b>Table 11:</b> Recidivism rates for the ODARA after a 5-year time at risk (n = 185) and comparisons with normative risk rates.....	135
<b>Table 12.</b> DyRiAS risk category distribution and recidivism rates for IPV offenders at 3 months (n = 168), 6 months (n = 167), 1 year (n = 166) and 5 years (n = 146) time at risk.....	149
<b>Table 13.</b> Level of intervention for offenders of the 3-months subsample being assigned to the high-risk DyRiAS categories, which were issued by the police subsequent to the index assault.....	151

---

## Abbildungsverzeichnis

<b>Figure 1.</b> Flowchart depicting the results of a systematic search for replication studies of three commonly used risk assessment instruments (June 2012).....	57
<b>Figure 2.</b> Number of characteristics matched between development and replication studies of the VRAG.....	64
<b>Figure 3.</b> Number of characteristics matched between development and replication studies of the SORAG. ....	64
<b>Figure 4.</b> Number of characteristics matched between development and replication studies of the Static-99. ....	65
<b>Figure 5.</b> Number of characteristics matched between development and replication studies of three commonly used actuarial risk assessment instruments (VRAG, SORAG, Static-99).....	65
<b>Figure 6.</b> Receiver operating characteristic (ROC) graph displaying the discrimination of the SORAG risk bins in the ZSOP .....	80
<b>Figure 7.</b> Estimated receiver operating characteristic (ROC) graph displaying the discrimination of the SORAG risk bins in the development sample.....	80
<b>Figure 8.</b> Comparing expected and observed recidivism rates by applying Bayes' theory.....	81
<b>Figure 9.</b> Absolute differences in percentiles between the SORAG development sample and the Zurich sex offender population (ZSOP).....	82
<b>Figure 10.</b> Percentiles corresponding to SORAG total risk scores for the tool's development sample and the Zurich sex offender population (ZSOP).....	82
<b>Figure 11.</b> Flow chart depicting the process of fulfilling the ODARA's including criteria and applying a fixed time at risk of five years. ....	128
<b>Figure 12.</b> Observed IPV recidivism rates within a time at risk of five years surrounded by Bayesian credible intervals calculated by using the Jeffreys' prior for the Beta distribution. The observed rates were compared to the normative data (Hilton, Harris, & Rice, 2010). ....	136
<b>Abbildung 13:</b> Zielpersonen von Todesdrohungen (Warren et al., 2011).....	157
<b>Abbildung 14:</b> Hoch-Risiko-Drohung.....	160
<b>Abbildung 15:</b> Beurteilung des Risikopotenzials der Drohung auf einer 3-stufigen Skala (O'Toole 2000).....	161

---

## Abkürzungsverzeichnis

ARAI	-	Aktuarisches Risk-Assessment Instrument [Actuarial Risk Assessment Instrument]
AUC	-	Area under the curve
CI	-	Konfidenzintervall [Confidence Intervall]
DA	-	Danger Assessment
DVRAG	-	Domestic Violence Risk Appraisal Guide
DyRiAS	-	Dynamisches Risiko-Analyse-System
HCR-20	-	Historical Clinical Risk Management-20
ICC	-	Intraclass Correlation
IPH	-	Intimpartnertötung [Intimate Partner Homicide]
IPV/IPG	-	Intimate Partner Violence/Intimpartnergewalt
IQR	-	Inter quartile range
M	-	Mittelwert
Md	-	Modalwert
N.A.	-	Nicht zutreffend [Not applicable]
ODARA	-	Ontario Domestic Assault Risk Assessment
ROC	-	Receiver Operating Characteristic
SD	-	Standardabweichung [Standard deviation]
SAM	-	Stalking Assessment and Management Checklist
SARA	-	Spousal Risk Appraisal Guide
SORAG	-	Sex Offender Risk Appraisal Guide
SPJ	-	Strukturierte professionelle Urteilsbildung [Structured Professional Judgment]
VRAG	-	Violence Risk Appraisal Guide
WAVR-21	-	Workplace Assessment and Targeted Violence Risk

## **Zusammenfassung**

Risikobeurteilungen bei Gewalt- und Sexualstraftätern spielen eine bedeutende Rolle im polizeilichen und justiziellen Kontext. Dabei werden verschiedene Methoden angewandt, die sich hauptsächlich im Ausmaß ihrer Strukturierung voneinander unterscheiden. Nach einer Grundsatzkritik an frei und intuitiv durchgeführten Risikobeurteilungen in den 1970iger Jahren, wurden bis heute eine Vielzahl von strukturierten Instrumenten zur Risikobeurteilung (sogenannte Risk-Assessment Instrumente) entwickelt, von denen sich einige durch einen hochstandardisierten Mechanismus bei der Erhebung und Auswertung von risikorelevanten Merkmalen auszeichnen und andere zwar den Datenerhebungsprozess standardisieren, die Auswertung jedoch flexibel im Sinne einer individuellen vom Beurteilenden abhängigen Gesamteinschätzung bleibt.

Die wissenschaftliche Auseinandersetzung mit der Validität dieser verschiedenen Risk-Assessment Instrumente ist umfangreich und häufig von kontroversen Diskussionen geprägt. Der Großteil von Replikationsstudien betrifft die Trennschärfe der Risk-Assessment Instrumente bezüglich des Kriteriums Rückfälligkeit und weist im Durchschnitt auf die Überlegenheit standardisierter Verfahren gegenüber freien und intuitiven Einschätzung hin. Die Replikationsstudien sind jedoch häufig durch Mängel im Studiendesign, wie zum Beispiel durch Abweichungen bezüglich der Untersuchungspopulation, Operationalisierung von Rückfälligkeit oder Länge des Beobachtungszeitraumes, gekennzeichnet und ihre Aussagekraft daher mitunter fraglich. Darüber hinaus werden weitere Validitätsaspekte, wie z.B. die Kalibrierung, die eine wesentliche Bedeutung für die Interpretation der verschiedenen über die Risk-Assessment Instrumente ausgewiesenen Risikokategorien aufweist, meistens vernachlässigt.

Auf der Grundlage dieser Erkenntnisse war es Ziel der insgesamt sechs Studien der vorliegenden Dissertation verschiedene methodische Aspekte der Validität von standardisierten Risk-Assessment Instrumenten aufzuwerfen und anhand der Befunde von empirischen Untersuchungen zu diskutieren. Zunächst wurden Mindestanforderungen für Replikationsstudien erarbeitet und im Weiteren auf diesen basierend die Kriteriumsvalidität von drei standardisierten Risk-Assessment Instrumenten an Täterpopulationen im Kanton Zürich (Schweiz) überprüft. Diese wurden zur Risikobeurteilung von Sexualstraftätern (Sex Offender Risk Appraisal Guide [SORAG]) und für den Bereich der Intimpartnergewalt (Ontario Domestic Assault Risk Assessment [ODARA] und Dynamisches Risiko-Analyse-Systems [DyRiAS]) entwickelt. Eine der zwölf erarbeiteten Mindestanforderungen bezieht sich auf die manuellkonforme Anwendung von Risk-Assessment Instrumenten. Da die meisten Risk-Assessment Instrumente nicht im deutschsprachigen Raum entwickelt wurden, ergibt sich daraus die Forderung nach wissenschaftlichen Übersetzungen, die bezüglich des in Kanada entwickelten ODARA Inhalt einer weiteren Studie der vorliegenden Dissertation war. Zusammenfassend wiesen die untersuchten Risk-Assessment Instrumente im Kanton Zürich eine nur unzureichende bis moderate Fähigkeit, zwischen rückfälligen und nicht rückfälligen Straftätern zu diskriminieren sowie eine unzulängliche Kalibrierung beim SORAG und ODARA auf. Dieses Ergebnis ist im Allgemeinen auf eine Überschätzung des Rückfallrisikos zurückzuführen, wobei es offen bleibt, inwiefern risikorelevante dynamische Entwicklungen im Verlauf der Beobachtungszeit oder eine grundlegende inhaltliche Problematik des Risikomodells die geringe Spezifität der Instrumente herbeiführen. Im Bereich der Intimpartnergewalt fällt diese Überschätzung weniger stark in den unteren als in den Hochrisiko-Kategorien aus, weshalb sich für die Optimierung des Risk-Assessments Überlegungen zu einer stufenweise Risikobeurteilung ergeben, die zunächst ein

Screening und darauffolgend vertiefende Abklärungen bei als Hochrisiko-Täter eingeschätzten Personen vorsehen könnten. Prospektive Studien zur Erfassung dynamischer Prozesse sowie spezifische Analysen innerhalb von Hochrisikopopulationen sollten Gegenstand zukünftiger Studien sein.

Nicht für alle Bereiche, die einer Risikobeurteilung bedürfen, liegen schon Risk-Assessment Instrumente vor, deren Validität für die Anwendung in anderen regionalen Kontexten zunächst im Rahmen von Replikationsstudien überprüft werden könnte. Vor allem bezüglich der Entwicklung von Instrumenten zur Einschätzung des Ausführungsrisikos von Drohungen ist die Befundlage trotz langjähriger Forschung noch vage. Ziel der sechsten Studie war es daher auf der Grundlage einer umfangreichen Literaturlaufarbeitung ein allgemeines Modell zur Triagierung zwischen Niedrigrisiko- und Hochrisikodrohungen vorzustellen. Hierbei wurden vier risikorelevante Bereiche – Charakteristika der Drohung, Charakteristika der drohenden Person, Warnverhalten und aktuelle Belastungsfaktoren – identifiziert, die die Grundlage für eine strukturierte Erfassung von Risikomerkmale einer Drohung bilden. In zukünftigen Untersuchungen sollten eindeutige Operationalisierungen der Risikomerkmale erarbeitet und das Modell empirisch überprüft werden.

## Leistungsnachweis

### **1. Current obstacles in replicating risk assessment findings: A systematic review of commonly used actuarial instruments**

- *Autoren:* Astrid Rossegger, Juliane Gerth, Katharina Seewald, Frank Urbaniok, Jay P. Singh und Jerome Endrass
- *Publikationsstatus:* Publiziert in „Behavioral Sciences & the Law“ (2013), Volume 31, Issue 1, Seiten 154-164
- *Eigener Beitrag:* Mitwirkung am Studiendesign, der Datenerhebung, -aufbereitung und -auswertung sowie an allen Kapiteln der Manuskripterstellung

### **2. Examining the predictive validity of the SORAG in Switzerland**

- *Autoren:* Astrid Rossegger, Juliane Gerth, Jay P. Singh und Jerome Endrass
- *Publikationsstatus:* Publiziert in „Sexual Offender Treatment“ (2013), Volume 8, Issue 2
- *Eigener Beitrag:* Mitwirkung am Studiendesign, der Datenerhebung, -aufbereitung und -auswertung sowie an allen Kapiteln der Manuskripterstellung

### **3. Das Ontario Domestic Assault Risk Assessment (ODARA) – Validität und autorisierte deutsche Übersetzung eines Screening-Instruments für Risikobeurteilungen bei Intimpartnergewalt**

- *Autoren:* Juliane Gerth, Astrid Rossegger, Frank Urbaniok und Jerome Endrass
- *Publikationsstatus:* Im Druck bei „Fortschritte der Neurologie - Psychiatrie“
- *Eigener Beitrag:* Erarbeitung des Studiendesigns, Leitung des Übersetzungsprozesses, Koordination der Rückübersetzung, federführend in der Manuskripterstellung

#### **4. Assessing the discrimination and calibration of the Ontario Domestic Assault Risk**

##### **Assessment in Switzerland**

- *Autoren:* Juliane Gerth, Astrid Rossegger, Elisabeth Bauch, Jérôme Endrass
- *Publikationsstatus:* Eingereicht bei „Violence Against Women“
- *Eigener Beitrag:* Maßgebliche Mitwirkung bei der Erarbeitung des Studiendesigns, der Datenerhebung (einschließlich Anleitung und Aufsicht weiteren beteiligten Personen), eigenständige Datenaufbereitung und -auswertung, federführend in der Manuskripterstellung

#### **5. Assessing the Risk of Severe Intimate Partner Violence: Validating the DyRiAS in**

##### **Switzerland**

- *Autoren:* Juliane Gerth, Astrid Rossegger, Jay P. Singh und Jerome Endrass
- *Publikationsstatus:* Eingereicht bei „Archives Forensic Psychology“
- *Eigener Beitrag:* Maßgebliche Mitwirkung bei der Erarbeitung des Studiendesigns, der Datenerhebung (einschließlich Anleitung und Aufsicht weiteren beteiligten Personen), eigenständige Datenaufbereitung und -auswertung, federführend in der Manuskripterstellung

#### **6. Identifikation von Hoch-Risiko-Drohungen**

- *Autoren:* Juliane Gerth und Catherine Graber
- *Publikationsstatus:* Als Buchkapitel publiziert in Interventionen bei Gewalt- und Sexualstraftätern: Risk-Management, Methoden und Konzepte der forensischen Therapie (2012), J. Endrass, A. Rossegger, F. Urbaniok, B. Borchard (Eds.), Berlin: Medizinisch wissenschaftliche Verlagsgesellschaft, Seiten 393-401

- *Eigener Beitrag*: Erarbeitung des Konzepts für den Beitrag, Konzeption des Modells, Aufarbeitung und Integration der Literatur zum Modell, federführend bei der Manuskriptenterstellung

## 1. Einleitung

### 1.1. Einführung ins Risk-Assessment bei Gewalt- und Sexualstraftätern

Im Umgang mit Gewalt- und Sexualstraftätern kommt der Beurteilung des Rückfallrisikos in der Praxis eine entscheidende Rolle zu. So spielt beispielsweise in Deutschland (neben Aspekten der Schuldfähigkeit) das Rückfallrisiko eine substantielle Rolle, wenn über die Unterbringung von Straftätern in einem psychiatrischen Krankenhaus, das Gewähren von Vollzugslockerungen oder die Anordnung einer sichernden Maßregel entschieden werden soll (Boetticher et al., 2007). Risikobeurteilungen werden aber nicht nur im Kontext des Strafvollzugs vorgenommen, sondern finden auch in angrenzenden Bereichen wie bei polizeilichen Ermittlungsbehörden Einsatz, zum Beispiel bei der Beurteilung der Ausführungsgefahr von Drohungen oder einem risikoorientierten Umgang mit Fällen von Intimpartnergewalt (z.B. Kantonspolizei Zürich, 2014).

Während im deutschen Sprachraum der Begriff der Prognose für Risikobeurteilungen geläufig ist (Boetticher et al., 2007), verdeutlicht der im Englischen gebräuchliche Begriff des ‚risk assessment‘, dass es sich dabei um einen Prozess zur *Abschätzung des Risikos*, dass eine Ereignis eintritt, nicht aber um die *Vorhersage* darüber, *ob* das Ereignis eintritt oder nicht, handelt. Denn beim Prozess des forensischen Risk-Assessment geht es um die Sammlung risikorelevanter Informationen von Straftätern und deren Zusammenführung zu einer möglichst präzisen Schätzung über die Wahrscheinlichkeit, dass Straftäter erneut Delikte begehen (Lurigio & Taxman, 2013).

Die generelle Zuverlässigkeit von Risikobeurteilungen, aber auch die kontextuelle Angemessenheit verschiedener Methoden der Risikobeurteilung werden bis heute kontrovers diskutiert (Hilton, Harris, & Rice, 2006; Monahan, 1996; Skeem & Monahan, 2011).

Während Risikobeurteilungen bis weit in die zweite Hälfte des 20. Jahrhunderts hauptsächlich in unstrukturierter und intuitiver (in sogenannt unstrukturiert-klinischer) Form (siehe Kapitel 1.1.1) vorgenommen wurden, haben sich in den letzten 25 Jahren zwei weitere Methoden etabliert: die mechanische Methode, die sich durch eine standardisierte Datenerhebung und Datenauswertung auszeichnet (siehe Kapitel 1.1.2) und die strukturiert-klinische Methode, deren Ziel es ist, eine Brücke zwischen den zwei vorangehend genannten Methoden zu schlagen (siehe Kapitel 1.1.3). Hauptsächlich unterscheiden sich die drei Methoden also im Ausmaß der Standardisierung des Beurteilungsprozesses.

### **1.1.1. Die unstrukturiert-klinische Methode**

Risikobeurteilungen, die der unstrukturiert-klinischen Methode folgen, zeichnen sich durch intuitive und freihändige (unstrukturierte) Einschätzungen eines Experten aus. Damit ist nicht gemeint, dass es sich um ein spontanes oder gar laienhaftes Urteil handelt, sondern dass unstrukturiert-klinische Risikobeurteilungen keinem standardisierten Regelwerk folgen und die einzelfallorientierten Risikobeurteilungen durch Subjektivität geprägt sind: Art, und Gewichtung von Risikomerkmale liegen im Ermessen des Beurteilenden (vgl. Rossegger, Endrass, & Gerth, 2012).

Die unstrukturiert-klinische Methode war bis Mitte des 20. Jahrhunderts die in der Praxis vorherrschende Methode bei der Beurteilung des Rückfallrisikos von Straftätern. Gleichzeitig war die Praxis der Risikobeurteilungen durch eine fast ausschließliche Fokussierung auf die Ausprägung von Symptomen psychiatrischer Krankheitsbilder geprägt und es fand im Prinzip

---

eine – später viel kritisierte – Gleichsetzung von ‚Gefährlichkeit‘ und ‚psychischer Krankheit‘ statt (Monahan, 1984; Steadman & Cocozza, 1974). Dass eine solch starke Fokussierung auf Krankheitsbilder zur Einschätzung der ‚Gefährlichkeit‘ nicht gerechtfertigt ist, war spätestens nach der 1974 publizierten Arbeit von Steadman und Cocozza nicht mehr abzustreiten: Ein Gerichtsurteil des obersten Gerichts der USA führte 1966 zur Entlassung von 967 psychisch kranken Straftätern, die nach Verbüßen ihrer Haftstrafe unter Verweis auf das Vorliegen einer psychischen Störung als hochgefährlich beurteilt worden waren und daraufhin eine Sicherungsverwahrung für sie angeordnet wurde ("Baxstrom v Herold," 1966). Das Gericht entschied, dass es als verfassungswidrig gelte, Straftäter ausschließlich aufgrund einer weiterhin vorliegenden psychischen Störung zu verwahren, ohne dass deren Gefährlichkeit nach Ablauf der Haftstrafe erneut überprüft würde ("Baxstrom v Herold," 1966). Die Entlassung dieser Straftäter ermöglichte eine außergewöhnliche Untersuchung der Validität dieser Risikobeurteilungen, die zur Anordnung der Sicherungsverwahrung geführt hatten. In den vier Jahren nach der Entlassung wurden nur 21% der ursprünglich als ‚Hochrisikopopulation‘ bezeichneten psychisch kranken Straftätergruppe mit gewalttätigen Übergriffen rückfällig. Die Wiederverurteilungsrate für ein Gewaltdelikt lag sogar nur bei 2% (Steadman & Cocozza, 1974). Die Güte der ‚Gefährlichkeits‘-Beurteilung der Psychiater, die für die Empfehlung der Verwahrung in diesen Fällen verantwortlich waren, war ernüchternd. Die Diagnose einer psychischen Störung als alleiniges Kriterium zur Risikobeurteilung heranzuziehen, erwies sich als ungeeignet.

In der großangelegten ‚Bridgewater‘-Katamnesestudie bezogen Kozol, Boucher, and Garofalo (1972) über die Diagnose einer psychiatrischen Störung hinaus zwar weitere psychiatrische Merkmale (wie zum Beispiel ‚schwach ausgeprägte Empathie‘) und Tat- und

Opfermerkmale in die Beurteilung von ‚Gefährlichkeit‘ mit ein. Die Operationalisierung und Gewichtung dieser Merkmale erfolgte jedoch unstandardisiert und führte im Ergebnis gemessen an der Rückfälligkeit ebenso zu einer Fehl- und Überschätzung der ‚Gefährlichkeit‘ – spezifisch in etwa zwei von drei Fällen (Kozol et al., 1972).

Aus den beiden exemplarisch dargestellten Studien (‚Baxstrom‘ und ‚Bridgewater‘) ergeben sich Hinweise auf die wesentlichen Kritikpunkte an der unstrukturiert-klinischen Methode: Überschätzung der Relevanz von Merkmalen für die Beurteilung des Rückfallrisikos und fehlende Transparenz und Systematik des Entscheidungsprozesses, der zu einer intra- und interpersonellen Instabilität von Risikobeurteilungen führen kann (Andrews & Bonta, 2010, pp. 311-312). Schon Goldberg konnte 1970 im Bereich der Persönlichkeitsdiagnostik aufzeigen, dass Experten zwar durchaus über geeignete theoretische Beurteilungsmodelle verfügen, von diesen aber in der praktischen Umsetzung abzuweichen scheinen, was die Güte der Einschätzungen reduziert („Goldberg-Paradox“; Goldberg, 1970). Darüber hinaus untermauerten Studien von Steadman and Cocozza (1978) und Quinsey and Amtman (1979) die fehlende Spezifität von Experten im Rahmen unstrukturiert-klinischer Risikobeurteilungen. So konnten Steadman and Cocozza (1978) zeigen, dass unter Experten keine Interraterreliabilität bezüglich der Relevanz spezifischer Merkmale für das Rückfallrisiko vorlag, d.h. ganz unterschiedliche Merkmale für das Rückfallrisiko als relevant erachtet wurden: Bis auf die Art des Indexdeliktes korrelierte kein einziges Merkmal signifikant mit der Höhe der Gefährlichkeitsbeurteilung. Quinsey and Amtman (1979) wiesen in einer experimentellen Versuchsanordnung aus, dass sich Experten gegenüber Laien in der Methodik und Güte ihrer Beurteilungen nicht auszeichneten, d.h. dass ihre Beurteilungen durch keine spezifische Vorgehensweise gekennzeichnet waren und sie zu keinen zuverlässigeren Einschätzungen als Laien führten.

Letztlich sind unstrukturierte Beurteilungen auch im forensischen Bereich nicht vor den typischen Urteilsfehlern gefeit, die sich durch Fehler in der menschlichen Informationsverarbeitung beispielsweise durch die Salienz bestimmter Informationen (Hilton, Harris, Rawson, & Beach, 2005) oder kritische Referenzsysteme (Hilton, Carter, Harris, & Sharpe, 2008) ergeben.

### **1.1.2. Die mechanische Methode**

Die Kritik an der unstrukturiert-klinischen Methode, wie sie im vorangegangenen Kapitel exemplarisch dargestellt wurde, führte in den 1970er Jahren zu unterschiedlichen Reaktionen. Während einige Wissenschaftler grundsätzlich in Frage stellten, das Konstrukt der Gefährlichkeit überhaupt zuverlässig beurteilen zu können – ein prominenter Vertreter dieses elementaren Zweifels war Diamond (1974) – wurde von anderen Wissenschaftlern vor allem eine Methodenkritik formuliert. Ein Teil der Methodenkritik bezog sich auf das Design bisheriger Validierungsstudien: Beispielsweise wurde die Validität der für die Beurteilung von Rückfälligkeit herangezogenen Kriterien in Frage gestellt (vgl. Rabkin, 1979) und damit die Frage aufgeworfen, inwiefern die vorliegenden Befunde tatsächlich Aussagen über die Validität von Risikobeurteilungen zulassen. Ein anderer Teil der Methodenkritik bezog sich auf das eigentliche Vorgehen bei der Risikobeurteilung, also die Methode der unstrukturierten Urteilsbildung (vgl. Hanson, 2005; Monahan, 1984; Quinsey, Harris, Rice, & Cormier, 2006). So wurde vorgeschlagen, Risikomerkmale empirisch zu identifizieren und sie in standardisierte Modelle zur Abbildung von Risikopopulationen einfließen zu lassen (vgl. Wenk, Robison, & Smith, 1972). Die mit Rückfälligkeit assoziierten Risikomerkmale sollten in Form von Listen zusammengestellt und für den Beurteilenden in einer klar definierten Weise zur standardisierten Erhebung und Auswertung zur Verfügung stehen, um somit den Einfluss subjektiver

Verzerrungen weitestgehend auszuschließen (Andrews, 1989; Hanson, 2005; Quinsey et al., 2006).

Dies resultierte in der Entwicklung mechanischer Risk-Assessment Instrumente, die sich durch einen standardisierten Katalog mit Rückfälligkeit assoziierter Merkmale sowie vordefinierte Antwortkategorien und Auswertungsstrategien auszeichnen (vgl. Rossegger et al., 2012).

Das heißt, die Zusammenstellung relevanter Merkmale und die Gesamtauswertung der zusammengetragenen Informationen folgen einem invariablen Algorithmus (Latessa, Listwan, & Koetzle, 2013; Quinsey et al., 2006), wobei das Gesamturteil meistens über die Einordnung des zu Beurteilenden in eine von mehreren Risikokategorien (z.B. ‚niedrig‘, ‚moderat‘ oder ‚hoch‘; oder ‚1‘ bis ‚5‘) gefällt wird. Im Zuge der Entwicklung von mechanischen Instrumenten wurde von den Autoren auch von dem eher weitgefassten Begriff der ‚Gefährlichkeit‘ Abstand genommen. Ziel eines Risk-Assessments anhand mechanischer Risk-Assessment Instrumente ist es vielmehr, die Höhe des Risikos für einen Rückfall mit einem ähnlich gelagerten Delikt einzuschätzen (Heilbrun, Douglas, & Yasuhara, 2009).

Einige dieser mechanischen Instrumente weisen eine empirische Entwicklungsgrundlage auf, d.h. Eingang fanden nur jene Merkmale, die in einer konkreten, aber umfangreichen und repräsentativen Stichprobe signifikant mit Rückfälligkeit korrelierten (Quinsey et al., 2006). Kennzeichnend für aktuarische Risk-Assessment Instrumente ist es darüber hinaus, dass zur Interpretation des Ergebnisses spezifische Rückfallwahrscheinlichkeiten pro Risikokategorie angegeben werden, die die Verteilung der Rückfälligen pro Risikokategorie in umfangreichen Normstichproben widerspiegeln (vgl. Rossegger et al., 2012).

Eines der ersten aktuarischen Risk-Assessment Instrumente, das nach der substanziellen Methodenkritik der 1970iger Jahre zur Risikobeurteilung von Gewalt- und Sexualstraftätern entwickelt wurde, ist der Violence Risk Appraisal Guide (VRAG; Quinsey et al., 2006), der bis heute eines der in der Praxis am weitesten verbreiteten Risk-Assessment Instrumente ist (Fazel, Singh, Doll, & Grann, 2012). Er beinhaltet nur wenige (zwölf) und hauptsächlich statische, d.h. unveränderliche Risikomerkmale. Der überwiegend statische Charakter der Items lässt sich durch das gewählte Design bei der Entwicklung des Instruments erklären: die potenziellen Prädiktorvariablen wurden retrospektiv anhand von Aktenmaterial erhoben, das Informationen bis zum Zeitpunkt des Anlassdelikts enthielt. Der Sanktionsverlauf war nicht Bestandteil der Informationsgrundlage (Latessa et al., 2013; Quinsey et al., 2006). Dieser Art zur Entwicklung aktuarischer Instrumente folgten einige andere verbreitete Risk-Assessment Instrumente wie z.B. der Sex Offender Risk Appraisal Guide (SORAG; Quinsey et al., 2006) und der Static-99 (Hanson & Thornton, 2000) zur Risikobeurteilung bei Sexualstraftätern (Fazel et al., 2012). Obwohl eine Vielzahl von Untersuchungen konsistent zeigen konnte, dass die mechanische im Vergleich zu unstrukturierten, intuitiven Einschätzungen eine im Durchschnitt zuverlässigere Unterscheidung zwischen rückfälligen und nicht rückfälligen Straftätern ermöglicht (Ægisdóttir et al., 2006; Bonta, Law, & Hanson, 1998; Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Hanson & Morton-Bourgon, 2009; Meehl, 1954), waren die wesentlichen Kennzeichen der am weitesten verbreiteten aktuarischen Risk-Assessment Instrumente zugleich Gegenstand neuer Kritik. Kritisiert wurde, dass sie lediglich ein statisches Korrelationsmodell darstellen, das zum einen einer einzelfallspezifischen Risikobeurteilung nicht gerecht werde und zum anderen keine Erfassung dynamischer Prozesse ermögliche (vgl. Falzer, 2013; Guy, 2008).

Dies sowie der ‚atheoretische‘ Ansatz lassen darüber hinaus die Ableitung eines Risikomodells, das Rückschlüsse auf mögliche Präventionsstrategien bieten würde, nicht zu (Bonta & Andrews, 2007; Quinsey et al., 2006).

Dieser Kritik begegnend haben sich innerhalb der mechanischen Methode ebenso Risk-Assessment Instrumente etabliert, die theoriegeleitet entwickelt wurden und dynamische Merkmale explizit aufgreifen. Grundlegend wird davon ausgegangen, dass veränderbare Merkmale einen protektiven oder risikoerhöhenden Einfluss auf das Rückfallrisiko haben können (z.B. Stabilität einer Beziehung oder Trennungssituation), wobei darunter Merkmale der Persönlichkeit oder Situation gefasst werden (z.B. psychische Störungsbilder oder Verlust der Arbeitsstelle), die einen akuten oder eher stabileren Charakter aufweisen können (z.B. Alkoholintoxikation oder Alkoholabhängigkeit) und auch Anhaltspunkte für die Fallführung von Straftätern geben (Bonta & Andrews, 2007). Theoriegeleitete und dynamische Risk-Assessment Instrumente sind eng an das 1990 erstmals formalisierte Risk-Needs-Responsivity Prinzip von Andrews, Bonta, and Hoge (1990) angelehnt, wonach die Art und Intensität von Interventionen gemäß Art und Relevanz der das Rückfallrisiko bestimmenden ‚kriminogenen Bedürfnisse‘ (englisch: criminogenic needs) und der individuellen Ansprechbarkeit des Straftäters auf Interventionsmaßnahmen ausgerichtet werden sollten (Bonta & Andrews, 2007; Latessa et al., 2013). Wesentlicher Unterschied zu den vorangehend dargestellten Beispielen aktuarischer Risk-Assessment Instrumente ist damit, dass der Leitgedanke der Entwicklung nicht ‚schlicht‘ auf die Beurteilung des Rückfallrisikos, sondern theoriegeleitet auf das ‚Risikomanagement‘ fokussiert und Autoren dieser Instrumente weniger eine ‚prädiktionsgeleitete‘ als eine ‚präventionsgeleitete‘ Auffassung vom Risk-Assessment haben (vgl. Mossman, 2013). Beispiele mechanischer Instrumente, die dynamische Risikomerkmale einbeziehen, sind das Level of

Service Inventory Revised (LSI-R; Andrews & Bonta, 2001), der STABLE-2007 (Hanson, Harris, Scott, & Helmus, 2007) und das Forensische Operationalisierte Therapie-Risiko-Evaluations-System (FOTRES; Urbaniok, 2007).

### **1.1.3. Die strukturiert-klinische Methode**

Die strukturiert-klinische Methode (häufig auch ‚strukturierte professionelle Urteilsbildung‘ oder englisch: ‚structured professional judgment‘ [SPJ] genannt) stellt ebenso wie die mechanischen Risk-Assessment Instrumente mit einem dynamischen Charakter die Etablierung eines Präventionsmodells zum Management des Rückfallrisikos in den Fokus eines Risk-Assessment Prozesses (Hart & Logan, 2011). Der grundlegende Unterschied besteht jedoch in der zu diesem Ziel führenden Methode. Während zwar Regeln zur Erhebung der in die Instrumente einbezogenen Risikomerkmale, die theoriegeleitet auf der Grundlage von Literaturlaufarbeitungen und Diskussionen innerhalb von Expertenforen ermittelt wurden (vgl. von Franqué, 2013), vorliegen, wird jedoch kein Auswertungsalgorithmus für die Gesamtbeurteilung vorgegeben. Vielmehr obliegt dem Beurteilenden die Zusammenfügung und Gewichtung der Risikomerkmale zu einem Gesamturteil. Das heißt, die Risikomerkmale können einzelfallspezifisch als mehr oder weniger relevant erachtet und miteinander kombiniert werden (Hart & Logan, 2011). Dabei wird durch Leitlinien ein strukturierter Rahmen für den fallspezifischen Beurteilungsprozess zur Verfügung gestellt (Hart & Logan, 2011). Ziel der SPJ-Instrumente ist es, damit eine Brücke zwischen hochstandardisierten und gänzlich unstrukturierten Beurteilungsmethoden zu schlagen und damit die jeweils kritischen Eigenschaften durch eine Kombination beider auszuloten (Hart & Logan, 2011).

In der Praxis verbreitete Beispiele für diese Instrumente sind der HCR-20, der seit 2013 in seiner dritten überarbeiteten Version vorliegt und zur Risikobeurteilung von erneutem

gewalttätigen Verhalten entwickelt wurde (Douglas, Hart, Webster, & Belfrage, 2013; Webster, Eaves, Douglas, & Wintrup, 1995) und der Spousal Risk Appraisal Guide (SARA; Kropp, Hart, Webster, & Eaves, 1998), ein Instrument zur Risikobeurteilung bei Intimpartnergewalt.

Bezugnehmend auf die Erkenntnisse der Auseinandersetzung mit der unstrukturiert-klinischen Methode stehen SPJ-Instrumente häufig in der Kritik, wissenschaftlich abgestützte Befunde über die Problematik wenig strukturierter Beurteilungen zu ignorieren, da der Beurteilungsprozess anhand von SPJ-Instrumenten keiner standardisierten Gesamtbewertung folgt und damit für subjektive Verzerrungen anfällig ist (Hilton, Harris, & Rice, 2010, p. 31).

## **1.2. Validierung von Risk-Assessment Instrumenten**

Für die Anwendung der Risk-Assessment Instrumente in der Praxis kommt der Frage nach der Validität der Verfahren – an der üblicherweise die Effektivität und Nützlichkeit des Assessments ausgemacht wird – eine besondere Bedeutung zu. Wenngleich es verschiedene Formen der Validität gibt, war primär die Kriteriumsvalidität Gegenstand bisheriger Validierungsstudien forensischer Risk-Assessment Instrumente. Dabei wird üblicherweise Rückfälligkeit als Kriterium definiert.

Wie valide welche Methoden des Risk-Assessments und im Spezifischen einzelne Risk-Assessment Instrumente sind und welche Schlussfolgerungen die empirischen Befunde zulassen, ist aufgrund der unter Kapitel 1.1 dargestellten Methodenunterschiede, aber auch unterschiedlichen Möglichkeiten, die Validität eines Instrumentes überprüfen zu können, Gegenstand einer fortlaufenden wissenschaftlichen Kontroverse (Falzer, 2013). Vor dem Hintergrund der jeweiligen Kritik an den verschiedenen Methoden wurde zur Bestimmung der zuverlässigeren Methode eine Vielzahl von Untersuchungen durchgeführt. Wenngleich es seit den 1950iger Jahren klare Hinweise auf die Unterlegenheit der unstrukturiert-klinischen

Methode bezüglich der Kriteriumsvalidität gibt (Meehl, 1954), weisen aktuelle Arbeiten immer wieder darauf hin, dass Vergleiche zwischen den verschiedenen strukturierten Methoden zu weniger eindeutigen Ergebnissen führen, die die Bevorzugung der einen oder anderen Methode rechtfertigen könnten (vgl. Skeem & Monahan, 2011). Werden Unterschiede berichtet, so zu Gunsten der aktuarischen Methode, wobei diese im Durchschnitt marginal ausfallen (vgl. Falzer, 2013; Guy, 2008) und im Mittel über eine moderate Kriteriumsvalidität nicht hinausgehen (Bowen, 2011; Hanson & Morton-Bourgon, 2009; Kilvinger, Rossegger, Urbaniok, & Endrass, 2012). Fraglich ist bei diesen zusammenfassenden Bewertungen der Kriteriumsvalidität der Instrumente allerdings zum einen, inwiefern Design und Qualität der Primärstudien eine Aussage darüber überhaupt zulassen, d.h. inwiefern Replikationsstudien in ihrem Design die Kernkriterien des Instruments aufgreifen und darüber hinaus wissenschaftlichen Standards von longitudinalen (follow-up) Studien genügen (Rossegger, Gerth, Seewald, et al., 2013). Zum anderen sollte gewährleistet sein, dass der Unterschiedlichkeit der Methoden bezüglich ihres theoretischen Ansatzes und Anwendungskontextes bei der Überprüfung und Interpretation der Validität Rechnung getragen wird (vgl. Falzer, 2013; Guy, 2008).

Strittig ist daher, inwiefern die höhere Kriteriumsvalidität aktuarischer Risk-Assessment Instrumente gegenüber anderen strukturierten Instrumenten eine tatsächliche Überlegenheit dieser Methode widerspiegelt (Falzer, 2013; Swanson, 2008).

Anliegen der vorliegenden Dissertation war es daher, sich unter einem praxisrelevanten Blickwinkel mit methodischen Aspekten der Validität von Risk-Assessment Instrumenten auseinanderzusetzen und in diesem Zusammenhang eine systematische Überprüfung der Qualität bisheriger Replikationsstudien von als gemeinhin valide geltenden aktuarischen Risk-Assessment Instrumenten vorzunehmen und Mindestanforderungen für aussagekräftige

Replikationen zu erarbeiten; das Augenmerk auf Validitätsaspekte von aktuarischen Risk-Assessment Instrumenten zu legen, die weitestgehend übersehen werden, jedoch eine wesentliche praktische Relevanz für die Anwendung der Instrumente aufweisen; die Kriteriumsvalidität von Instrumenten der mechanischen Methode in einem spezifischen Anwendungskontext – nämlich Intimpartnergewalt – zu überprüfen und die Relevanz der strukturiert-klinischen Methode im Kontext der Beurteilung des Ausführungsrisikos von Drohungen darzustellen.

### **1.2.1. Mindestanforderungen an Replikationsstudien**

Um ein Risk-Assessment Instrument als valide bezeichnen zu können, reicht es nicht aus, dass seine Validität im Entwicklungskontext nachgewiesen wurde. Vielmehr sind Replikationen an unabhängigen Stichproben notwendig, die den Erstbefund bestärken. Diese Replikationsstudien sind als Validierung des Instrumentes wiederum nur dann aussagekräftig, wenn sie dem Design der Entwicklungsstudie entsprechen und das Instrument manalkonform angewendet wird. Vor allem bei aktuarischen Risk-Assessment Instrumenten, die aufbauend auf korrelativen Zusammenhängen in einer spezifischen Stichprobe entwickelt wurden und konkrete Rückfallraten zur Interpretation des Assessments zur Verfügung stellen, sind sorgfältige Replikationen in vergleichbaren, aber unabhängigen Stichproben zur Beurteilung der Validität notwendig (Rossegger, Gerth, Seewald, et al., 2013). In den letzten 20 Jahren waren die in der Praxis am häufigsten angewendeten aktuarischen Risk-Assessment Instrumente (vgl. R. P. Archer, Buffington-Vollum, Stredny, & Handel, 2006; Viljoen, McLachlan, & Vincent, 2010) – wie der VRAG (Quinsey et al., 2006), der SORAG (Quinsey et al., 2006) und der Static-99 (Hanson & Thornton, 2000) – Gegenstand einer Vielzahl von Replikationsstudien und einiger Meta-Analysen.

Mit Bezug auf diese Studien wird häufig darauf hingewiesen, dass alle drei Instrumente gemeinhin als valide Verfahren zur Schätzung des Rückfallrisikos gelten. Auffällig ist jedoch, dass die entsprechenden Primärstudien eigentlich eine hohe Heterogenität aufweisen, d.h. bezüglich ihrer Operationalisierung verschiedener Kernkriterien von denen der Entwicklungsstudie abweichen, und die Bedeutung der Ergebnisse als Validitätsnachweis kritisch zu diskutieren ist. Ziel der ersten für die vorliegende Dissertation relevanten Studie war es, diesem Eindruck nachzugehen und alle bis dato publizierten Replikationsstudien zu den Risk-Assessment Instrumenten VRAG, SORAG und Static-99 systematisch zu erfassen. Eine erste Übersicht bestätigte diese Wahrnehmung vorläufig, so dass auf Grundlage einer Literaturrecherche Mindestanforderungen für Replikationsstudien abgeleitet wurden, um folglich die Güte von Replikationsstudien zu gewährleisten.

**Empirische Studie: Current obstacles in replicating risk assessment findings: A systematic review of commonly used actuarial instruments**

Als Ergebnis der Literaturlaufarbeitung wurden gesamthaft zwölf Kriterien zur Beurteilung der wissenschaftlichen Güte von Replikationsstudien definiert: Damit eine Replikationsstudie zur Sicherung der Validität eines Instrumentes als aussagekräftig gelten kann, sollte eine Übereinstimmung zwischen Entwicklungs- und Replikationsstudie bezüglich 1) des Geschlechts, 2) der Altersgruppe (z.B. Jugendliche oder Erwachsene)<sup>1</sup>, 3) des Anlassdeliktes (z.B. Gewalt- oder Sexualdelikt), 4) der Länge des Beobachtungszeitraumes<sup>2</sup> und des Rückfallkriteriums in 5) Art (z.B. Gewalt- oder Sexualdelikt)<sup>3</sup> und 6) rechtlichem Status (z.B. polizeiliche Registrierung

---

<sup>1</sup> Alter ist negativ mit Rückfälligkeit korreliert. Werden deutlich voneinander abweichende Altersgruppen untersucht, ist der Vergleich von Rückfallraten kritisch (Sampson & Laub, 2003)

<sup>2</sup> Werden in der Replikationsstudie andere Zeiträume untersucht, so sind Rückfallraten nicht vergleichbar (Harris & Rice, 2003; Harris et al., 2003; Quinsey et al., 2006).

<sup>3</sup> Bei einer fehlenden Übereinstimmung des Rückfallkriteriums muss davon ausgegangen werden, dass unterschiedliche Phänomene erhoben werden, die zudem die Basisrate deutlich beeinflussen können.

oder Verurteilung) vorliegen. Es ist darüber hinaus erforderlich, dass 7) Akteninformationen herangezogen werden<sup>4</sup>, 8) reliables Werten garantiert ist (durch die Schulung der Anwender oder einen Nachweis der Interrater-Reliabilität), 9) keine Itemwertungen angepasst und 10) keine Items systematisch ausgelassen werden, 11) auf Stichprobenschwund (durch Inhaftierungen, Abschiebung, Tod oder Namensänderung) kontrolliert<sup>5</sup> und 12) möglichst ein fixer Beobachtungszeitraum bestimmt wird.

Um die Replikationsgüte der bis Juni 2012 publizierten Replikationsstudien des VRAG ( $k = 38$ ), SORAG ( $k = 21$ ) und Static-99 ( $k = 49$ ) systematisch zu ermitteln, wurde die Erfüllung der formulierten Mindestanforderungen anhand einer standardisierten Erhebung geprüft. Dabei zeigte sich, dass durchschnittlich etwas mehr als die Hälfte ( $M = 6.6$ ) der zwölf Mindestanforderungen erfüllt wurden, jedoch keine der Studien allen Anforderungen gerecht wurde. Besonders kritisch für die Aussagekraft der Studien als Replikation ist, dass über die Hälfte der Studien (56%) Straftäter einschlossen, die jünger als 18 Jahre alt waren, wobei die Instrumente an ausschließlich erwachsenen Straftätern entwickelt worden waren. Bei einem Fünftel der Studien (21%) lagen keine Information zum Geschlecht der Straftäter vor, bei nur einem Drittel der Studien wurde auf Stichprobenschwund kontrolliert (32%) oder der rechtliche Status des Deliktes berücksichtigt (34%). Nur in knapp mehr als der Hälfte der Studien (58%) wurde eine gute Interrater-Reliabilität oder die Schulung der Beurteiler ausgewiesen.

Auf Grundlage dieser Ergebnisse muss kritisch diskutiert werden, inwiefern die bisherigen Replikationsstudien überhaupt aussagekräftig für die Validität der Instrumente sind. Es muss angezweifelt werden, ob die Robustheit des Risikomodells, wie es in der Entwicklungsstudie herausgearbeitet wurde, anhand von Studien nachgewiesen werden kann, die von den

---

<sup>4</sup> Allein auf Selbstberichte zurückzugreifen genügt nicht, da deren Maß an Objektivität als zu gering angesehen werden muss.

<sup>5</sup> Wird nicht auf Stichprobenschwund kontrolliert, kann Rückfälligkeit unterschätzt werden (Harris & Rice, 2007).

Kernmerkmalen des Modells abweichen. Konkret kann diese Problematik an zwei Beispielen dargestellt werden:

1) Würden ein oder mehrere Items des Instrumentes in modifizierter Form erhoben, so würde dies zu einer systematischen Anpassung der in das Risikomodell einfließenden Risikofaktoren führen. Eine sich im Rahmen einer Validierung möglicherweise ergebende hohe Validität würde dann zwar die Güte des neuen Modells, aber keine Bestätigung der Validität des ursprünglichen Entwicklungsmodells darstellen.

2) Wird im Rahmen einer Validierung ein signifikanter Zusammenhang zwischen dem Summenwert des Instrumentes und dem Rückfallkriterium ausgewiesen, wenn sich gleichzeitig die Definitionen des Rückfallkriteriums jedoch deutlich zwischen den Studien unterscheiden, so wäre die Schlussfolgerung, es handele sich um einen Nachweis der Validität des Instrumentes, nicht plausibel. Auf Grundlage eines solchen Befundes stellt sich eher die Frage, was genau das Instrument eigentlich misst und inwiefern eine praktische Relevanz für dessen Anwendung bei der konkreten Fragestellung nach dem Rückfallrisiko gegeben ist, wenn es mit ganz unterschiedlichen Facetten devianten Verhaltens korreliert.

Obwohl die Validität der weitverbreitet angewendeten Risk-Assessment Instrumente VRAG, SORAG und Static-99 mit Bezug auf eine Vielzahl von Replikationsstudien in der Literatur kaum in Frage gestellt wird, lässt die mangelnde Güte dieser Studien als ‚wahre‘ Replikationen an der Robustheit des Befundes zweifeln. Dies ist in Abweichungen von zentralen Kriterien der Entwicklungsstudie, im Nichteinhalten von wissenschaftlichen Standards oder in der lückenhaften Dokumentation des Studiendesigns begründet, das sich damit einer Überprüfbarkeit entzieht.

### **1.2.2. Methodische Aspekte zur Untersuchung der Kriteriumsvalidität**

Wie präzise das Rückfallrisiko anhand eines Instrumentes geschätzt wird, kann mit zwei Teilaspekten der Kriteriumsvalidität ausgewiesen werden: Erstens mit der Trennschärfe des Instrumentes, die sich im aktuellen Kontext auf die Fähigkeit, zwischen rückfälligen und nicht rückfälligen Straftätern zuverlässig diskriminieren zu können, bezieht. Zweitens mit der Kalibrierung des Instrumentes, die eine Aussage über die Übereinstimmung der erwarteten Rückfallwahrscheinlichkeiten innerhalb verschiedener Risikokategorien (die sich z.B. über die Kombination zutreffender Risikofaktoren ergeben) und den in der Replikationsstichprobe tatsächlich beobachteten Risikoraten trifft (Falzer, 2013; Rossegger, Gerth, Singh, & Endrass, 2013).

#### **Trennschärfe**

Zur Überprüfung der Trennschärfe hat sich das Verfahren der Receiver Operating Characteristic (*ROC*) als Standardmethode etabliert. Als basisratenunabhängiges Verfahren hat es Vorteile gegenüber anderen Methoden wie zum Beispiel der punktbiserale Korrelation zwischen dem Ergebnis eines Risk-Assessment Instrumentes und dem Kriterium (Fawcett, 2006; Mossman, 2013), da Rückfallraten mit Gewalt- und Sexualdelikten (innerhalb eines mittleren Beobachtungszeitraumes von fünf bis sechs Jahren) bei durchschnittlich 14% bezüglich erneuter Sexualdelikte und 25% bezüglich erneuter Gewalt- (inkl. Sexual-)delikte liegen, ein selten auftretendes Ereignis ist (Hanson, 2005).

Die *ROC* kann als Funktion der Spezifität und Sensitivität eines Instruments über die möglichen cut-off-Werte des Entscheidungskriteriums (z.B. Itemsummenwert des Instruments) hinweg betrachtet werden, wobei die unter ihrem Funktionsgraphen erfasste Fläche (area under the curve [*AUC*]) das Effektmaß für die Trennschärfe darstellt. Konkret erfasst die *ROC* die

Wahrscheinlichkeit, mit der bei einem zufällig gewählten Rückfälligen anhand des Instruments auch ein höheres Rückfallrisiko ermittelt wurde als bei einem zufällig gewählten nicht rückfälligen Straftäter (Fawcett, 2006). D.h., die *AUC* gibt den Anteil aller möglichen Zufallspaare von Rückfälligen und Nicht-Rückfälligen an, bei denen die Klassifizierung im obigen Sinn korrekt vorgenommen wurde, wobei eine *AUC* von 1.0 einer perfekten Trennschärfe und eine *AUC* von .50 einem Zufallsbefund entspricht (Mossman, 2013). Die praktische Relevanz des Ergebnisses einer *ROC*-Analyse ist auf die Aussage zur Diskriminationsfähigkeit des Instrumentes hinaus begrenzt. Inwiefern die erwarteten Rückfallraten innerhalb der Risikokategorien eines Instrumentes mit den tatsächlich beobachteten in der Replikationsstichprobe übereinstimmen, ist eine Frage der Kalibrierung eines Instrumentes (Schmid & Griffith, 2005).

### **Kalibrierung**

Explizite, verschiedenen Risikokategorien hinterlegte Risikoraten sind eine praxisrelevante Stärke aktuarischer Risk-Assessment Instrumente. Zwar können Assessment Ergebnisse anderer Generationen häufig ebenso einer Risikokategorie zugeordnet werden, jedoch weisen diese dann meist einen non-numerischen Charakter (z.B. ‚niedriges‘, ‚moderates‘ oder ‚hohes‘ Risiko) auf. Aus der Literatur ist bekannt, dass non-numerische Risikokategorien jedoch einen hohen Interpretationsspielraum zulassen. Einerseits wird eine Vielzahl unterschiedlicher Begriffe für dieselbe Risikoeinschätzung verwendet (Grann & Pallvik, 2002), andererseits zeigt sich eine starke Heterogenität in der wahrgenommenen Bedeutung dieser non-numerischen Risikokategorien (Hilton, Carter, et al., 2008). Gefragt nach der probabilistischen Spezifizierung der zwei Grenzübergänge zwischen einem ‚niedrigen‘, ‚moderaten‘ und ‚hohen‘ Rückfallrisiko für ein Gewaltdelikt innerhalb von zehn Jahren, variierten die angegebenen

Wahrscheinlichkeiten in einer Untersuchung von Hilton, Carter, et al. (2008) mit 60 klinisch-forensischen Experten zwischen 8% und 54% bzw. 38% und 95%. Es wird deutlich, wie heterogen non-numerische Risikobegriffe verstanden werden und zwischen verschiedenen Experten zu substantiell anderen Interpretationen führen können. Daraus ergibt sich die praktische Relevanz von probabilistischen Normwerten zur Kommunikation eines Rückfallrisikos. Die Überprüfung beobachteter und erwarteter Rückfallraten bleibt jedoch im Rahmen der meisten Replikationsstudien aktuarischer Instrumente aus (Rossegger, Endrass, Gerth, & Singh, 2014; Rossegger, Gerth, Singh, et al., 2013), weshalb noch weitgehend Unklarheit darüber besteht, ob dieser Vorteil aktuarischer Risk-Assessment Instrumente auch praktische Relevanz hat und die Risikokommunikation bei der Fallbearbeitung darauf abgestützt werden kann.

Aus diesem Grund lag der Fokus der zweiten für die vorliegende Dissertation relevanten Studie darauf, die Übereinstimmung von Risikonormen eines in Kanada entwickelten aktuarischen Risk-Assessment Instrumentes mit den risikokategorienspezifischen Rückfallraten einer Zürcher Stichprobe von Sexualstraftätern zu überprüfen.

### **Empirische Studie: Examining the predictive validity of the SORAG in Switzerland**

Der SORAG ist ein aktuarisches Risk-Assessment Instrument, welches zur Risikobeurteilung erneuter hands-on Gewalt- und Sexualdelikte bei hands-on Sexualstraftätern 1998 von der Forschungsgruppe um Vernon L. Quinsey entwickelt wurde (Quinsey et al., 2006). Er setzt sich aus vierzehn Items zusammen, wobei diese zu einem Summenwert addiert werden und dieser wiederum einer von neun Risikokategorien zugeordnet werden kann. Den Risikokategorien sind jeweils Risikonormen hinterlegt, wonach die Rückfallwahrscheinlichkeit mit der Höhe der Risikokategorie positiv korreliert (Quinsey et al., 2006). Obwohl in den letzten dreizehn Jahren

eine Vielzahl von Studien zur Validierung des SORAG publiziert wurden, wiesen nur vier Studien beobachtete Rückfallraten pro Risikokategorie zumindest deskriptiv aus, wobei keine einzige ihre Übereinstimmung mit den erwarteten Raten inferenzstatistisch überprüfte.

In einer Gesamtstichprobe von 137 Sexualstraftätern, die sich aus zwei Substichproben zusammensetzte, einerseits aus einer Substichprobe aller im Jahr 2000 im Schweizer Kanton Zürich registrierten Gewalt- und Sexualstraftäter, die ein Strafmaß von mindestens 10 Monaten oder eine gerichtliche Therapieanordnung erhielten, und andererseits aus Sexualstraftätern, die zwischen 1997 und 2009 eine Therapie beim Psychiatrisch-Psychologischen Dienst des Kantons Zürich aufnahmen, wurden neben der Trennschärfeanalyse verschiedene Berechnungen zur Kalibrierung des SORAG durchgeführt. Dabei wurde explizit darauf geachtet, ein aussagekräftiges Studiendesign über die Erfüllung aller zwölf Mindestanforderungen an Replikationsstudien zu realisieren.

Die Analysen zur Kalibrierung des SORAG wiesen auf eine deutliche Abweichung von durchschnittlich 21% über alle Risikokategorien hinweg hin. Dies zeigte sich im Spezifischen in der signifikanten Abweichung der Likelihood ratios in fünf der neun Risikokategorien. Eine Bayes'sche basisratenabhängige Anpassung der Risikonormen führte zu einer zwar substantiellen Verbesserung (8%), aber immer noch zu einer unzureichenden Übereinstimmung der Rückfallraten zwischen Entwicklungs- und der Zürcher Validierungsstichprobe. Detaillierte Analysen weiterer Stichprobenkennwerte wiesen entsprechend auf einen signifikanten Unterschied bezüglich Summenwert- und Risikokategorienverteilung zwischen den beiden Straftäterstichproben hin ( $t(136) = -5.54, p < .001$  bzw.  $D$  [Kolomogorov-Smirnoff-Teststatistik] = 0.25,  $p < .001$ ).

Zusammengefasst zeigt sich, dass die Verwendbarkeit der derzeit publizierten Risikonormen des SORAG kritisch scheint. Dieses Ergebnis steht im Einklang mit wenigen anderen Replikationsstudien, die eine Überprüfung der Kalibrierung des aktuarischen Risk-Assessment Instruments VRAG (Quinsey et al. 2006) vorgenommen haben (vgl. Rossegger, 2014).

### **1.3. Validität von Risk-Assessment Instrumenten bei Intimpartnergewalt**

Ca. ein Drittel aller Frauen ist laut des 2013 erschienenen Berichts der Weltgesundheitsorganisation weltweit von Intimpartnergewalt betroffen (World Health Organisation, 2013). Andere, regionale Umfragen zusammenfassende Studien schätzen das Ausmaß an Betroffenen unter Männern ähnlich hoch ein (J. Archer, 2002; Straus, 2009). Die Konsequenzen der Gewalt sind für die weiblichen Opfer jedoch im Durchschnitt schwerer (J. Archer, 2000; Greenfeld et al., 1998; Straus, 2009; Swan, Gambone, Caldwell, Sullivan, & Snow, 2008; Tjaden & Thoennes, 2000). Auf Ebene der Anzeigestatistiken zeigt sich, dass zwischen ca. 80% und 90% aller Täter polizeilich registrierter Vorfälle von Intimpartnergewalt männlich sind (z.B. Bundesamt für Statistik Schweiz, 2014; Melton & Sillito, 2012). Ähnlich wie beim Risk-Assessment in anderen Bereichen der Gewaltdelinquenz wurden daher auch bei Intimpartnergewalt hauptsächlich Instrumente zur Risikobeurteilung von männlichen Tätern entwickelt. Seit im Verlauf der 1980iger Jahre in verschiedenen Ländern wie den USA, Kanada und Australien ‚pro-arrest‘- und ‚pro-charging‘-Strategien eingeführt wurden, erlangte die Polizei zunehmenden Handlungsspielraum im Umgang mit Fällen von Intimpartnergewalt (Tutty et al., 2008). Weitere Entwicklungen, wie z.B. die Erklärung der Generalversammlung der Vereinten Nationen (1993) zur ‚Beseitigung von Gewalt gegen Frauen‘, führten in Österreich als erstem europäischen Land 1997 zum Inkrafttreten eines sogenannten Gewaltschutzgesetzes, welches die rechtlichen Voraussetzungen zum Schutz von Betroffenen häuslicher Gewalt schuf

(Nationalrat Republik Österreich, 1997). Auch in Deutschland traten 2002 auf Bundesebene und in der Schweiz auf Kantonsebene – z.B. im Kanton Zürich (2007) – Gewaltschutzgesetze in Kraft, die Gewalt im häuslichen Bereich als Offizialdelikt fassen und gleichzeitig in den meisten Fällen den polizeilichen Verantwortungsbereich in der praktischen Umsetzung durch die Anpassung des Polizeigesetzes vergrößern, indem beispielsweise gefährdende Personen unabhängig von einem Haftbefehl zunächst in Gewahrsam genommen und Schutzmaßnahmen wie mehrwöchige Kontakt- oder Rayonverbote ausgesprochen werden können (Bundestag Bundesrepublik Deutschland, 2001; Kantonsrat Kanton Zürich Schweiz, 2006). Im Zuge dieser Entwicklungen wuchs die Notwendigkeit zuverlässiger Risk-Assessment Strategien als Entscheidungsgrundlage zur Implementierung der polizeilichen Interventionsmaßnahmen, wobei in diesem Kontext besondere Anforderungen an das Risk-Assessment gestellt werden: es sollte sich um ein sensibles und in der Anwendung einfaches, d.h. ökonomisches und auf leicht zugänglichen Informationen basierendes Instrument handeln, um eine systematische und schnelle Triagierung zwischen Tätern mit niedrigem und jenen mit hohem Rückfallrisiko, die vertieft abgeklärt werden sollten, zu ermöglichen (Hilton, Harris, & Rice, 2010). Nicht alle der bis heute entwickelten Risk-Assessment Instrumente für Intimpartnergewalt sind für das sogenannte frontline-Assessment geeignet, da sie vom Anwender z.B. psychiatrische Vorkenntnisse verlangen, die Interpretation des Assessment Ergebnisses offen lassen oder Zugriff auf eine umfangreiche Informationsgrundlage benötigen. Unter den infrage kommenden deuten erste Befunde vorsichtig darauf hin, dass das Ontario Domestic Assault Risk Assessment eines der bisher am besten validierten Instrumente ist, wonach es in einer aktuellen Studie von Messing & Thaller (2013) im Durchschnitt eine moderate, aber anderen Instrumenten überlegene Trennschärfe von  $AUC = .67$  aufweist. Das ODARA wurde 2004 in Kanada entwickelt und ist

ein aktuarisches, dreizehn dichotome Items beinhaltendes Screening-Instrument. Die Summe aller zutreffenden Items kann einer von sieben Risikokategorien zugeordnet werden. Von den bis Juni 2014 anhand einer systematisch durchgeführten Literaturrecherche identifizierten Studien, die verschiedene Formen der Validität des ODARA untersuchten (z.B. Kriteriums-, Konstruktvalidität oder inkrementelle Validität), beschäftigten sich fünf mit der Trennschärfe des ODARA. *AUC*-Werte zwischen .64 (Hilton, Harris, Popham, & Lang, 2010) und .74 (Hilton & Harris, 2009) wiesen eine moderate bis gute Trennschärfe aus. Im Schweizer Kanton Zürich wurde im Rahmen einer Evaluation des Gewaltschutzgesetzes (Endrass, Rossegger, 2012) die Anwendung eines standardisierten Risk-Assessment Instrumentes zur Optimierung des polizeilichen Triagierungs-Prozesses empfohlen. In Bezug auf die oben in Kürze dargestellten Befunde einer Literaturrecherche wurde das ODARA zur Implementierung im Kanton Zürich vorgesehen – vorbehaltlich seiner Validierung, da Replikationsstudien, die unabhängig von den Autoren des Instrumentes durchgeführt wurden, bisher noch ausstanden. In diesem Zusammenhang war es Ziel der dritten und vierten für die vorliegende Dissertation relevanten Studien, zunächst eine autorisierte deutsche Übersetzung des ODARA für die verbreitete Anwendung im deutschsprachigen Raum anzufertigen und nachfolgend die Trennschärfe und Kalibrierung des Instrumentes an einer Zürcher Stichprobe zu überprüfen.

### **1.3.1. Übersetzung des Ontario Domestic Assault Risk Assessment (ODARA)**

Um eine optimale Voraussetzung für die manuellkonforme Anwendung des ODARA im deutschsprachigen Raum zu schaffen, wurde das Instrument wissenschaftlich übersetzt.

#### **Übersichtsstudie und Übersetzung: Das Ontario Domestic Assault Risk Assessment (ODARA) – Validität und autorisierte deutsche Übersetzung eines Screening-Instrumentes für Risikobeurteilungen bei Intimpartnergewalt**

Die wissenschaftliche Übersetzung folgte einem standardisierten Vorgehen, bei dem die englische Originalversion zunächst von der Erst- und Zweitautorin der Publikation in die deutsche Sprache übersetzt wurde. Anschließend erfolgte eine Rückübersetzung durch zwei die deutsche Sprache beherrschende englische Muttersprachler, denen die Items des ODARA nicht bekannt waren. Diese englische Version wurde den Autoren der Originalversion vorgelegt und auf Unverständlichkeiten oder inhaltliche Abweichungen überprüft. Eine einmalige Wiederholung dieses Vorgehens war aufgrund hilfreicher Anmerkungen der Autoren, die zur Anpassung einiger Passagen in der deutschen Version führten, notwendig.

Die deutsche Übersetzung des ODARA umfasst folgende dreizehn Items: 1) Früherer häuslicher Vorfall, 2) Früherer nicht-häuslicher Vorfall, 3) Frühere Freiheitsstrafe von 30 Tagen oder mehr, 4) Versagen bei früherer bedingter Entlassung, 5) Drohung, beim Index-Übergriff zu verletzen oder zu töten, 6) Einsperren der Partnerin beim Index-Übergriff, 7) Besorgnis des Opfers, 8) Mehr als ein Kind, 9) Leibliches Kind des Opfers von einem früheren Partner, 10) Gewalt gegen andere, 11) Substanzmittelmissbrauch, 12) Übergriff auf das Opfer während der Schwangerschaft und 13) Barrieren der Opferunterstützung.

Mit dieser autorisierten deutschen Übersetzung des ODARA steht nun für den deutschsprachigen Raum eine Version zur Verfügung, die den Einsatz des ODARA in der Praxis und Forschung vereinfacht und gemäß den Ansprüchen der Originalversion vereinheitlicht.

### **1.3.2. Kriteriumsvalidität des Ontario Domestic Assault Risk Assessment**

#### **(ODARA)**

Da die einzige bisher in Europa, und zwar in Österreich, durchgeführte Validierungsstudie (Rettenberger & Eher, 2013) ein substantiell von der Entwicklungsstudie abweichendes

Studiendesign aufwies und ihr Nutzen als Replikationsstudie daher fraglich ist, war es Ziel der vierten für die vorliegende Dissertation relevanten Studie, eine umfassende Prüfung der Kriteriumsvalidität des ODARA vorzunehmen, um seine Eignung für die Anwendung als Screening-Instrument bei Intimpartnergewalt im Kanton Zürich zu testen. Dabei wurde besonderer Wert auf die Berücksichtigung der unter Kapitel 1.2 vorgestellten methodischen Aspekte der Validierung und konzeptionellen Aspekte der Validität von Risk-Assessment Instrumenten gelegt.

### **Empirische Studie: Assessing the discrimination and calibration of the Ontario**

#### **Domestic Assault Risk Assessment in Switzerland**

Die Untersuchungsstichprobe setzte sich aus 185 aller 2008 bei der Stadtpolizei Zürich registrierten häuslichen Gewaltvorfälle zusammen, die die ODARA Einschlusskriterien erfüllten (d.h. männlich und mindestens 18jährig waren, einen physischen Übergriff gegenüber ihrer (Ex-)Intimpartnerin verübt hatten oder sie mit einer Waffe in der Hand in ihrer Anwesenheit mit dem Tod bedroht hatten und für die maximal fünf ODARA-Items aufgrund unzureichender Informationsgrundlage nicht gewertet werden konnten) und einen fixen fünfjährigen Beobachtungszeitraum erreichten. Der durchschnittliche ODARA Summen- und Kategorienwert war mit  $M = 5.1$ . bzw.  $M = 5.5$  in dieser Stichprobe signifikant höher als in der Entwicklungsstichprobe ( $M = 2.9$ ,  $p = .000$ ). Innerhalb eines fixen Beobachtungszeitraumes von fünf Jahren lag die Rückfallrate für erneute polizeilich registrierte physische Übergriffe oder Drohungen gegenüber einer (Ex-)Intimpartnerin bei 32% – eine Basisrate, die mit der Entwicklungsstichprobe vergleichbar ist (30%). Mit einer  $AUC$  von .63 ( $p = .004$ ) wies das ODARA in der Zürcher Stichprobe eine zwar signifikante, aber höchstens knapp moderate und im Vergleich zu anderen Replikationsstudien die geringste Trennschärfe auf. Bezüglich der

Rückfallraten lag über alle ODARA Kategorien hinweg ein durchschnittlicher Fehler von 21% pro Risikokategorie vor. Dazu trugen vor allem die zwei höchsten Kategorien bei, in denen die Rückfallraten signifikant geringer ausfielen. Zusammenfassend stellt die Zürcher Stichprobe gemäß ODARA-Modell aufgrund des hohen durchschnittlichen Summenwerts eine Hoch-Risikopopulation dar. Diese Auffassung spiegelt sich jedoch nicht in einer erhöhten Rückfallrate wider, d.h. ein substantieller Anteil von Tätern der zwei höchsten Risikokategorien (durchschnittlich 60%) wird letztlich gar nicht rückfällig. Dies kann mehrere Gründe haben: Erstens könnte die hohe Interventionsquote (gegenüber beinahe allen Tätern [99%] wurden Schutzmaßnahmen ausgesprochen, ein vergleichsweise hoher Anteil [76%] wurde darüber hinaus im Zuge des Indexdeliktes festgenommen und über ein Viertel [28%] inhaftiert) einen Einfluss auf die Rückfallraten gehabt haben und damit die vergleichsweise hohen ODARA-Werte bei durchschnittlich vergleichbarer Rückfallrate erklären.

Zweitens ist es ebenso plausibel anzunehmen, dass Intimpartnergewalt vor allem bei spezifischen Tätersubgruppen wie beispielsweise beim ausschließlich gegenüber der (Ex-)Intimpartnerin gewalttätigen oder beim psychiatrisch auffälligen Tätern (vgl. Tätertypologien gemäß Holtzworth-Munroe, Meehan, Herron, Rehman, & Stuart, 2003) besonders stark dynamischen Prozessen unterliegt, die das Rückfallrisiko substantiell modifizieren können (Birdsey & Snowball, 2013; European union agency for fundamental rights, 2014; Rennison & Welchans, 2000; Tjaden & Thoennes, 2000).

Letztlich könnte die Eignung des ODARA-Modells in der Zürcher Stichprobe grundsätzlich in Frage gestellt werden, da nur drei der dreizehn ODARA Items signifikant mit Rückfälligkeit korrelieren. Es ist durchaus vorstellbar, dass systemische und gesellschaftliche Unterschiede

zwischen verschiedenen Ländern die Übertragbarkeit aktuarischer Risikomodelle einschränken (Helmus & Bourgon, 2011).

Zusammengefasst lässt sich festhalten, dass die Eignung des ODARA als Screening-Instrument in einer Schweizer Population polizeilich registrierter Fälle von Intimpartnergewalt momentan fraglich ist. Zwar ist der positive Zusammenhang zwischen ODARA-Kategorie und Rückfälligkeit signifikant, jedoch ist die Diskrimination zwischen Rückfälligen und Nicht-Rückfälligen in den ‚Hochrisiko‘-Kategorien kritisch. Da jedoch die meisten Täter der Kohorte vor allem in diese Kategorien fallen und sich die Rückfallraten in den unteren Kategorien kaum voneinander unterscheiden, ist ein plausibler Grenzwert zur Triagierung auf Grundlage der aktuellen Daten strittig und letztlich eine Frage der Ressourcen und Akzeptanz des Anteils sogenannter „Falsch-Negativer“. Inwiefern der Befund den Anteil aller tatsächlich hoch rückfallgefährdeten Täter widerspiegelt oder eine Überschätzung des Risikos darstellt, ist auf Grundlage der aktuellen Daten nicht nachzuvollziehen und muss Gegenstand weiterer, möglichst prospektiver Untersuchungen sein. Ob der Einbezug dynamischer Faktoren zu einer Erhöhung der Trennschärfe und Spezifität eines Risk-Assessment Instrumentes innerhalb kürzerer Beobachtungszeiträume führt, war hingegen Fragestellung der fünften für die vorliegende Dissertation relevanten Studie – einer Studie zur Validierung des dynamischen Risk-Assessment Instruments DyRiAS (Hoffmann & Glaz-Ocik, 2012).

### **1.3.3. Trennschärfe und Spezifität des Dynamischen Risiko-Analyse-Systems**

#### **(DyRiAS)**

Dynamische Risikobeurteilungen sind durch die Einbeziehung von Risikomerkmale, die im Laufe der Zeit veränderbar sind, gekennzeichnet und greifen damit die Kritik an den aktuell

geläufigen aktuarischen Risk-Assessment Instrumenten auf, als statisches Risikomodell diesen Merkmalen nicht gerecht zu werden (Hart & Logan, 2011). Diese Herangehensweise ist – wie eingangs beschrieben – weniger mit dem Ziel verknüpft, lediglich eine Aussage über das Rückfallrisiko zu treffen, sondern vielmehr damit, Risikomanagementstrategien anhand von proximalen Risikomerkmale zu entwickeln, diese Strategien zu evaluieren und je nach Ausprägung der Risikomerkmale fortlaufend zu adaptieren (Guy, Packer, & Warnken, 2012). Dieser Logik folgend sieht dynamisches Risk-Assessment die Wiederholung von Risikobeurteilungen vor, um mögliche Veränderungen über Re-klassifizierungen erfassen zu können (Falzer, 2013). Denn würden dynamische Risikomerkmale nicht kontinuierlich überprüft, nähmen sie letztlich einen statischen Charakter an (Quinsey et al., 2006).

Entsprechend sollten diese Instrumente bei relativ kurzen Beobachtungszeiträumen als valide gelten (Falzer, 2013; Guy et al., 2012). In der folgenden Studie wurde aus diesem Grund die Trennschärfe des Dynamischen Risiko-Analyse-Systems (DyRiAS; Hoffmann & Glaz-Ocik, 2012) für Rückfälle mit schwerer Gewalt (Körperverletzung, Gefährdung des Lebens und [versuchter] Tötung) innerhalb vier verschiedener Beobachtungszeiträume ermittelt (drei Monate, sechs Monate, ein Jahr und fünf Jahre) und die Spezifität der Hochrisiko-Kategorien des Instruments unter Berücksichtigung angeordneter polizeilicher Interventionen analysiert.

### **Empirische Studie: Assessing the Risk of Severe Intimate Partner Violence: Validating the DyRiAS in Switzerland**

Das DyRiAS (Hoffmann & Glaz-Ocik, 2012) ist ein in Deutschland entwickeltes und 2012 veröffentlichtes Verfahren, das zur Risikobeurteilung von schwerer Gewalt, d.h. tötungsnaher oder tödlicher Gewalt, durch einen Mann gegenüber seiner (Ex-)Intimpartnerin anzuwenden ist.

Dabei kann es sich beim Indexvorfall um einen bereits stattgefundenen physischen Übergriff handeln oder aber auch ausschließlich um eine Androhung dessen. Auf Grundlage von 39 Items, von denen einige statischer und andere dynamischer Natur sind, wird für ein unmittelbares, relativ kurzes Zeitfenster von mehreren Tage bis Monaten eine Risikoeinschätzung anhand der Zuordnung des Täters zu einer von sechs Risikokategorien vorgenommen. Die Items wurden theoriegeleitet anhand einer Literaturlaufarbeitung zum Thema der Intimpartnertötung zusammengestellt und die Auswertung folgt einem hierarchischen Entscheidungsprozess, der Interaktionen zwischen Risiko- und protektiven Merkmalen einbezieht. Die Validität des DyRiAS ist noch nicht ausreichend untersucht. Bisher liegt ausschließlich eine Studie zur konkurrenten Validität des Instrumentes vor, in der eine retrospektive Überprüfung von Männern, die ihre (Ex-)Intimpartnerinnen töteten oder sie zu töten versuchten, vorgenommen wurde und zusammenfassend 82% der Täter der höchsten und zweithöchsten Risikokategorie zugeordnet wurden. Kritisch an dieser Untersuchung ist, dass die Rater vor der DyRiAS-Wertung über die Art des Deliktes und das Ziel der Studie informiert waren (Hoffmann & Glaz-Ocik, 2012).

Im Gegensatz dazu wurde die Unvoreingenommenheit der Rater der vorliegenden Zürcher Studie sichergestellt, indem eine gesamte Jahreskohorte von Männern erhoben wurde, die gegenüber ihren (Ex-)Intimpartnerinnen gewalttätig geworden waren und für die zunächst keine Informationen über Rückfälligkeit vorlagen. Rückfälligkeit wurde zu einem späteren Zeitpunkt unabhängig von den vorherigen Wertungen erfasst. Aufgrund unzureichender Informationsgrundlage konnte das DyRiAS nicht für alle Täter der Kohorte gewertet werden, so dass sich die Untersuchung auf 174 Männer, die 2008 bei der Zürcher Stadtpolizei wegen Intimpartnergewalt registriert wurden, bezog. Keiner der Täter wurde den Risikokategorien 0

---

und 1 zugeordnet, Mittel- und Modalwert lagen bei der dritten Kategorie und 4.0% der Stichprobe wurden der höchsten, d.h. Kategorie 5, zugeordnet.

Nach weiterem Ausschluss von Tätern, die in ihr Heimatland ausgewiesen, zwischenzeitlich inhaftiert oder verstorben waren und damit die für die Überprüfung der Trennschärfe des DyRiAS relevante Mindestdauer des Beobachtungszeitraumes nicht erreichten, ergaben sich vier Substichproben (drei Monate:  $n = 168$ , sechs Monate:  $n = 167$ , ein Jahr:  $n = 166$ , und fünf Jahre:  $n = 146$ ). Die Basisrate von Rückfälligkeit mit einem schweren Gewaltdelikt war sehr gering und lag je nach Beobachtungszeitraum zwischen 0.6% und 8.9% (drei Monate bzw. fünf Jahre). Kein Täter, der der Kategorie 5 zugeordnet wurde, wurde rückfällig und kein Nicht-Rückfälliger wurde als Niedrig-Risiko-Täter erfasst. Zwar ergaben sich moderate bis gute Trennschärfekoeffizienten über die vier Substichproben hinweg ( $.64 \leq AUC \leq .85$ ), jedoch wurde keine davon signifikant, d.h. in keiner der Stichproben fiel die Trennschärfe des DyRiAS signifikant besser als aus, es durch eine zufällige Zuteilung möglich gewesen wäre. Unter Berücksichtigung des dynamischen Charakters des DyRiAS und der Intensität von polizeilichen Interventionen stellte sich die Frage nach der Spezifität des Instrumentes in der Substichprobe des dreimonatigen Beobachtungszeitraumes, wobei zu erwarten wäre, dass Täter, die als Hoch-Risiko-Täter eingeschätzt wurden, aber keine oder nur ein geringes Ausmaß an Intervention erhielten, eher rückfällig würden und sich dies in den Rückfallraten der entsprechenden Risikokategorien widerspiegeln müsste. Während alle Nichtrückfälligen der Risikokategorie 5 festgenommen, inhaftiert, an die Staatsanwaltschaft vermittelt wurden und eine Schutzmaßnahme erhielten, war das nur bei 82.1% der zweithöchsten Kategorie der Fall. Im Gegenteil, keiner derjenigen der zweithöchsten Risikokategorie, die weder festgenommen, noch inhaftiert wurden, wurde rückfällig. Das DyRiAS weist daher auch für einen sehr kurzen

Zeitraum nach Deliktbegehung und unter Berücksichtigung polizeilicher Interventionen eine geringe Spezifität auf. Für die Praxis bedeutet dies, dass von einer starken Überschätzung des Rückfallrisikos anhand des DyRiAS-Ergebnisses ausgegangen werden muss. Das wiederum wirkt sich auf die Verhältnismäßigkeit von Management-Strategien aus, würde man diese an der Höhe der zugeordneten Risikokategorie ausrichten. Es muss daher kritisch hinterfragt werden, inwiefern die im Instrument einbezogenen Merkmale und der hinterlegte Auswertungsalgorithmus tatsächlich das aktuelle Risiko für schwere Gewalt abbilden.

#### **1.4. Assessment-Strategien bei Drohungen**

Unter Drohungen werden Äußerungen über den Wunsch oder die Absicht einer Person verstanden, eine andere Person zu schädigen und ihre körperliche Unversehrtheit zu verletzen (Meloy, 2001). Risikobeurteilungen, die im Drohungskontext durchgeführt werden, unterscheiden sich von Beurteilungen zur Schätzung des Rückfallrisikos dahingehend, dass keine Aussage über die Wahrscheinlichkeit einer Tatwiederholung sondern über die Wahrscheinlichkeit der Ausführung einer angedrohten Tat getroffen wird. Die überwiegende Mehrheit von Drohungen wird gegenüber gut bekannten Personen der drohenden Person ausgesprochen (Meloy, 2001). Warren, Mullen, and Ogloff (2011) wiesen innerhalb einer klinischen Stichprobe von Drohenden aus, dass in über zwei Dritteln der Fälle Personen aus dem näheren Umkreis des Drohenden und in jedem dritten Fall der Intimpartner von der Drohung betroffen waren.

Im Kontext von Intimpartnergewalt kommt der Beurteilung der Ausführungsgefahr von Drohungen in der Praxis eine hohe Relevanz zu. Nicht immer ist es schon zu einem physischen Übergriff gekommen, wenn sich Opfer von Intimpartnergewalt an die Polizei wenden. Manchmal liegen ausschließlich verbale (oder auch schriftliche) Drohungen vor, die die Opfer

veranlassen, Hilfe zu suchen. Dabei muss davon ausgegangen werden, dass der prädiktive Wert der Drohung als solche gering ist, denn die Mehrheit der Drohungen wird nicht umgesetzt, vor allem, wenn es um die Androhung schwerer Gewalt geht: Bisherige empirische Studien, die den weiteren Verlauf von Fällen, in denen gedroht wurde, analysierten, wiesen aus, dass nur 4% der Drohungen gegenüber Justizbeamten ausgeführt wurden (Calhoun, 1998) und einem Drittel von Todesdrohungen später zumindest eine Gewalttat folgte, wobei nur in 10% der Fälle die ursprünglich bedrohte Person betroffen war (Warren et al., 2011). Spezifisch für Todesdrohungen stellten Warren et al. (2011; 2008) fest, dass auf Todesdrohungen folgende (versuchte) Tötungsdelikte sogar nur in 1-3% der Fälle registriert wurden. Der Leitgedanke des Drohungsassessments besteht in der Identifikation von Risikomerkmalen, an die ein Risikomanagement knüpfen und damit eine Eskalation der Situation und Gefährdung der bedrohten Person verhindert werden kann (Meloy, Hart, & Hoffmann, 2014). Dynamische, d.h. situative und Verhaltensmerkmale weisen dabei eine besondere Bedeutung auf, da häufig das Risiko einer unmittelbaren Drohungsausführung Gegenstand des Assessments ist. Darüber hinaus lassen sich aber auch personenbezogene eher statische Merkmale erkennen, die eine grundlegende Risikodisposition mitbestimmen (McNiel & Binder, 1989; Meloy, Hoffmann, Guldemann, & James, 2012; Mullen et al., 2009; Warren et al., 2011).

Seit Mitte der 1990iger Jahre sind theoretische Erklärungsmodelle – sogenannte ‚pathway‘-Modelle (z.B. Fein & Vossekuil, 1999) – entwickelt worden, die der Frage nach sukzessiven Einflussfaktoren (behaviorale, kognitive, emotionale und situative) auf dem Weg von der Drohung zur Umsetzung der Tat nachgegangen sind. Diese Modelle beziehen sich meist auf spezifische Formen von Drohungen, so z.B. auf die Bedrohung von öffentlichen oder berühmten Personen (Fein & Vossekuil, 1999) oder die Ankündigung von ‚School Shootings‘ (Hoffmann &

Roshdi, 2013) und ihre empirische Absicherung steht noch weitestgehend aus (Meloy et al., 2014). Ebenso stehen nur für spezifische Kontexte von Drohungen, wie z.B. Gewalt am Arbeitsplatz (Workplace assessment and targeted violence risk [WAVR-21]; Meloy, White, & Hart, 2013), Stalking (Stalking assessment and management checklist [SAM]; Kropp, Hart, & Lyon, 2008), oder Intimpartnertötung (DyRiAS; Hoffmann & Glaz-Ocik, 2012) Instrumente zur Verfügung, auf die im Rahmen eines Drohungsassessments zurückgegriffen werden kann (Meloy et al., 2014). Gesamthaft ist die empirische Absicherung von Modellen um das Drohungsassessment eher schwach, wobei als wesentlicher Grund dafür die geringe Basisrate von in die Tat umgesetzten Drohungen gesehen werden kann, die einen empirischen Nachweis der Zuverlässigkeit der Modelle erschwert (Meloy et al., 2014). Darüber hinaus adressieren bestehende Modelle eher spezifische und häufig auch seltenere Phänomene, wie z.B. School-Shootings oder die Bedrohung öffentlicher Personen ab, während das „Tagesgeschäft“ der Polizei, wie zum Beispiel die Beurteilung der Ausführungsgefahr bei Drohungen im häuslichen Kontext, und Strategien zur Triagierung zwischen Hochrisiko- und Niedrigrisikodrohungen in diesem Bereich noch weitestgehend ausstehen (Meloy et al., 2014). Auf unterster Stufe einer ‚Management-Pyramide‘ ist die Polizei gefordert, an erster Stelle abzuklären, *ob* ein Drohungsmanagement notwendig ist und an zweiter Stelle zu erarbeiten, *wie* dieses gestaltet werden sollte. Wann birgt also eine Drohung ein hohes Risiko, so dass die Implementierung eines Drohungsmanagements gefordert ist? Und wann kann das Ausführungsrisiko als gering eingestuft und der Fall mit relativ geringem Aufwand abgeschlossen werden?

Ziel der sechsten für die vorliegende Dissertation relevanten Studie war es daher, einen Ansatz zur Triagierung zwischen Hoch-Risiko- und Niedrig-Risiko-Drohungen zu entwickeln, der die bisherigen, zum Zeitpunkt der Studie vorliegenden wissenschaftlichen Erkenntnisse in

einem allgemeinen Risikoschema zusammenfasst. Auf Grundlage der umfassenden Literaturrecherche zeigte sich, dass dieser Ansatz der strukturiert-klinischen Methode folgen würde, d.h. sich auf wenige empirisch nachgewiesene Zusammenhänge sowie auch auf theoretische, bisher nicht ausreichend evaluierte Modelle stützen müsse.

### **Übersichtsstudie: Identifikation von Hoch-Risiko-Drohungen**

Die von verschiedenen Autoren in den letzten 25 Jahren aufgeworfenen Risikomerkmale lassen sich in ein Modell – bestehend aus vier übergeordneten risikorelevanten Bereichen – integrieren, wonach Charakteristika der Drohung, Charakteristika der drohenden Person, Warnverhalten und aktuelle Belastungsfaktoren die Ausführungsgefahr einer Drohung beeinflussen (Gerth & Graber, 2012).

- Charakteristika der Drohung betreffen ihre Erscheinungsform und ihre Inhalte. Eine Drohung kann je nach Ausprägung gemäß O'Toole (2000) einer von drei Risikokategorien (‚niedrig‘, ‚moderat‘ und ‚hoch‘) zugeordnet werden. So wird eine detailreiche Darstellung von Motiven, Mitteln zur Tatumsetzung, Opfer und Tatort als risikoreich eingeschätzt, sobald die Aussagen einem den Umständen entsprechenden realistischen Szenario gerecht werden (O'Toole, 2000).
- Für die Ausführungsgefahr relevante Charakteristika der drohenden Person sind eine dissoziale Persönlichkeitsstruktur (McNiel & Binder, 1989; Mullen et al., 2009), wahnhaftes Erleben (Warren et al., 2008), frühere oder aktuelle Gewalthandlungen (Warren et al., 2011), Waffeneinsatz und Waffenaffinität (Meloy et al., 2012; Warren et al., 2011), Substanzmittelproblematik (Warren et al., 2011) und Suizidalität (Fein & Vossekuil, 1999).

- Risikoerhöhend ist des Weiteren, wenn die drohende Person Formen von Warnverhalten zeigt. Unter Warnverhalten sind einerseits Vorbereitungshandlungen und andererseits eine zunehmende Fixierung bzw. Wahrnehmungseinengung auf spezifische Personen oder Konflikte zu verstehen (Meloy et al., 2012). Dabei kann das Verhalten verschiedene Formen annehmen – von sich intensivierendem militärischen Interesse und „Testläufen“ gewalttätigen Verhaltens über Kontaktaufnahmen mit dem Opfer, einem wachsendem Mitteilungsbedürfnis gegenüber Dritten bis hin zu einer fatalistischen Rechtfertigung des Vorhabens aus Mangel an Alternativen (Meloy et al., 2012).
- Letztlich spielen situative Faktoren, die eine Belastungssituation für die drohende Person darstellen, eine wichtige Rolle für die Beurteilung des akuten Ausführungsrisikos. Zu unterscheiden sind dabei Faktoren, die im Zusammenhang mit der formulierten Drohung stehen (u.a. eine Zuspitzung des Konflikts z.B. durch den bevorstehenden gerichtlichen Entscheid über einen Rechtsstreit oder wegen des Zugzwangs, eine konditional formulierte Drohung umzusetzen, da die entsprechende Situation eingetreten ist) und Faktoren, die die soziale oder persönliche Situation verschärfen, z.B. Abwenden von nahestehenden Personen oder finanzielle Sorgen, Absetzen von Medikamenten oder Entwicklung einer akuten Psychose (Meloy et al., 2012).

Auf Grundlage der bisherigen Befunde zu diesen Merkmalen ist davon auszugehen, dass Auffälligkeiten in einem oder mehreren der genannten Bereiche die Ausführungsgefahr einer Drohung erhöhen. Je mehr Bereiche bei der drohenden Person als auffällig gelten, desto höher könnte das Ausführungsrisiko der Drohung sein. Die Bereiche subsumieren statische und

dynamische Risikofaktoren, weshalb sie zum einen auf eine grundlegende Risikodisposition hinweisen, zum anderen Merkmale identifizieren, die das Risiko akut beeinflussen könnten und damit Anknüpfungspunkte für das Risikomanagement bieten. In der aktuellen Form ist das Modell ein theoriegeleiteter, minimal strukturierter Leitfaden, dessen Weiterentwicklung zu einem strukturierten Instrument noch aussteht. In der praktischen polizeilichen Arbeit der Fachgruppe des Bedrohungsmanagements des Kantons Zürich (Schweiz) hat es sich jedoch bereits als nützlich bei der Triagierung und dem Management von Personen, die drohen, erwiesen (persönliches Gespräch mit Dr. Astrid Rossegger am 14.10.2014).

### **1.5. Ausblick**

Ausgehend von der grundlegenden Kritik, dass das Rückfallrisiko von Gewalt- und Sexualstraftätern anhand der unstrukturiert-klinischen Methode nicht zuverlässig eingeschätzt werden kann, haben sich in den letzten vier Jahrzehnten weitere methodische Ansätze etabliert. Dass der mit diesen neuen methodischen Ansätzen einhergehende Grad an Strukturierung sinnvoll war, wird durch eine Vielzahl von Studien gestützt, die die Überlegenheit mechanischer und im Spezifischen aktuarischer Risk-Assessment Instrumente gegenüber der unstrukturiert-klinischen Methode ausgewiesen haben (z.B. Hanson & Morton-Bourgon, 2009). Das Hinzuziehen standardisierter Risk-Assessment Instrumente wurde beispielsweise in Deutschland zur Qualitätssicherung inzwischen in die Mindestanforderungen von ‚Prognosegutachten‘ aufgenommen (Boetticher et al., 2007). Wenngleich es eine breite empirische Absicherung für die Überlegenheit der mechanischen und aktuarischen Risk-Assessment Instrumente gegenüber der ‚freihändig‘ durchgeführten Risikobeurteilung gibt, muss beachtet werden, dass auch diese Instrumente im Durchschnitt nur ‚moderate‘ Trennschärfen erzielen (z.B. Hanson & Morton-Bourgon, 2009; Kroner & Mills, 2001; Langton et al., 2007; Messing & Thaller, 2013;

Rettenberger, Matthes, Boer, & Eher, 2009). Ferner müssen die berichteten Validitätswerte vor dem Hintergrund heterogen gestalteten Studiendesigns kritisch diskutiert werden. Dazu gehört auch, dass die praktische Relevanz der für die publizierten Risikonormen zweifelhaft ist, da die wenigen, die Kalibrierung untersuchenden Studien die Übertragbarkeit der Normen auf andere Stichproben derzeit in Frage stellen. In den Zürcher Untersuchungen zum SORAG und ODARA ergibt sich bis auf eine allgemeine Überschätzung der Rückfälligkeit allerdings kein einheitliches Bild bezüglich der kategorienspezifischen Abweichungen der Rückfallraten. Während in der Stichprobe der Sexualstraftäter schon von vornherein ein grundlegender Unterschied in der Basisrate von Rückfälligkeit zwischen Entwicklungs- und Replikationsstichprobe besteht, liegt anhand des ODARA bei einer vergleichbaren Basisrate erneuter Intimpartnergewalt vor allem eine Überschätzung der Rückfälligkeit in den hohen Risikokategorien vor. Das heißt, dass es eine „zu geringe“ Rückfallrate in den höheren Risikokategorien ist, die die Trennschärfe beeinträchtigt.

Eine mögliche Erklärung für die Abweichungen könnte in der atheoretischen Konstruktionsweise liegen, durch die das Instrument kein inhaltliches Konstrukt, sondern lediglich Korrelate mit Rückfälligkeit über einen langen Beobachtungszeitraum hinweg abbildet und damit dynamische, risikoverändernde Prozesse nicht erfasst werden können. Dies scheint vor allem deshalb relevant, da es sich in beiden Zürcher Stichproben um vergleichsweise interventionsintensive Stichproben handelt. In diesem Sinne läge dann zwar keine initiale Fehlbeurteilung des Rückfallrisikos durch das Instrument vor, aber eine Unterschätzung dynamischer Prozesse, die im weiteren Verlauf den Zusammenhang zwischen Rückfallrisiko und Rückfälligkeit beeinflussen könnten. So konnten Belfrage et al. (2012) zeigen, dass die Höhe des Rückfallrisikos mit der Intensität von Interventionen korreliert und diese das Vorkommen von

Rückfälligkeit höchstwahrscheinlich medieren. Allerdings steht noch weitgehend infrage, welche durch Interventionen veränderbare Merkmale tatsächlich mit Rückfälligkeit zusammenhängen (vgl. Quinsey et al., 2006). Zu dieser Skepsis geben auch die Befunde der vorgestellten Zürcher Studie zum DyRiAS Anlass sowie eine Übersichtsarbeit zum dynamischen Risk-Assessment Instrument STABLE-2007 (Hanson et al., 2007), die zusammenfassend beschreibt, dass der Gewinn der Erhebung dynamischer Risikomerkmale zur Bestimmung rückfallrelevanter Veränderungen noch nicht aufgezeigt werden konnte (Eher, Matthes, & Rettenberger, 2012). Sollten dynamische Risikomerkmale das Ergebnis eines Risk-Assessments zu Beginn des Beobachtungszeitraums in gewisser Weise von den tatsächlichen Rückfalldaten am Ende des Beobachtungszeitraumes „entkoppeln“, ist es plausibel anzunehmen, dass die Kriteriumsvalidität dieser Risk-Assessment Instrumente in Abhängigkeit der Intensität der Interventionen variiert und daher weder die Instrumente für eine Anwendung bei Therapiepopulationen im Verlauf noch die Therapiepopulationen für die Überprüfung der Validität geeignet sind. Vor diesem Hintergrund wäre eine Erweiterung der bestehenden ODARA- und SORAG-Risikomodelle um dynamische Risikomerkmale, wie es schon anhand des STABLE-2007 vorgeschlagen wurde, zu überdenken. Der STABLE-2007 wird mit seinen dynamischen Risikomerkmale, die sich auf die ‚kriminogenen Bedürfnisse‘ gemäß RNR-Prinzip beziehen, als Ergänzung zum Static-99 angewendet und stellt damit eine Ergänzung zum ausschließlich statischen Risk-Assessment dar.

Eine weitere Erklärung für die unzureichende Kriteriumsvalidität könnte darin liegen, dass die konkreten Risikomerkmale, die sich in den kanadischen Stichproben mit Rückfälligkeit als zusammenhängend erwiesen, nicht im gleichen Ausmaß zur Risikobeurteilung im Kanton Zürich geeignet sind und wie im Beispiel des ODARA zu einer geringen Spezifität des Instrumentes und

damit Überschätzung des Risikos führen. Während die Rückfallrate in der niedrigsten Kategorie des ODARA in der Zürcher Stichprobe bei 0% liegt, steigt sie bis zur höchsten Kategorie nur auf 45% an. Das heißt, weniger als die Hälfte der Täter der ‚Hochrisiko‘-Kategorie werden rückfällig und die über das ODARA erfassten Merkmale sind für diese Kategorie offensichtlich zu unspezifisch. Bis jetzt ist es noch unklar, ob ein valideres Modell anhand einer Adaption oder Erweiterung des ODARA um andere Riskomerkmale zu ermöglichen wäre oder nicht eher die Erarbeitung mehrstufiger Risk-Assessment Strategien notwendig wäre, bei denen zunächst eine grundlegende Risikodisposition im Rahmen eines Screenings zur Triage ermittelt werden und darauffolgend bei als ‚Hochrisiko‘-Tätern eingeschätzten Personen eine vertiefte Abklärung unter Einbezug spezifischerer Merkmale erfolgen könnte. Ein erster Vorschlag diesbezüglich wurde von den Autoren des ODARA vorgebracht, indem im Anschluss des ODARA bei Erreichen eines spezifischen Grenzwertes der Domestic Violence Risk Appraisal Guide (DVRAG; Hilton, Harris, & Rice, 2010) angewendet werden soll, der zum einen eine spezifischere Gewichtung der ODARA Items und zum anderen den Einbezug der Psychopathy Checklist – einer Checkliste zur Erfassung des Konstrukts der Psychopathie gemäß Hare (2003) – vorsieht. Allerdings liegt noch keine einzige peer-reviewed publizierte Replikationsstudie vor, die die Eignung des mehrstufigen ODARA/DVRAG Risk-Assessment Systems bestätigen würde.

Prospektive Studien zur Erfassung dynamischer Prozesse sowie spezifische Analysen innerhalb von Hochrisikopopulationen sollten Gegenstand zukünftiger Studien zur Optimierung der Risk-Assessment Strategien im Kanton Zürich sein.

## 2. Eigene Arbeiten

### 2.1. Current obstacles in replicating risk assessment findings: A systematic review of commonly used actuarial instruments

#### 2.1.1. Abstract

*Background:* An actuarial risk assessment instrument can be considered valid if independent investigations using novel samples can replicate the findings of the instrument's development study. In order for a study to qualify as a replication, it has to adhere to the methodological protocol of the development study with respect to key design characteristics, as well as ensuring that manual-recommended guidelines of test administration have been followed.

*Methods:* A systematic search was conducted to identify predictive validity studies ( $N = 84$ ) on three commonly used actuarial instruments: the Violence Risk Appraisal Guide (VRAG), the Sex Offender Risk Appraisal Guide (SORAG), and the Static-99. Sample (sex, age, criminal history) and design (follow-up, attrition, recidivism) characteristics, as well as markers of assessment integrity (scoring reliability, item omissions, prorating procedure), were extracted from 84 studies comprising 108 samples.

*Results:* None of the replications matched the development study of the instrument they were attempting to cross-validate with respect to key sample and design characteristics. Furthermore none of the replications strictly followed the manual-recommended guidelines for the instruments' administration.

*Conclusion:* Additional replication studies that follow the methodological protocols outlined in actuarial instruments' development studies are needed before claims of generalizability can be made.

### 2.1.2. Introduction

Violence is a major public health issue, which has led to concern regarding the recidivism risk of offenders. To assess violence risk, structured risk assessment instruments are increasingly used. One form of structured risk assessment follows the actuarial approach, in which statistical correlates of the antisocial behavior of interest are identified and weighted according to the strength and direction of association. Total scores are summed and converted via tables into probabilistic estimates of reoffending risk (Grove et al., 2000). There are numerous actuarial risk assessment instruments (ARAI) available for assessing recidivism risk in violent and sexual offenders. Among them the Violence Risk Appraisal Guide (VRAG; Quinsey et al., 2006), the Sex Offender Risk Appraisal Guide (SORAG; Quinsey et al., 2006), and the Static-99 (A. Harris, Phenix, Thornton, & Hanson, 2003) are the most commonly used according to recent surveys (R. P. Archer et al., 2006; Viljoen et al., 2010). These ARAIs were designed using a primarily atheoretical approach and as they were constructed using correlative associations, it is important to replicate initial findings to ensure generalization to similar samples. Replication studies differ from cross-validation studies in that they attempt to establish the predictive validity of risk assessment instruments when used as designed.

More specifically, a replication study should, at a minimum, consider the following key study characteristics when attempting to reproduce the predictive validity findings of ARAIs and to guarantee assessment integrity (e.g., following manual-recommended guidelines of test administration):

- *Sample:* Since age (Sampson & Laub, 2003) and sex (Benda, 2005) correlate with persistent offending, studies investigating the robustness of a specific ARAI should use the same inclusion criteria that were used for the development sample with respect to

age and sex. The development samples of the VRAG (Quinsey et al., 2006), the SORAG (Quinsey et al., 2006) and the Static-99 (A. Harris et al., 2003) consisted of adult male offenders. A further basic criterion with respect to sample composition is the type of offense that led to study inclusion (“index offense”). Some ARAIs were developed for violent and sexual offenders (e.g., VRAG), while others were specifically developed for sexual offenders (e.g., SORAG and Static-99). Sound replication studies use the same study inclusion criteria with regard to the type of index offense stated in the development study.

- *Follow-up.* As ARAIs produce estimates of recidivism risk within a specific timeframe, and as length of follow-up negatively influences base rate, replication studies should attempt to use fixed lengths of follow-up similar to those used in development studies (G. T. Harris & Rice, 2003; G. T. Harris et al., 2003; Quinsey et al., 2006).
- *Controlling for attrition.* To avoid potential bias associated with attrition, whether participants died, left the jurisdiction in which a study was conducted (by their own choosing or as result of deportation to their countries of origin), were institutionalized, or changed their names while at risk should be investigated (G. T. Harris & Rice, 2007).
- *Outcome.* Recidivism should have the same operationalization as the development study, especially with regard to type of persistent offending (e.g., violent [including sexual], violent [non-sexual], sexual) as well as legal status (charge, conviction, incarceration, self-report).
- *Assessment integrity.* ARAIs are based on a mechanical algorithm and do not allow for the alteration of items, the modification of item weights, or the use of interpretations of the final score other than those recommended by instrument authors. It is therefore

important that replication studies follow instrument manuals with regard to administering instruments as well as interpreting results. Replication studies should therefore use the ARAIs as originally designed without item approximation, without systematic item omissions, following manual-recommended “proratings” where permitted (G. T. Harris, Rice, & Camilleri, 2004; Quinsey et al., 2006), and using only information collected from reliable data sources (e.g., assessors should be trained in the application of the instrument (Hanson, 2012) and a satisfactory level of inter-rater reliability should be shown (G. T. Harris et al., 2004). The assessor should have access to an official criminal record as recorded by police, court, or correctional officials and should not solely rely on offender self-report (A. Harris et al., 2003).

### **Present study**

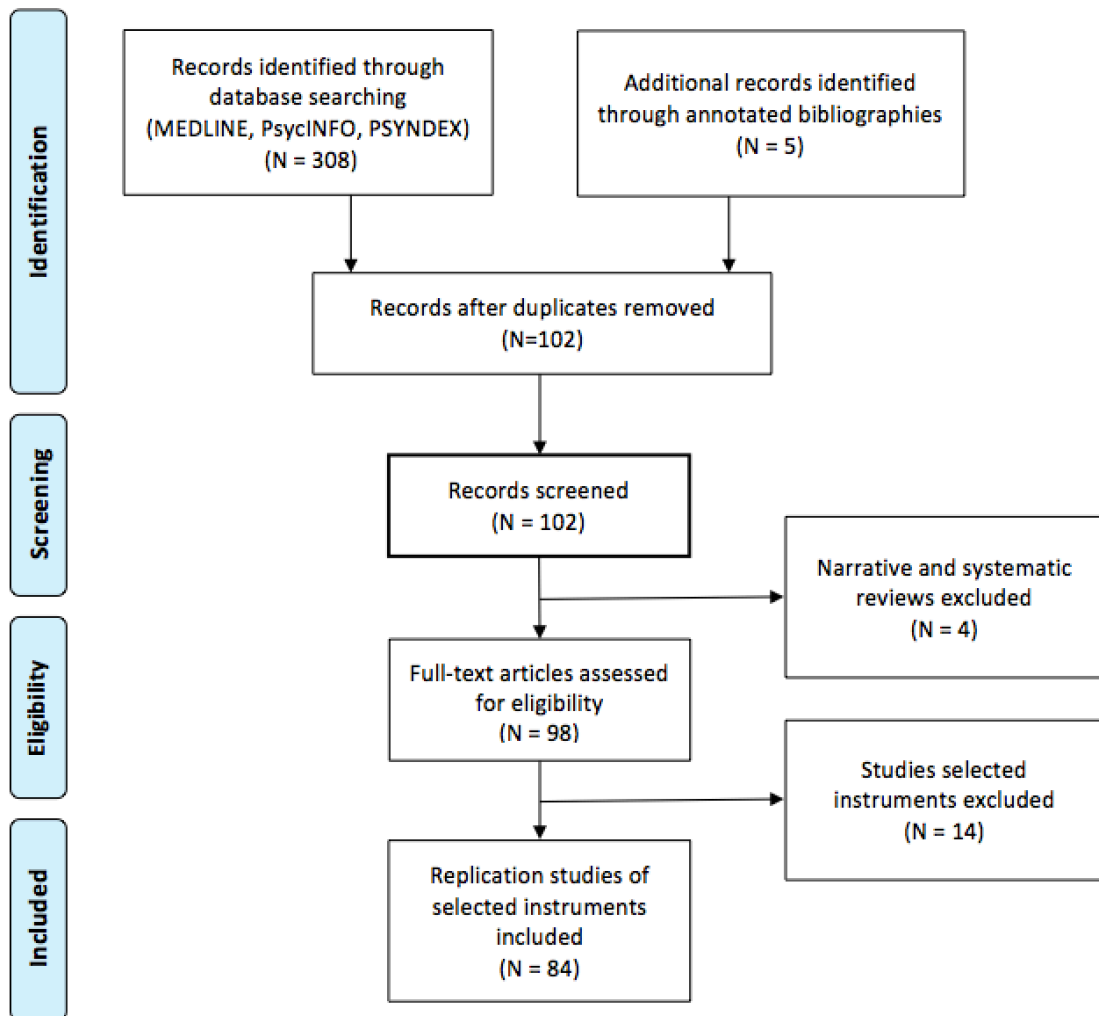
Despite the importance of the topic, the extent to which replication studies of ARAIs are similar to development studies has not been investigated in previous systematic reviews of the risk assessment literature (e.g. Singh, Grann, & Fazel, 2011). To address this gap, we examined the replication literature on three commonly used ARAIs (the VRAG, SORAG, and Static-99) to analyze the characteristics detailed earlier.

### **2.1.3. Methods**

#### **Study selection**

A systematic search of the predictive validity literature on the VRAG, SORAG and Static-99 was conducted using the MEDLINE, PsycINFO, and PSYINDEX databases and both the full names and acronyms of instruments as keywords. Peer-reviewed primary studies in any language from any country published until June 2012 were considered for inclusion. The initial search

identified 313 records. After excluding duplicates and reviews, 98 studies remained. A further 14 studies were excluded as they did not investigate the predictive validity of the instruments of interest, resulting in a final count of 84 studies comprising 108 samples (*Figure 1*).



*Figure 1.* Flowchart depicting the results of a systematic search for replication studies of three commonly used risk assessment instruments (June 2012)

**Data extraction**

Twelve criteria of similarity between development and replication studies as well as assessment integrity were extracted. For an overview of the VRAG, SORAG, Static-99 and characteristics of the respective development samples, see Table 1.

**Table 1.** Characteristics of three commonly used actuarial risk assessment instruments

Characteristic	Instrument		
	VRAG	SORAG	Static-99
Manual	Quinsey et al. (2006)	Quinsey et al. (2006)	A. Harris et al. (2003)
Offender sex	Male	Male	Male
Offender age (years)	Adults (18 at time of release)	Adults (18 at time of release)	Adults (18 at time of release)
Index offense	Violent/Sexual (contact offenses only)	Sexual (contact offenses only)	Sexual (contact offenses or non-contact offenses, exception: illegal pornography)
Items	12	14	10
Omissions allowed	4	4	1 (Item #2)
Prorating	Yes	Yes	No
Bins	9	9	4
Type of recidivism	Violent/Sexual (no arson, contact sex-offenses only)	Violent/Sexual (no arson, contact sex-offenses only)	Violent/Sexual (contact offenses or non-contact offenses, exception: illegal pornography)
Legal status of recidivism	Charge/Convictions	Charge/Convictions	Convictions
Length of follow-up (years) <sup>a</sup>	7, 10	7, 10	5, 10, 15

*Note.* VRAG, Violence Risk Appraisal Guide; SORAG, Sex Offender Risk Appraisal Guide.

<sup>a</sup>Periods for which probabilistic estimates of risk are provided.

The extracted matching criteria included:

- *Sample characteristics.* The study sample matches the development sample regarding sex (entirely male population), exclusion of juvenile offenders, and type of index offense.
- *Follow-up characteristics.* The replication study covers a similar follow-up period as presented in the norm-tables published in instrument manuals (mean follow-up deviates not more than six months from the follow-up time listed in the norms of the instruments). The length of follow-up was held fixed.
- *Controlling for attrition.* Controlling for attrition was reported, e.g., whether participants left the jurisdiction, changed their names, were institutionalized or died during follow-up.
- *Outcome.* The same operationalization of recidivism in the replication study as in the development study is used as far as type of offense (e.g., violent [including sexual], violent [non-sexual], sexual) and source of information (e.g., charge, conviction, incarceration, self-report) is concerned.
- *Assessment integrity.* The application of the instruments followed instrument manuals. Instruments were not solely administered on the basis of clinical interviews. At least one official file was considered (e.g., criminal record, correctional files, clinical files). The instrument was administered by either trained raters or assessors for whom a strong inter-rater reliability was established ( $ICC \geq .75$ , Krippendorff's  $\kappa \geq .75$ , Pearson's  $r \geq .80$ ). No items were approximated (i.e., the original, unaltered version was used). The authors who used non-English versions of the instruments used validated translations

(i.e., translation and back-translation conducted by independent translators). No items were systematically omitted.

Since the VRAG and the SORAG allow for the raters to use proratings, missing items should be replaced in accordance with the manual-recommended protocol. The algorithm for prorating (other than assigning a 0 for missing cases was first published by Quinsey et al. (2006) and hence was not available for VRAG/SORAG replication studies published before then. Therefore, it was not scored as a matching criterion.

Information regarding the 12 matching criteria was extracted from the *Abstract*, *Method*, and *Results* sections. Missing information was coded as a mismatch with the development study with the exception of the criteria regarding item approximation and systematic item omission. If not stated otherwise, it was assumed that study authors used the unaltered English-language versions of the instruments and did not systematically omit items.

### **Inter-rater reliability**

As a measure of quality control, characteristics of 45 studies were extracted by three of the authors (A.R., J.G., K.S.), all of whom had at least a masters degree in psychology. The inter-rater reliability of the data extracted was strong ( $\kappa > .80$ ). Disagreements were settled by consensus. The two psychologists with the most experience in administering ARAIs (A.R., J.G.) rated the remaining investigations.

### **Data Analysis**

Data were analyzed using STATA 12.0 for Windows (StataCorp, 2012).

#### **2.1.4. Results**

The majority of the 84 studies were published in English ( $N = 76$ ; 90.5%) followed by German ( $N = 6$ ; 7.1%) and French ( $N = 2$ ; 2.4%). Studies originated in Canada ( $N = 31$ ; 37.0%), the U.K. ( $N = 13$ ; 15.5%), the U.S.A. ( $N = 11$ ; 13.1%), Austria ( $N = 6$ ; 7.1%), Switzerland ( $N = 5$ ; 6.0%), Germany ( $N = 4$ ; 4.8%) and Belgium ( $N = 3$ ; 3.6%). In more than a third of the articles ( $N = 32$ ; 38.1%), predictive validity was investigated using overlapping samples. Of the 108 included samples, 17 (15.7%) included fewer than 100 participants, and 44 (40.7%) fewer than 200 participants. Twenty-four samples (22.2%) included more than 500 participants.

##### **Replication match**

**Sample characteristics.** Approximately two-thirds ( $k = 71$ ; 65.7%) of samples consisted only of males, and participants were 18 years or older in approximately half of them ( $k = 48$ ; 44.4%). There was no information regarding sex composition for about a fifth of the samples ( $k = 23$ ; 21.3%), and there was no information regarding participant age for approximately one-half ( $k = 45$ ; 41.7%). Over half of the samples ( $k = 68$ ; 63.0%) matched the development study with regard to the nature of participants' index offense. For approximately a fifth of the samples ( $k = 19$ ; 17.6%), it was not possible to determine the index offense that participants had committed.

*Follow-up characteristics.* For 28 (25.9%) samples, a similar follow-up period as in the instruments' manuals was included. The length of follow-up was held fixed for 34 (31.5%) samples.

**Controlling for sample attrition.** Approximately a third of the samples ( $k = 34$ ; 31.5%) controlled for sample attrition (e.g., controlling for participant death [ $k = 29$ ; 26.9%] and emigration/deportation [ $k = 20$ ; 18.5%]).

Table 2 shows samples controlled for attrition stratified for each of the three instruments.

**Table 2.** Handling attrition

Source of attrition	Total ( <i>k</i> = 108)	VRAG ( <i>k</i> = 38)	SORAG ( <i>k</i> = 21)	Static-99 ( <i>k</i> = 49)
Death	29 (26.9%)	11 (29.0%)	4 (19.1%)	12 (28.6%)
Immigration	20 (18.5%)	8 (21.1%)	3 (14.3%)	9 (18.4%)
Name change	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Other	13 (12.0%)	0 (0%)	2 (9.5%)	8 (16.3%)

*Note.* VRAG, Violence Risk Appraisal Guide; SORAG, Sex Offender Risk Appraisal Guide; *k* = Number of samples

**Outcome.** In over half of the samples (*k* = 60; 63.9%), the type of recidivism was the same as in the instrument's development study (e.g. violent [including sexual], violent [non-sexual], sexual) while this was the case in only 34.3% (*k* = 37) with regard to the source of information used to detect the outcome criterion (e.g. charge, conviction, incarceration, self-report).

**Assessment integrity.** In two-thirds (*k* = 83; 76.9%) of the included samples, instruments were scored based on clinical interviews supplemented by file review. Reliable scoring (either using trained assessors or establishing a good inter-rater reliability) was reported in over half of the samples (*k* = 63; 58.3%). In the majority of samples (*k* = 93; 86.1%), there was no indication that items had been omitted systematically. Prorating guidelines according to instruments' manuals were explicitly reported for 2 (3.4%) of the 59 samples investigating the predictive validity of the VRAG or the SORAG. In 8 (16.3%) of the 49 samples investigating the predictive validity of the Static-99, the authors explicitly stated that there were no missing items.

**Total matching score.** On average, replication studies considered roughly two-thirds of the 12 matching criteria ( $M = 6.6$  [ $SD = 2.0$ ],  $Md = 7$ ,  $IQR = 5.5-8$ ; VRAG:  $M = 6.1$  [ $SD = 1.7$ ],  $Md$

CURRENT OBSTACLES IN REPLICATING RISK ASSESSMENT FINDINGS

= 6, *IQR* = 5-7; SORAG: *M* = 6.0 [*SD* = 2.3], *Md* = 6, *IQR* = 5-7; Static-99: *M* = 7.3 [*SD* = 1.8], *Md* = 8, *IQR* = 6-9; Table 3).

**Table 3.** Correspondence between development and replication studies of three commonly used actuarial risk assessment instruments

Replication match	Total ( <i>k</i> = 108)	VRAG ( <i>k</i> = 38)	SORAG ( <i>k</i> = 21)	Static-99 ( <i>k</i> = 49)
Offender sex (entirely male)	71 (65.4%)	22 (57.9%)	13 (61.9%)	36 (73.5%)
Offender age (entirely adult)	48 (44.4%)	16 (42.1%)	10 (47.6%)	22 (44.9%)
Index offense	68 (63.0%)	16 (42.1%)	11 (52.4%)	41 (83.7%)
Using file information <sup>a</sup>	83 (76.9%)	33 (86.8%)	17 (81.0%)	33 (67.4%)
Reliable scoring <sup>b</sup>	63 (58.3%)	25 (65.8%)	13 (61.9%)	25 (51.0%)
No item approximations <sup>c</sup>	88 (81.5%)	33 (86.8%)	14 (66.7%)	41 (83.7%)
No systematic item omission	93 (86.1%)	35 (92.1%)	15 (71.4%)	43 (87.8%)
Length of follow-up	28 (25.9%)	7 (18.4%)	4 (19.1)	17 (34.7%)
Fixed length of follow-up	34 (31.5%)	17 (44.7%)	3 (14.3%)	14 (28.6%)
Controlling for attrition	34 (31.5%)	12 (31.6%)	5 (23.8%)	17 (34.7%)
Type of recidivism	69 (63.9%)	10 (26.3%)	13 (61.9%)	46 (93.9%)
Legal status of recidivism	37 (34.3%)	6 (15.8%)	7 (33.3%)	24 (48.9%)

*Note.* VRAG, Violence Risk Appraisal Guide; SORAG, Sex Offender Risk Appraisal Guide; *k* = Number of samples

<sup>a</sup>Scoring was not solely based on clinical interviews.

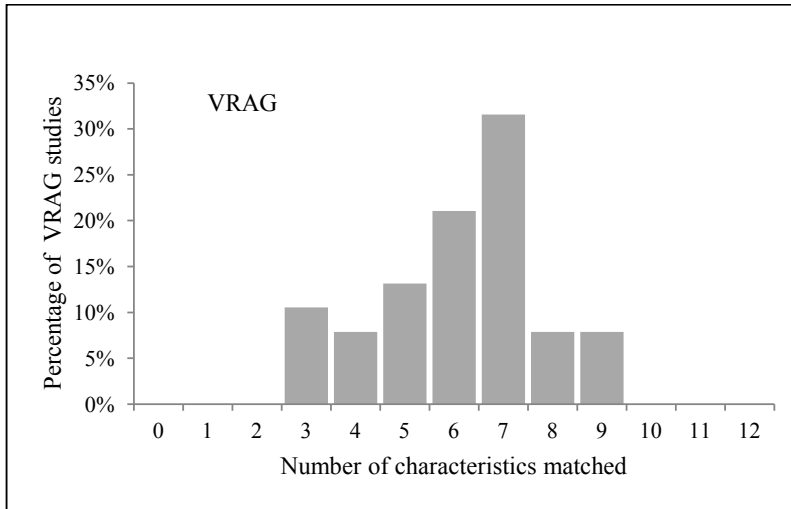
<sup>b</sup>Trained raters or excellent inter-rater reliability (Krippendorff's  $\kappa \geq .75$ ; *ICC*  $\geq .75$ ; Pearson's *r*  $\geq .80$ ).

<sup>c</sup>Operationalizing of items has not been changed or no other than original version or a published translation of the instrument has been used.

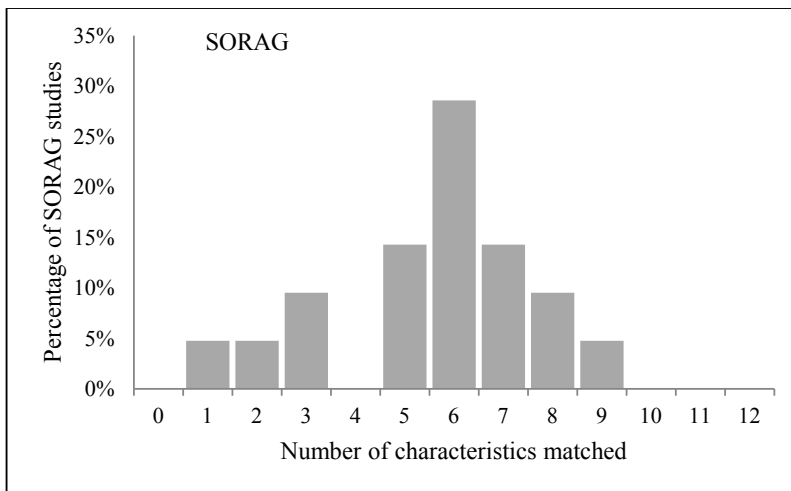
None of the 108 samples scored a perfect replication match regarding the 12 criteria, although one study measuring the predictive validity of the Static-99 did meet 11 criteria (Barbaree, Langton, Blanchard, & Cantor, 2009). See Figure 2 to Figure 5.

## CURRENT OBSTACLES IN REPLICATING RISK ASSESSMENT FINDINGS

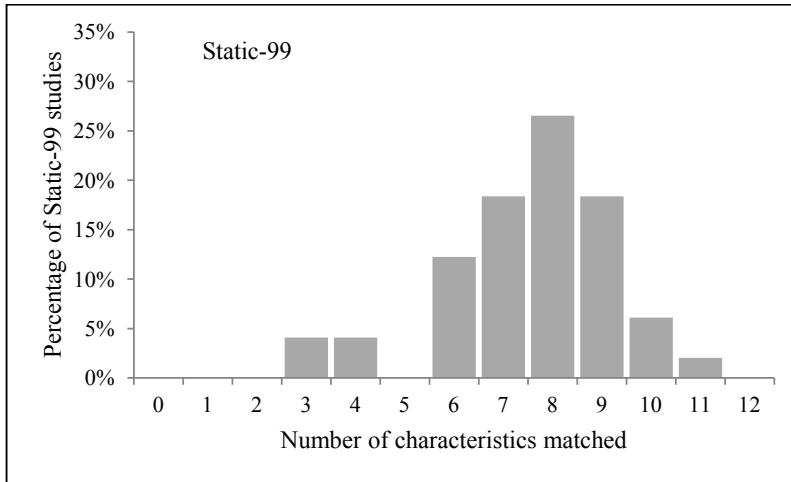
---



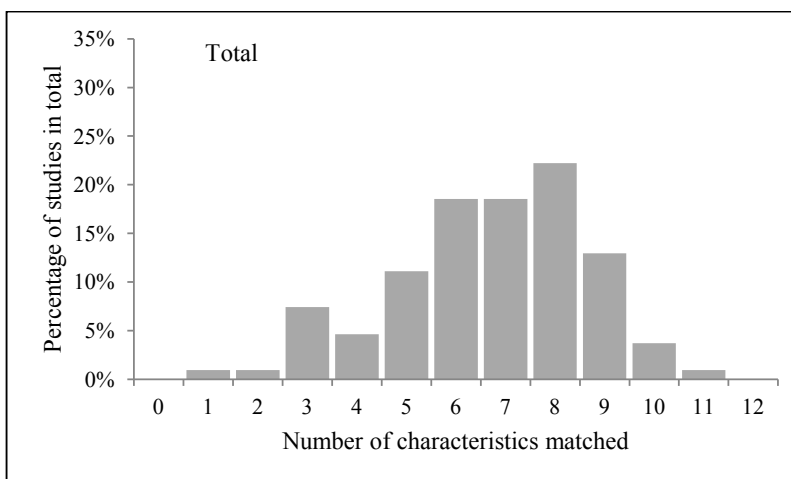
**Figure 2.** Number of characteristics matched between development and replication studies of the VRAG.



**Figure 3.** Number of characteristics matched between development and replication studies of the SORAG.



**Figure 4.** Number of characteristics matched between development and replication studies of the Static-99.



**Figure 5.** Number of characteristics matched between development and replication studies of three commonly used actuarial risk assessment instruments (VRAG, SORAG, Static-99).

### 2.1.5. Discussion

According to recent surveys the VRAG, SORAG and Static-99 are the most commonly used ARAIs in clinical practice (R. P. Archer et al., 2006; Viljoen et al., 2010). ARAIs can be considered valid if independent studies can replicate the findings of the development study. In order for a study to qualify as a replication, it has to follow the study protocol of the original study with respect to key study characteristics as well as guidelines for test administration.

Instrument authors claim that there are several replication studies corroborating the validity of ARAIs (Phenix, Hanson, Harris, & Thornton; Waypoint Centre for Mental Health Care, 2012). It is unclear however, whether these replication studies matched the development studies with respect to key characteristics of study protocol as well as the administration of the ARAIs.

The goal of the present study was to examine the similarity between development and replication studies and to investigate assessment integrity. A systematic search of the predictive validity literature on the VRAG, SORAG, and Static-99 identified 84 peer-reviewed investigations comprising 108 samples. Correspondence between the development and replication studies was assessed on the basis of sample characteristics, instrument administration, follow-up, controls for attrition and outcome definition.

The replication match varied considerably across these criteria. The highest replication match could be found regarding assessment integrity, namely whether assessors did not systematically omit (86.1%) or alter items (81.5%). This score has to be interpreted with caution though, since assessment integrity was assumed as given if not otherwise explicitly stated. In 76.9% of the study samples – as required by instrument manuals - the authors consulted file information. This finding shows that assessment integrity was considerably violated in roughly one out of four cases, since the basis of information did not meet the minimum requirement for test administration.

The remaining matching criteria were considered in less than two-thirds of the studies. Offender sex composition matched the development sample in 65.7% of cases, type of recidivism in 63.9% and type of index offense in 63.0%. Reliability of instrument administration was considered for 58.3% of the study samples either by using trained raters or by documenting inter-rater reliability. The remaining five criteria were explicitly considered by less than 50% of

the study samples: 44% percent of the samples consisted solely of adult offenders. Finally, controlling for attrition and fixed length of follow-up periods were the least frequently reported criteria (31.5% each).

Of the 108 samples investigated, no study could be identified with a perfect replication match. Roughly half of the study samples matched the development study in two-thirds or fewer of the relevant criteria, although the replication match was better for Static-99 studies than for SORAG/VRAG validation studies.

It is unclear however, whether the replication study authors neglected to report information regarding sample characteristics, application of instrument, follow-up, attrition, and outcome definition or did not, in fact, consider these characteristics fully in their empirical design. If the finding of this paper is that we are dealing with a mere reporting phenomenon, study authors should be reminded about sticking to the scientific protocol in relation to the reporting of results. If, however, the relatively poor replication match was the result of altered methodological designs, the studies extracted for this investigation cannot be considered as true replication studies. Accordingly, the results of these studies should be interpreted with caution and cannot serve as a direct corroboration of the predictive validity of ARAIs.

Since the majority of the extracted investigations found a significant association with an outcome, the included studies are commonly interpreted as proof of the robustness of the VRAG, SORAG, and Static-99 (e.g. Hastings, Krishnan, Tangney, & Stuewig, 2011; Kröner, Stadtland, Eidt, & Nedopil, 2007) though the purported replication investigations differed with respect to the follow-up period (Quinsey, Book, & Skilling, 2004; Rettenberger, Matthes, et al., 2009), the composition of the study sample (G. T. Harris, Rice, & Cormier, 2002; Hastings et al., 2011; Snowden, Gray, & Taylor, 2010), and the definition of the outcome criterion (Endrass,

Rossegger, Frischknecht, Noll, & Urbaniok, 2008; G. T. Harris & Rice, 2003; G. T. Harris et al., 2004; Hastings et al., 2011; Kroner & Mills, 2001; Lindsay et al., 2008; Loza, Villeneuve, & Loza-Fanous, 2002; Storey, Watt, Jackson, & Hart, 2012). Deviations from the methodology used in the development study were interpreted as corroboration of model robustness (G. T. Harris & Rice, 2007; G. T. Harris et al., 2004). Even though on other occasions ARAI authors stipulated specific requirement for replication studies that were also the basis for the current investigation, they still interpreted the results of studies using deviating methodological characteristics as corroboration for the ARAI (e.g. G. T. Harris & Rice, 2007; G. T. Harris et al., 2004).

Using a study that differs from the original investigation with respect to key study characteristics as corroboration for model robustness can lead to contradictory results. This is especially the case when aside from measures of accuracy (such as the “gold standard” Area under the curve (Mossman, 1994) measures of calibration are used (e.g. Endrass, Urbaniok, Held, Vetter, & Rossegger, 2009). If a replication study finds that for example 44% of the offenders in the risk-bin X reoffended within three years of follow-up (compared with the 7-year follow-up period in the original study), it is evidence of a poor calibration and not of the robustness of the model. The validity of the instrument is even more challenged, when the outcome is altered. If ARAIs correlate with all sorts of social maladaptive behavior, the question arises: what do these instruments really measure?

If ARAIs correlate with all sorts of behavior under different conditions, in different contexts, using different follow up periods and data bases, it cannot automatically serve as a corroboration for the (assumed) robustness of the instrument, as it remains unclear, what the instrument really measures. If maladaptive behavior and not proneness to sexual or violent offending was the

latent trait being assessed by an ARAI, it is questionable whether such an ARAI should be used in court to assess the risk for persistent sexual or violent offending. Whereas some deviation from the original study could suggest model robustness, a larger deviation jeopardizes its validity, especially if the deviation concerns the dependent variable.

#### **2.1.6. Conclusion**

There are only few validation studies of the most commonly used ARAIs with a high replication match. More replication studies are needed that follow the methodological protocols outlined in the original studies by instruments' authors.

## **2.2. Examining the predictive validity of the SORAG in Switzerland**

### **2.2.1. Abstract**

*Background:* The SORAG is one of the most commonly used actuarial risk assessment tools for sex offender evaluations. Although many studies have investigated the predictive validity of this instrument, few have examined whether its published expected rates of recidivism are useful in practice. The aim of the present investigation was to investigate predictive validity of the SORAG in Switzerland, considering both the discrimination and calibration components of predictive validity.

*Material and Methods:* The instrument was administered to two total cohorts of offenders ( $N = 137$ ) in the Canton of Zurich, Switzerland that were followed for a fixed period of seven years after discharge. Recidivism was defined as new charges and/or convictions for violent (including sexual) offenses. Discrimination was measured using receiver operating characteristic (ROC) curve analysis and calibration using Sanders' decomposition of the Brier score. A dependent  $t$ -test was used to examine the difference in expected and observed percentiles, and a Kolmogorov-Smirnov Test was conducted to compare the distribution of offenders in risk bins.

*Results:* ROC analyses revealed an acceptable level of discrimination for both SORAG total risk scores ( $AUC = 0.69$ , 95%  $CI = 0.56-0.82$ ) and risk bins ( $AUC = 0.67$ , 95%  $CI = 0.54-0.80$ ). The seven year recidivism rates both overall as well as for each risk bin were considerably lower than the published SORAG norms. An average forecast error of 20.9% for each risk bin of the SORAG suggests a large difference between expected and observed recidivism rates.

*Conclusion:* Should further studies in Switzerland replicate the current findings, a re-calibration of the SORAG may be needed before the instrument can be considered a valid method to assess recidivism risk in Swiss sex offender populations.

### 2.2.2. Introduction

There is strong empirical evidence suggesting that evaluations conducted using mechanical instruments result in superior assessments of risk compared to clinical judgments (Ægisdóttir et al., 2006; Grove et al., 2000; Quinsey et al., 2006). These mechanical (or “actuarial”) risk assessment instruments are composed of closed-ended questions with pre-determined responses. These responses are combined using objective, transparent, and inflexible rules proscribed by the tool authors, resulting in either a total risk score or risk classification.

Over the past 30 years, many actuarial risk assessment tools have been developed for the specific purpose of assessing the likelihood of recidivism in criminal offenders. Among these are a number of schemes designed for assessing the risk of sex offender recidivism. According to recent surveys (Jackson & Hess, 2007; Viljoen et al., 2010), the Sex Offender Risk Appraisal Guide (SORAG; Quinsey et al., 2006) is one of the more commonly applied sex offender recidivism risk assessment instruments. The SORAG is a Canadian tool composed of 14 items with a statistical association with new charges and/or convictions for violent (including sexual) offenses. The SORAG produces a total risk score that can be used to classify sex offenders into one of nine risk categories (i.e., “bins”), each of which has an affiliated expected rate of recidivism for 7 and 10 years follow-up. In addition, total risk scores can be used to compare a given offender to the population of sex offenders using published percentiles.

There is a considerable empirical base concerning the ability of the SORAG to discriminate between recidivists and non-recidivists which includes both North American ( $AUC = 0.64-0.78$ ) and European ( $AUC = 0.71-0.75$ ) investigations (Bartosh, Garby, Lewis, & Gray, 2003; Eher, Matthes, Schilling, Haubner-MacLean, & Rettenberger, 2012; Pham & Ducro, 2008; Rice & Harris, 2002). However, study authors routinely discount the guidance of the SORAG’s authors

concerning how the instrument should be administered and its findings interpreted (Rossegger, Gerth, Seewald, et al., 2013; Singh, Desmarais, & Van Dorn, 2013). Specifically, previous studies have not examined statistically whether the predicted recidivism rates for each risk bin are an accurate reflection of observed rates. And only four studies have descriptively reported the number of recidivists per risk bin (see Table 4). In addition, the accuracy of the percentile ranks offered in the SORAG manual has not been replicated. Thus, further research is necessary to clarify the usefulness of the SORAG for practitioners, especially those working in jurisdictions other than Canada, where expected recidivism rates for risk bins and score distributions may be different.

**Table 4.** Characteristics of previous studies reporting risk bin outcome information for the SORAG

	Quinsey et al. (2006) <sup>a</sup>	Nunes et al. (2002)	G. T. Harris et al. (2003)	Looman (2006)	Eher et al. (2008b), (2008a) <sup>b</sup> ; subgroup rapists)	Eher et al. (2008b), (2008a) <sup>b</sup> ; subgroup child molesters)
Replication match <sup>c</sup>	–	7	7	6	3	3
No item approximations	–	No	Yes	Yes	No	No
No systematic item omission	–	No	Yes	Yes	Yes	Yes
Reliable scoring	Yes	No	Yes	Yes	No	No
Controlling for attrition	Yes	No	No	No	No	No
File information used for scoring	Yes	Yes	Yes	Yes	Yes	Yes
Mean LoFU (years)	9.9	7.3	5.2	4.6	3.6	3.6
Legal status of recidivism	Charge + conviction	Charge + conviction	Charge + conviction	Conviction	Conviction	Conviction

*Note.* LoFU = length of follow-up; N = size of total study sample; NR = not reported.

<sup>a</sup>SORAG development sample.

<sup>b</sup>This publication is the corresponding German version of Eher et al. (2008b).

<sup>c</sup>Out of 12 matching criteria established by Rossegger, Gerth, Seewald, et al. (2013).

<sup>d</sup>Base rate of violent (including sexual) recidivism for offenders with a SORAG score.

**Table 4** continued. Characteristics of previous studies reporting risk bin outcome information for the SORAG

Risk bin	Quinsey et al. (2006) <sup>a</sup>	Nunes et al. (2002)	G. T. Harris, M. E. Rice, V. L. Quinsey, et al. (2003)	Looman (2006)	Eher et al. (2008b), (2008a) <sup>b</sup> ; subgroup rapists)	Eher et al. (2008b), (2008a) <sup>b</sup> ; subgroup child molesters)
	Recidivism rate					
1	7% (14)	8% (87)	19% (NR)	0% (6)	0% (9)	5% (22)
2	15% (23)	8% (52)	18% (NR)	17% (6)	0% (13)	3% (33)
3	23% (40)	15% (46)	29% (NR)	11% (19)	15% (20)	0% (22)
4	39% (58)	30% (37)	50% (NR)	13% (28)	25% (16)	7% (15)
5	45% (52)	39% (18)	55% (NR)	32% (38)	33% (15)	15% (13)
6	58% (46)	0% (12)	63% (NR)	36% (42)	6% (15)	9% (11)
7	58% (32)	25% (4)	63% (NR)	33% (33)	19% (16)	33& (9)
8	75% (18)	50% (2)	71% (NR)	57% (44)	45% (16)	40% (5)
9	100% (5)	0% (0)	76% (NR)	57% (26)	43% (11)	0% (0)
Recidivism rate <sup>d</sup> (N)	42% (288)	15% (258)	48% (205)	34% (242)	24% (123)	6% (130)
Mean total risk score	8.9 (SD = 11.3)	-3.2 (SD = 10.3)	10.0 (SD = 10.8)	16.6 (SD = 12.3)	NR	NR

Note. LoFU = length of follow-up; N = size of total study sample; NR = not reported.

<sup>a</sup>SORAG development sample.

<sup>b</sup>This publication is the corresponding German version of Eher et al. (2008b).

<sup>c</sup>Out of 12 matching criteria established by Rossegger, Gerth, Seewald, et al. (2013).

<sup>d</sup>Base rate of violent (including sexual) recidivism for offenders with a SORAG score.

### **The Present Study**

The aim of the present study was to conduct the first replication of the SORAG in the country of Switzerland. Two total forensic cohorts in the Canton of Zurich were followed for up to seven years, with novel charges and/or convictions for a violent (including sexual) offense used as criteria for recidivism. Both the discrimination and calibration of the SORAG were measured, and the distribution of total scores was compared to the percentile ranks published by the tool authors. As attention was paid to matching those participant and study design characteristics of the SORAG development study, high rates of predictive validity were hypothesized.

#### **2.2.3. Material and Methods**

##### **Sample**

The study sample consisted of two total cohorts of violent and sex offenders from the Canton of Zurich, Switzerland ( $N = 861$ ). The first cohort was taken from the Zurich Forensic Study (Urbaniok et al., 2007), which longitudinally followed all offenders with either a sentence of at least 10 months or court-ordered therapy who were supervised by the criminal justice system of the Canton as of August 2000 ( $N = 465$ ). The second cohort consisted of all forensic patients receiving treatment in the Psychiatric/Psychological Service who began treatment between January 1, 1997 and December 31, 2009 ( $N = 296$ ). The Psychiatric/Psychological Service is the largest provider of both outpatient and inpatient offender treatment, with approximately 250 violent and sexual offenders receiving services at any one time. To make the cohorts comparable to the SORAG development sample, only adult male offenders ( $n = 740$ ) with contact sex offenses being the index offense ( $n = 267$ ) and who were released into the community with a potential follow-up of seven years ( $n = 168$ ) were included. Upon excluding participants who

died ( $n = 5$ ), were deported before recidivating ( $n = 18$ ), or had more than four missing items on the SORAG ( $n = 8$ ), a final study sample of 137 offenders was obtained.<sup>6</sup>

### **Procedure**

The SORAG was coded based on clinical and criminal justice files by Master's-level psychologists who had attended an accredited Psychopathy Checklist-Revised (Hare, 2003) workshop and were blind to individual offenders' outcomes. The integrity of the SORAG assessments was ensured through the use of a validated and peer-reviewed translation of the instrument (Rossegger, Gerth, Urbaniok, Laubacher, & Endrass, 2010). As recommended by the SORAG authors (Quinsey et al., 2006), item 13 (phallometric test results) was substituted with diagnoses of pedophilia or sadism according to the *Diagnostic and Statistical Manual* of the American Psychiatric Association or the Screening Scale for Pedophilic Interests (Seto, Harris, Rice, & Barbaree, 2004). This systematic substitution was necessary, as phallometric assessments are not legally admissible in Switzerland. Further, there is no validated penile plethysmography test available in Switzerland (Cantonal Court of Zurich, 2012). When items were missing, the prorating algorithm suggested by the SORAG authors was followed. Using this administration strategy, trained raters reached a good interrater reliability of  $\kappa > 0.70$  (Fleiss, Levin, & Paik, 2003; Landis & Koch, 1977)

Following the SORAG manual, recidivism was defined as a new charge and/or conviction for a violent (including sexual) offense within seven years of release. Data on recidivism was based on criminal records, which were last reviewed in 2011. Acts of pseudo-recidivism (e.g., a new charge and/or conviction after an index offense that was precipitated by an incident prior to

---

<sup>6</sup> Offenders who changed their names were still able to be followed.

the index offense) were identified by the construction of crime trajectories for each offender and were not considered acts of recidivism (cf. Quinsey et al., 2006).

### **Statistical Analysis**

Both the discrimination and calibration components of predictive validity were investigated for the SORAG. Discrimination was measured using receiver operating characteristic (ROC) curve analysis and the area under the curve (*AUC*) parameter. To measure calibration, the expected seven year recidivism rates according to the SORAG's published norms (Quinsey et al., 2006) were compared to those rates observed in the present sample both overall and for each of the nine risk bins, individually. This was evaluated using the  $\chi^2$  test to assess the goodness-of-fit between expected and observed recidivism rates per risk bin, as well as Sanders' (1963) decomposition of the Brier score (1950) to provide an index of variation in forecasting. To further investigate the current controversy surrounding the usefulness of base rate-adjusted actuarial models (G. T. Harris & Rice, 2013; Mossman, 2006), we calculated the likelihood ratio (LR) for each SORAG risk bin and compared it to that established for the SORAG development sample (Rice, personal communication, July 17, 2013). Following the guidance of Mossman (2006), we also tested a calibration model with estimated rates obtained by applying Bayes' theory.

In addition to exploring discrimination and calibration validity, a one-sample *t*-test was used to examine the difference in expected and observed percentiles. Further, a Kolmogorov-Smirnov Test (K-S test) was conducted to compare the distribution of risk bins. The K-S test produces a *D* statistic and corresponding *p*-value which are not affected by scale changes but rather serve to capture information on the relative distribution of the SORAG data. All analyses were two-

tailed, used a significance threshold of  $\alpha = 0.05$ , and were conducted using STATA/IC 12.1 for Windows and OSX (StataCorp, 2012).

#### **2.2.4. Results**

##### **Sample Characteristics**

The present study sample was composed of offenders convicted of either child molestation ( $n = 83$ , 60.6%) or rape ( $n = 54$ , 39.4%), with a mean age at conviction of 39.4 years ( $SD = 11.8$ ). The majority of the sample ( $n = 105$ , 76.6%) was enrolled in a treatment program, with most of these offenders having had therapy mandated by the court ( $n = 92$ , 87.6%). Diagnostic criteria for a personality disorder were met by 60 (43.8%) offenders, while 16 (11.7%) met criteria for schizophrenia and 29 (21.2%) for substance abuse or dependency. The base rate for violent (including sexual) recidivism within seven years after release was 16.1% ( $n = 22$ ).<sup>7</sup>

##### **Predictive Validity of the SORAG**

The predictive validity of the SORAG was measured using both discrimination and calibration performance indicators. ROC analyses revealed an acceptable level of discrimination for both SORAG total risk scores ( $AUC = 0.69$ , 95%  $CI = 0.56-0.82$ ,  $p < 0.05$ ) and risk bins ( $AUC = 0.67$ , 95%  $CI = 0.54-0.80$ ,  $p < 0.05$ ). Five of the nine risk bin LRs fell outside of the 95% confidence intervals established for the SORAG development sample (Table 5). This is reflected by the irregular shape of the ROC graph for the ZSOP compared to the developmental sample (Figure 6 and Figure 7). The seven year recidivism rates both overall as well as for each risk bin were considerably lower than for the SORAG norms (Table 5).

---

<sup>7</sup> Stratified by offense type: child sexual abuse ( $n = 13$ , 9.5%), rape ( $n = 6$ , 4.4%), assault ( $n = 3$ , 2.2%), and homicide ( $n = 1$ , 0.7%).

**Table 5.** Normative and observed risk bin distribution and recidivism rates for the SORAG

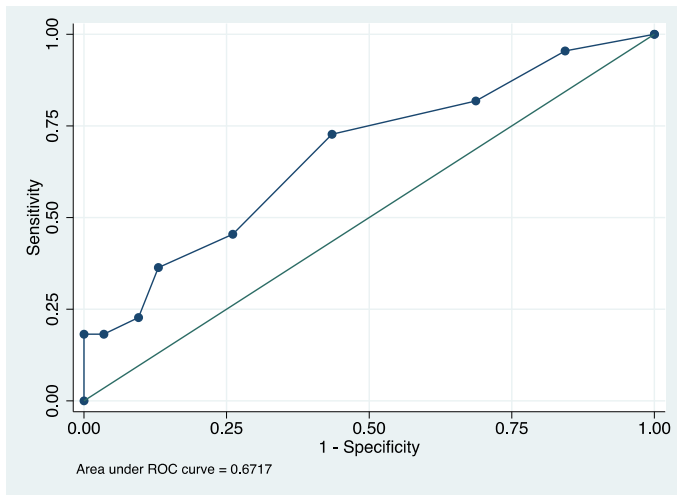
Risk bin	Total risk score	% of sample in each risk bin		Recidivism rate		LR of risk bin (95% intervall)	
		Quinsey et al. (2006) <sup>ab</sup>	ZSOP	Quinsey et al. (2006) <sup>at</sup>	ZSOP	Quinsey et al. (2006) <sup>ac</sup>	ZSOP
1	≤ -10	4.9%	13.9%	7%	5.2%	0.11 (0.01-0.80)	0.29
2	-9 to -4	8.0%	15.3%	15%	14.2%	0.21 (0.06-0.68)	0.87
3	-3 to +2	13.9%	22.6%	23%	6.4%	0.40 (0.20-0.81)	0.36
4	+3 to +8	20.1%	19.0%	39%	23.1%	0.87 (0.54-1.40)	1.57
5	+9 to +14	18.1%	12.4%	45%	11.8%	1.09 (0.67-1.80)	0.70
6	+15 to +19	16.0%	5.1%	58%	42.9%	1.79 (1.05-3.06)	3.92
7	+20 to +24	11.1%	5.8%	58%	12.5%	1.77 (0.92-3.43)	0.75
8	+25 to +30	6.3%	2.9%	75%	0.0%	4.83 (1.63-14.31)	0
9	≥ +31	1.7%	2.9%	100%	100.0%	NA	NA

*Note.* ZSOP = Zurich sex offender population. LR = Likelihood Ratio. N.A. = Not Applicable. Recidivism rate over seven years follow-up.

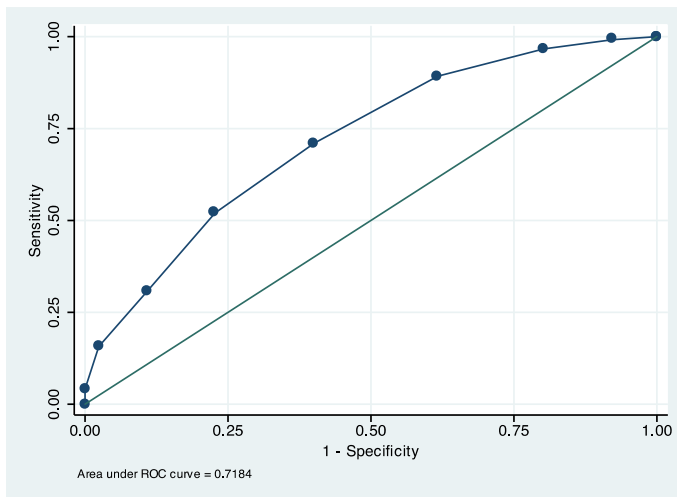
<sup>a</sup>SORAG development sample.

<sup>b</sup>M.E. Rice, personal communication, July 17, 2013.

<sup>c</sup>LRs of the development study were estimated on the basis of the distribution of SORAG risk bins (column two in Table 5) and recidivism rate within risk bins in the development sample (column five in Table 5).



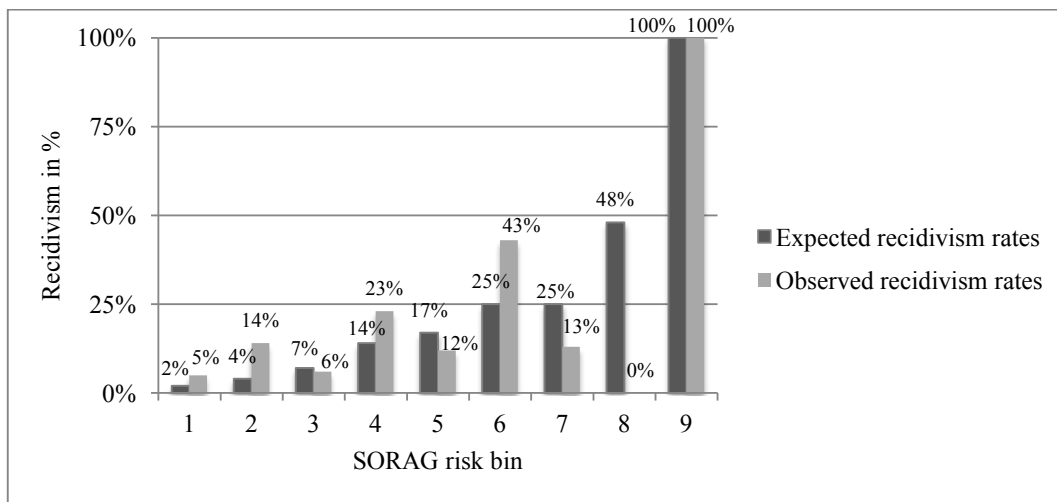
**Figure 6.** Receiver operating characteristic (ROC) graph displaying the discrimination of the SORAG risk bins in the ZSOP



**Figure 7.** Estimated receiver operating characteristic (ROC) graph displaying the discrimination of the SORAG risk bins in the development sample

*Note.* Estimation was based on the raw data provided by M.E. Rice (personal communication, July 17, 2013)

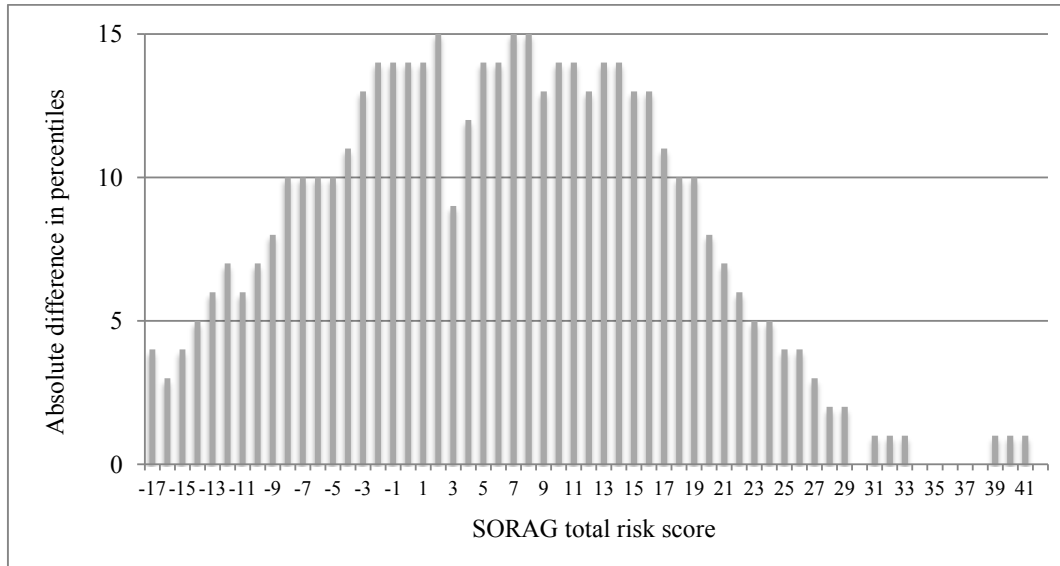
This lack of goodness-of-fit was reflected in a significant  $\chi^2$  statistic and a resolved Sanders’ decomposition score of 0.04, corresponding to an average forecast error of 20.9% per risk bin. The ratio of the excess forecast variance to the minimum forecast variance for the SORAG was 9.3, with ratios higher than 6.0 suggesting “considerably excess variation in forecasts” (Spiegelhalter, 1986, p. 427). When the norms were adjusted with respect to the base rate of the ZSOP sample by applying Bayes’ theory, comparison between the observed and estimated expected recidivism rates showed reduced but still substantial differences. As seen in Figure 8, recidivism rates were under-estimated as well as over-estimated (Sanders’ decomposition score = 0.007; average forecast error of 8.4% per risk bin; ratio of the excess forecast variance to the minimum forecast variance = 6.2).



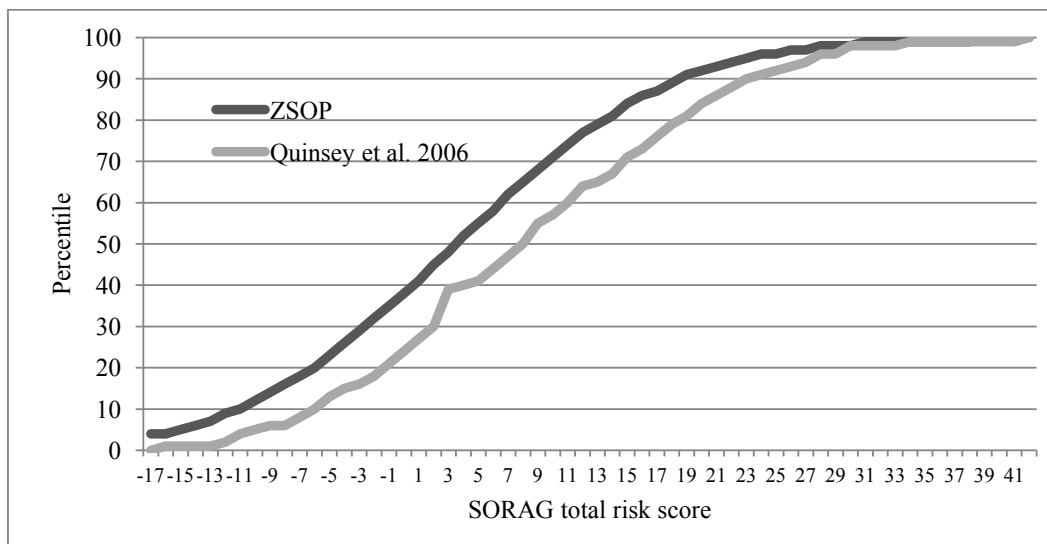
**Figure 8.** Comparing expected and observed recidivism rates by applying Bayes’ theory

The mean SORAG total risk score was +3.4 ( $SD = 11.6$ ), with scores ranging from -16 to +39. This mean score was significantly lower than that in the tool’s development study ( $M = 8.9$ ,  $SD = 11.3$ ),  $t(136) = -5.54$ ,  $p < 0.001$ . Further, the distribution of SORAG scores as exhibited by percentile ranks was significantly different than in the SORAG development study,  $t(59) = 11.1$ ,  $p < 0.001$  (Figure 9 and Figure 10). The mean deviation from the expected percentile was 7.6 ( $SD = 5.3$ ). When a K-S test was conducted to investigate the

difference in the distribution of offenders in the SORAG risk bins between the development study and the present study, a significant effect was found with approximately half (51.8%) of the offenders in the Zurich sex offender population classified into the first three risk bins ( $D = 0.25, p < 0.001$ ).



**Figure 9.** Absolute differences in percentiles between the SORAG development sample and the Zurich sex offender population (ZSOP).



**Figure 10.** Percentiles corresponding to SORAG total risk scores for the tool's development sample and the Zurich sex offender population (ZSOP).

### 2.2.5. Discussion

The aim of the present study was to provide the first examination of the predictive validity of the SORAG in Switzerland. The sample consisted of two total forensic cohorts from the Canton of Zurich, Switzerland that were followed up for seven years in the community. The SORAG was coded using file information for all sex offenders with strict adherence to the manual-based protocol. Recidivism was defined as a new charge and/or conviction for a violent (including sexual) offense. Given the match between the participant and study design characteristics of the present investigation and those of the SORAG development study, rates of predictive validity and percentile distributions were expected to be comparable to those in the tool's development sample.

There were three main findings of the study: First, ROC analyses revealed that the ability of the SORAG to discriminate between recidivists and non-recidivists was lower for the Zurich sex offender population than for the Canadian sex offender population. With an *AUC* of 0.69, it was of moderate performance and lower than in the instrument's development sample in Canada (*AUC* = .75; Quinsey et al., 2006). This finding supports previous research (e.g. Bartosh et al., 2003; Looman, 2006; Nunes et al., 2002) that has found lower *AUCs* for the SORAG compared to those found in the original study. According to the tool's authors, lower *AUCs* can be explained by lack of adherence to the administration protocol published in the SORAG's manual and/or substantial deviation from the study design used in the original study. As we used a total cohort of sex offenders and adopted a study design with a strong replication match, these explanations are not available for the present study. In addition, the comparatively lower base rate in Switzerland compared to Canada is not an explanation, as ROC analyses are robust to base rate variations. We also found that the likelihood ratios of five out of the nine SORAG risk bins differed substantially from those of the instrument's development sample, which led to considerable differences between the two

ROC plots. Hence, there is preliminary evidence suggesting a difference in the SORAG's ability to discriminate between recidivisms and non-recidivists in Switzerland compared to Canada.

Second, calibration analyses suggested that published expected recidivism rates produced by the SORAG should perhaps not be used as official predictions of recidivism rates for sex offenders in Switzerland at this time. Inconsistent fit was found between the expected and observed recidivism rates in all nine SORAG risk bins, with over-estimation errors ranging from 1.4% to 81.8%. This is consistent with prior findings in other German-speaking countries (Eher et al., 2008a; Eher et al., 2008b). Additionally, our findings suggested that calibration problems could not be solved by simply adjusting expected recidivism rates. This concurs with recent research by Harris and Rice (2013) who tested a re-calibration of a related risk assessment tool (the Violence Risk Appraisal Guide) using Bayesian methods. Given these differences, simply adjusting expected rates by taking base rate measures into account may not be as useful as undergoing a jurisdiction-specific re-calibration process.

Third, the distribution of SORAG total risk scores (and hence percentiles) differed from the distribution for the tool's development sample. Thus, while the supposed advantage of actuarial instruments is the provision of population norms to aid in the interpretation of assessment results, this finding suggests that Swiss sex offenders may not be accurately ranked using published guidelines. This is important, as whether an offender is considered to be at the 25<sup>th</sup> or the 75<sup>th</sup> percentile regarding his risk to recidivate may influence decision makers in determining therapeutic resource allocation and release conditions. Albeit published percentile rankings appear to have less use in Switzerland, independent research suggests that such rankings may still be useful in Canada (Barbaree, Langton, & Peacock,

2006). At this time, published expected recidivism rates and percentiles are not advised to be provided in reports on Swiss sexual offenders.

#### **2.2.6. Limitations**

Although we used a total cohort, the investigated sample was nevertheless small. Given the high prevalence of sex offenders receiving treatment in the Swiss criminal justice system, approximately three out of four offenders (76.6%) participated in a court-ordered treatment program. This figure is substantially higher than in the development sample of the SORAG (36.1%). Since studies investigating the effectiveness of sex offender treatment programs have shown preliminary evidence that specific psychotherapeutic interventions – especially those adhering to Risk-Needs-Responsivity principles (see Hanson, Bourgon, Helmus, & Hodgson, 2009 for a review) – may be able to reduce recidivism rates, it could be that the lower recidivism rate in Switzerland is the result of the routine implementation of treatment programs (cf. Lipsey & Cullen, 2007). However, future research into the efficacy of sex offender-specific treatment programs is warranted (Dennis et al., 2012), and it should be kept in mind that the present sample appears to have been an overall lower risk sample than the SORAG development sample. The latter may have affected the base rate of recidivism, as may have the fact that we did not subtract time spent in a criminal justice or mental health institution for a non-violent crime from the total follow-up period for each offender. Hence, appropriate caution is warranted before further research by independent research can be conducted in Switzerland to replicate our findings.

## **2.3. Das Ontario Domestic Assault Risk Assessment (ODARA) – Validität und autorisierte deutsche Übersetzung eines Screening-Instruments für Risikobeurteilungen bei Intimpartnergewalt**

### **2.3.1. Zusammenfassung**

Weltweit wird durchschnittlich jede dritte Frau Opfer von Intimpartnergewalt. Viele Frauen sind wiederholt betroffen. Um dem Risiko erneuter Gewalt effizient zu begegnen, ist ein reliables und valides Instrument zur Risikobeurteilung notwendig. Bis heute wurden mehrere Risk-Assessment Instrumente spezifisch für den Kontext der Intimpartnergewalt publiziert. Eine besondere Anforderung an diese Instrumente ist der breite Kontext, in dem die Instrumente in der Praxis zur Anwendung kommen (z.B. bei Polizei, in Kliniken und Opferberatungsstellen), um entsprechende Fälle zuverlässig triagieren zu können.

Ziel der Entwicklung des kanadischen Ontario Domestic Assault Risk Assessment (ODARA) war es, ein ökonomisches Instrument für die Triage von Niedrigrisiko- und Hochrisiko-Tätern von Intimpartnergewalt bereitzustellen. Bisherige Validierungsstudien weisen auf eine moderate bis gute Trennschärfe des ODARA hin, wobei robuste Befunde für den deutschsprachigen Raum noch ausstehen. In der vorliegenden Arbeit wird eine wissenschaftliche und von den Autoren der Originalversion autorisierte Übersetzung des Instrumentes vorgestellt.

### **2.3.2. Prävalenz von Intimpartnergewalt**

In ihrem Bericht aus dem Jahr 2013 wies die Weltgesundheitsorganisation (WHO) auf Grundlage repräsentativer Befragungen von Frauen aus, dass weltweit durchschnittlich jede dritte Frau mindestens einmal im Leben physische Gewalt durch ihren Partner erfährt (World Health Organisation, 2013). In der Europäischen Union ist jede fünfte Frau betroffen, wobei Deutschland mit einer Prävalenz von 22% im EU-Durchschnitt liegt (Agentur der

Europäischen Union für Grundrechte, 2014). Die überwiegende Mehrheit dieser Gewaltvorfälle bleibt im Dunkelfeld. So werden gemäß der aktuellen EU-Erhebung nur 14% aller gewalttätigen Vorfälle von Männern gegenüber ihren Intimpartnerinnen polizeilich registriert und gelangen somit ins Hellfeld der Justiz. 2012 machte Gewalt zwischen Intimpartnern in Deutschland 13% ( $N = 128'838$ ) aller polizeilich registrierten Delikte gegen Leib und Leben aus (Bundesministerium des Innern, 2013) und in der Schweiz waren es im Jahr 2013 insgesamt 7'345 Fälle von Intimpartnergewalt (inkl. Drohung, Tötlichkeiten und Beschimpfungen), die zu einer Anzeige bei der Polizei führten (Bundesamt für Statistik Schweiz, 2014).

Intimpartnergewalt stellt häufig kein singuläres Ereignis dar, sondern findet in vielen Fällen wiederholt und über einen langen Zeitraum hinweg statt. Eine in der Schweiz durchgeführte repräsentative Befragung von Patientinnen einer Klinik für Geburtshilfe und Gynäkologie konnte zeigen, dass fast jede dritte Frau, die zum Zeitpunkt der Befragung in stärkerem Ausmaß von Intimpartnergewalt betroffen war, diese Gewalt bereits seit drei Jahren oder länger erlebte (Gloor & Meier, 2004). Spezifisch für sexuelle Gewalt in Partnerschaften wies die aktuelle EU-weite Erhebung aus, dass über 50% der betroffenen Frauen die Gewalt bereits mehrfach innerhalb derselben Partnerschaft erlebt hatten (Agentur der Europäischen Union für Grundrechte, 2014). Die (fortgesetzten) Gewalterfahrungen können sich langfristig auf die Gesundheit der Frauen auswirken. Bei Frauen, die Intimpartnergewalt ausgesetzt sind, werden im Vergleich zu Frauen, die keine Gewalterfahrungen gemacht haben, doppelt so häufig psychische Störungen wie Depression, Angststörung und Posttraumatische Belastungsstörung beobachtet (World Health Organisation, 2013). Die Prävalenz psychischer Störungen und jene von Funktionsstörungen (z.B. Konzentrations- und Schlafstörungen) sind bei Frauen, die Opfer von Intimpartnergewalt geworden sind, auch gegenüber den Frauen erhöht, die ausschließlich

außerhalb von intimen Beziehungen Gewalt erfahren haben. Dies gilt insbesondere für das Erleben sexueller Gewalt in Partnerschaften (Agentur der Europäischen Union für Grundrechte, 2014).

Neben den psychischen Langzeitfolgen von Intimpartnergewalt ist auch die Gefahr schwerwiegender physischer Verletzungen groß. Weltweit werden 38% aller Tötungsdelikte an Frauen von einem (ehemaligen) Intimpartner verübt (World Health Organisation, 2013). In Deutschland waren es 2012 17% aller an Frauen begangenen Tötungsdelikte (Bundesministerium des Innern, 2013). Für die Schweiz hat das Bundesamt für Statistik ausgewiesen, dass sich über die Hälfte aller zwischen 2000 und 2004 registrierten (versuchten) Intimpartnertötungen (58%) in einer bestehenden Partnerschaft ereignet hatten (Bundesamt für Statistik Schweiz, 2008). 41% der betroffenen Frauen wurden schon vor der Tat durch den späteren Täter tötlich angegriffen, 53% bedroht oder tötlich angegriffen. Immerhin in über einem Drittel dieser Fälle (39%) war die frühere Gewalt(-drohung) polizeilich bekannt (Bundesamt für Statistik Schweiz, 2008).

### **2.3.3. Risk-Assessment bei Intimpartnergewalt**

Die Prävention von Intimpartnergewalt und ein risikoorientierter Umgang mit den Tätern ist angesichts der hohen Fallzahlen eine Herausforderung für die Polizei und das Justizsystem. Empirische Untersuchungen an gewalttätigen Männern weisen darauf hin, dass Intimpartnergewalt unterschiedlichen Dynamiken folgt und sich distinkte Typen von gewalttätigen Männern unterscheiden lassen. Holtzworth-Munroe et al. (2003) schlugen auf Grundlage einer faktorenanalytischen Untersuchung über verschiedene Merkmale von insgesamt 102 Männern vier Tätertypen vor, die sich im Hinblick auf Rückfallraten, psychiatrische Morbidität und Therapieansprechbarkeit unterscheiden. Diese Befunde legen nahe, dass nicht alle Täter von Intimpartnergewalt ein hohes Rückfallrisiko aufweisen, zur Identifizierung jener jedoch vertiefte Abklärungen notwendig sind und die Implementierung

spezifischer risikosenkender Interventionen angezeigt ist. Doch wie lassen sich abklärungsbedürftige Täter mit einem hohen Rückfallrisiko identifizieren?

Für eine Risikobeurteilung stehen in der Praxis verschiedene Instrumente zur Verfügung: In einer systematischen Übersichtsarbeit wiesen Kilvinger, Rossegger, Urbaniok und Endrass (2012) sechs Risk-Assessment Instrumente aus, die in einer wissenschaftlichen Zeitschrift mit Peer-Review-Verfahren publiziert und darüber hinaus empirisch validiert worden waren. Zu den sechs Instrumenten zählten das Danger Assessment (DA; Campbell, Webster, & Glass, 2009), das Spousal Assault Risk Assessment (SARA; Kropp, Hart, Webster, & Eaves, 1995), das Domestic Violent Screening Instrument (DVSI; Williams & Houghton, 2004), der Violence Risk Appraisal Guide (VRAG; Quinsey et al., 2006), das Ontario Domestic Assault Risk Assessment (ODARA; Hilton, Harris, & Rice, 2010) und der Domestic Violence Risk Appraisal Guide (DVRAG; Hilton, Harris, & Rice, 2010). Während bei den drei erstgenannten trotz strukturierter Datenerfassung die Gesamtbewertung über das Risiko dem Anwender obliegt, ist bei aktuarischen Risk-Assessment Instrumenten – wie VRAG, ODARA und DVRAG – hingegen eine konkrete Auswertungsstrategie vorgegeben (durch a priori definierte Gewichtungen der Items und standardisierte Zusammenführung dieser zu einem Gesamtwert), anhand derer eine Risikoeinschätzung vorgenommen wird. Aktuarische Risk-Assessment Instrumente zeichnen sich zusätzlich dadurch aus, dass der erzielte Gesamtwert darüber hinaus in eine Risikokategorie überführt werden kann, für die Normwerte mit erwarteten Rückfallraten vorliegen (Hanson & Morton-Bourgon, 2009; Hilton, Harris, & Rice, 2010).

Welches Instrument am besten geeignet ist, trennscharf zwischen rückfälligen und nicht rückfälligen Tätern zu unterscheiden, war Gegenstand einer aktuellen Übersichtsarbeit von Messing und Thaller (Messing & Thaller, 2013). Wenngleich die Anzahl bisheriger

Replikationsstudien noch gering ist, wies das ODARA durchschnittlich die höchste Trennschärfe bezüglich erneuter Intimpartnergewalt auf.

#### **2.3.4. Das Ontario Domestic Assault Risk Assessment – ODARA**

##### **Anforderungen aus der Praxis**

Anlass der Entwicklung des ODARA war das Tötungsdelikt eines Mannes in der Provinz Ontario in Kanada, das das öffentliche Interesse in besonderer Weise auf sich zog: Vor dem Tötungsdelikt hatte das Opfer bei der Polizei Anzeige wegen Intimpartnergewalt erstattet. Das Gericht ordnete ein Kontakt- sowie Waffenverbot an und entließ den Täter auf Kautions. Die Anzeige und das eingeleitete Verfahren hielten den Täter jedoch nicht davon ab, weiterhin gewalttätig gegenüber seiner Partnerin zu sein. Er sprach Todesdrohungen aus, kaufte sich eine Waffe, mit der er zunächst seine Partnerin und schließlich sich selbst erschoss. Im Nachgang des Tötungsdelikts wurden Fragen danach laut, weshalb die Gefährlichkeit des Mannes nicht erkannt worden und es misslungen war, das Opfer, das sich bereits hilfeschend an die Polizei gewendet hatte, zu schützen. In der eingeleiteten Aufarbeitung des Falls wurden zwei grundsätzliche Schwachstellen im Umgang mit Intimpartnergewalt identifiziert. Zum einen basierte die Beurteilung des Gerichts auf einer ungenügenden Informationsgrundlage. So war dem Gericht nicht bekannt, dass der Täter vor dem Tötungsdelikt wiederholt (sowohl gegenüber dem späteren Opfer des Tötungsdeliktes als auch gegenüber früheren Partnerinnen) gewalttätig geworden und u.a. wegen häuslicher Gewalt, Waffenbesitz und Nichteinhalten von Bewährungsaufgaben schon bei verschiedenen Behörden registriert war. Zum anderen wurde erkannt, dass mit dem VRAG zwar ein Instrument zur Verfügung steht, anhand dessen sich das Rückfallrisiko gewalttätiger Männer beurteilen lässt, der VRAG aber für einen flächendeckenden Einsatz als Screening-Instrument zum Beispiel durch die Polizei aber auch durch Klinikpersonal als erste Stufe

einer Risikoanalyse aufgrund der für die Anwendung erforderlichen Informationsgrundlage (z.B. Einschätzung über psychiatrische Merkmale des Täters) nicht geeignet ist. Die Regierung der Provinz Ontario veranlasste schließlich erstens die Überarbeitung von Kommunikationsstrukturen des Justizsystems mit dem Ziel, einen lückenlosen Informationsaustausch zu etablieren. Zweitens gab sie die Entwicklung eines spezifischen Instruments für Risikobeurteilungen im Kontext von Intimpartnergewalt in Auftrag. Das Instrument sollte einen breiten Einsatz in der Praxis erlauben (d.h. einem sogenannten „front-line“-Instrument gerecht werden) und eine geeignete Grundlage für eine systematische Fall-Triagierung darstellen. Es galt also ein sensitives Screening-Instrument zu entwickeln, das es erlaubt, jene Täter zu identifizieren, bei denen eine vertiefte Abklärung indiziert ist. Um dem Anspruch eines „front-line“-Instruments gerecht zu werden, sollte das Instrument in der Anwendung ökonomisch sein (kurz und einfach anzuwenden); die für die Anwendung notwendigen Informationen sollten einfach zugänglich sein (Fokus auf Informationen zum Vorfall häuslicher Gewalt und Informationen, die über das Opfer eruiert sind); vertiefte Informationen, wie beispielsweise psychiatrische oder psychologische Einschätzungen, sollten zunächst nicht abgefragt werden, um ein schnelles Screening zu ermöglichen (Hilton, Harris, & Rice, 2010).

### **Entwicklung des ODARA**

Grundlage für die Entwicklung des ODARA bildete eine empirische Untersuchung über Prädiktoren für Rückfälligkeit bei Intimpartnergewalt. Da bei der überwiegenden Mehrheit der polizeilich registrierten Fälle von Intimpartnergewalt der Täter männlich und das Opfer weiblich ist, entschieden sich die Autoren des ODARA bei der Entwicklung des Instruments ausschließlich männliche Täter zu berücksichtigen (Hilton, Harris, & Rice, 2010). Ausgangslage für die Entwicklung des Instruments bildete ein Variablenpool, der auf Grundlage der in der Literatur beschriebenen Prädiktoren definiert wurde. Hierbei handelte

sich um Variablen zur Soziodemographie und kriminellen Vorgeschichte des Täters, zur Beziehung zwischen Täter und Opfer sowie zu konkreten Tatmerkmalen und Charakteristika des Opfers, wobei all jene 38 Variablen weiter berücksichtigt wurden, die sich in der Praxis einfach erheben lassen. Der Zusammenhang der Variablen mit dem Rückfallkriterium wurde mittels multivariabler Regressionsanalysen untersucht. Dafür wurde eine repräsentative Stichprobe von 589 Tätern aus der Provinz Ontario in Kanada verwendet. In das finale multivariable Modell, welches den dreizehn Items des ODARA entspricht, wurden jene Variablen aufgenommen, die erstens in der Mehrheit gezogener Zufallsstichproben und zweitens auch in der Gesamtstichprobe in einem signifikanten Zusammenhang mit dem Rückfallkriterium standen und darüber hinaus einen eigenständigen von den anderen Items unabhängigen Erklärungswert bezüglich des Rückfallkriteriums aufwiesen (Hilton et al., 2004).

### **Anwendungsbereich des ODARA**

Der Anwendungsbereich des ODARA ist eng definiert: Das ODARA kann in allen polizeilich registrierten Fälle häuslicher Gewalt angewendet werden, bei denen der Täter männlich und das Opfer eine aktuelle oder frühere Lebenspartnerin des Täters ist. Ferner muss es sich um einen physischen Übergriff oder einer Todesdrohung unter Verwendung einer Waffe handeln (Hilton, Harris, & Rice, 2010). Das ODARA kann demnach beispielsweise nicht angewendet werden, wenn der Täter ausschließlich Gegenstände in der Wohnung zerstört (Sachbeschädigung) oder per Telefon physische Gewalt angedroht hat (Hilton, Harris, & Rice, 2010).

### **Anwendung des ODARA**

Das ODARA setzt sich aus dreizehn Items zusammen, die auf einer dichotomen Skala („ja“ [1 Punkt] versus „nein“ [0 Punkte]) beurteilt werden. Maximal können im ODARA

entsprechend dreizehn Punkte erreicht werden. Allen Items gemein ist, dass sie einen statischen Charakter aufweisen, d.h. ausschließlich Informationen betreffen, die sich auf den Zeitraum bis zum Index-Übergriff beziehen und damit unveränderlich sind (Hilton, Harris, & Rice, 2010).

Zur Auswertung des ODARA werden die in den dreizehn Items erzielten Punkte addiert. Der Summenwert kann in eine von sieben Risikokategorien überführt werden, die positiv mit dem Rückfallrisiko korrelieren. Für die Risikokategorien stehen erwartete Rückfallraten für einen Beobachtungszeitraum von fünf Jahren nach dem Index-Übergriff zur Verfügung (siehe Tabelle 6), wobei Rückfälligkeit als erneute Intimpartnergewalt definiert wurde. Darüber hinaus kann anhand einer Tabelle abgelesen werden, welcher Anteil der Täter aus der Entwicklungsstichprobe einen niedrigeren, den gleichen oder einen höheren Summenwert im ODARA erzielt hat als der durch den Anwender beurteilte Täter (Hilton, Harris, & Rice, 2010).

**Tabelle 6:** Normwerte des ODARA\*

ODARA-Wert	Rückfallrate (%)	Anteil in derselben Kategorie (%)	Anteil in einer niedrigeren Kategorie (%)	Anteil in einer höheren Kategorie (%)
0	7	9	0	91
1	17	17	9	74
2	22	21	26	53
3	34	20	47	33
4	39	13	67	20
5-6	53	14	80	6
7-13	74	6	94	0

\*gemäss Hilton, Harris, and Rice (2010)

Bei der Anwendung des ODARA dürfen bis zu fünf Items wegen fehlender Informationen ausgelassen werden („Missings“). Der Summenwert muss dann unter

Berücksichtigung der Anzahl ausgelassener Items und des in den beantworteten Items erzielten Summenwerts korrigiert werden. Für diesen Auswertungsschritt liegt eine entsprechende Kreuztabelle vor, die den Summenrohwert und die Anzahl der ausgelassenen Items gegenüberstellt und korrigierte Summenwerte enthält (siehe Tabelle 2; Hilton, Harris, & Rice, 2010). Siehe Tabelle 7.

**Tabelle 7:** Korrigierte Summenwerte des ODARA bei fehlender Information\*

ODARA-Rohwert	Anzahl der ausgelassener Items („Missings“)				
	1	2	3	4	5
0	0	0	0	0	0
1	1	1	1	1	2
2	2	2	3	3	3
3	3	4	4	4	5
4	4	5	5	6	7+
5	5	6	7+	7+	7+
6	7+	7+	7+	7+	7+

\* gemäss Hilton, Harris, and Rice (2010)

### **Interpretation des ODARA-Ergebnisses**

In der Entwicklungsstudie zeigte sich, dass der Summenwert nicht nur positiv mit einschlägiger Rückfälligkeit korreliert, sondern auch mit der Schwere und Häufigkeit von Rückfällen (Hilton et al., 2004).

Darüber hinaus bieten die für die Risikokategorien hinterlegten Rückfallraten eine Interpretationshilfe an: Der Einzelfall kann mit einer Gruppe von Tätern verglichen werden, die in dieselbe Risikokategorie fielen und bei denen der Verlauf der Legalbewährung bekannt ist. Entsprechend könnte eine Ergebnisdarstellung einer ODARA-Anwendung folgendermaßen lauten:

*„Herr S. erreichte einen ODARA-Summenwert von X. Dieser Summenwert entspricht der X. von 7 Risikokategorien. Gemäß der vorliegenden Normen (siehe Tabelle 6), erreichten weniger als XX% einer repräsentativen Vergleichsgruppe einen höheren ODARA-Summenwert als Herr S. Bei einer Beobachtungszeit von durchschnittlich fünf Jahren wurden XX% derjenigen Täter der Vergleichsgruppe mit einem erneuten polizeilich registrierten Vorfall von Intimpartnergewalt rückfällig, die derselben Risikokategorie wie Herr S. zugeordnet wurden.“*

Auf Grundlage einer Vielzahl von Forschungsbefunden ist davon abzuraten, den ODARA-Summenwert klinisch anzupassen (z.B. durch eine andere Gewichtung der Items im Einzelfall oder die Hinzunahme weiterer relevant erscheinender Merkmale). Beispielsweise konnten Hilton und Simmons (2001) eindrücklich aufzeigen, dass eine klinische Anpassung eines aktuarisch ermittelten Ergebnisses zu weniger validen Risikobeurteilungen führte. Als mögliche Ursache kann die Tendenz von Beurteilern in Betracht gezogen werden, spezifische Merkmale des Einzelfalls besonders zu gewichten, ohne dass dies ihrer tatsächlichen Relevanz für das Rückfallrisiko entspricht.

### **Anwendung in der Praxis**

Das ODARA kann im Falle eines Vorfalls von Intimpartnergewalt angewendet werden – unabhängig davon, ob die Voraussetzungen für das Einleiten eines Strafverfahrens erfüllt sind. Wichtige Informationsgrundlagen für die Anwendung des ODARA stellen zum einen Angaben des Opfers dar, zum anderen Dokumente, die über frühere Gewalthandlungen (im und außerhalb des häuslichen Kontextes) und Vorstrafen (einschließlich des Sanktions- und Bewährungsverlaufs) des Täters Auskunft geben. Eine angemessene Befragung des Opfers und der Zugriff auf Informationen aus dem Bundeszentralregister oder polizeilichen Informationssystemen sind daher für die Wertung des ODARA essentiell (Hilton, Harris, & Rice, 2010).

Gleichwohl das ODARA explizit als „front-line“-Instrument entwickelt wurde, das berufsgruppenunabhängig angewendet werden kann, empfehlen die Autoren ein spezifisches Anwendungstraining vor der Anwendung in der Praxis. Hilton et al. (2007) konnten diesbezüglich aufzeigen, dass ein absolviertes Training vor allem bei Wertungen, die auf Grundlage eines Interviews durchgeführt werden, die Anzahl von Anwendungsfehlern signifikant verringert.

### **Verwendung des ODARA als Screening-Instrument**

Das ODARA wurde spezifisch als Screening Instrument entwickelt und bildet die erste Stufe eines Risk-Assessment Systems, das sich aus dem ODARA und dem DVRAG zusammensetzt. Bei einer kombinierten Anwendung beider Instrumente wird eine vertiefte Abklärung für Fälle empfohlen, die in die Kategorien 3-7 des ODARA fallen, da die Spezifität des ODARA vor allem in den höheren Risikokategorien derjenigen des DVRAG unterlegen ist. Niedrig-Risiko-Täter werden hingegen zuverlässig über die ersten zwei Kategorien anhand des ODARA erfasst (Hilton, Harris, & Rice, 2010). Erscheint eine vertiefte Abklärung indiziert, werden anhand des DVRAG spezifische Informationen differenzierter erhoben und zusätzlich die Ausprägung psychopathischer Merkmale des Täters überprüft (Hilton, Harris, & Rice, 2010).

### **Limitationen des ODARA**

Bei der Interpretation des ODARA Ergebnisses gilt es stets zu berücksichtigen, dass es sich bei dieser Art der Beurteilung um ein Screening handelt. Eine Risikobeurteilung anhand des ODARA kann eine forensisch-psychiatrische Beurteilung nicht ersetzen. Das ODARA ist der Ausgangspunkt und nicht das Schlussergebnis eines möglicherweise umfangreichen Beurteilungsprozesses, der von Experten unterschiedlicher professioneller Provenienzen mitgestaltet wird.

### 2.3.5. Validität des ODARA

#### **Methodik der Literaturrecherche**

Um Aussagen über die Validität des ODARA treffen zu können, wurde eine systematische Literaturrecherche durchgeführt. Die Suche erfolgte im März 2014 in den Suchmaschinen „pubmed“, „PsycINFO“, „PsycINDEX“, „Medline“, „ERIC“ sowie „Google Scholar“ unter Verwendung des Stichworts „ODARA“ beziehungsweise „Ontario Domestic Assault Risk Assessment“. Weiter berücksichtigt wurden Publikationen, die Aspekte der Validität des ODARA empirisch untersucht haben und in einem peer-review Verfahren publiziert worden sind ( $n = 6$ ). Die so identifizierten Publikationen wurden um Publikationen ergänzt, die in der „ODARA-Bibliographie“ der Autoren des Instruments aufgeführt waren und den Einschlusskriterien entsprachen ( $n = 4$ ; Waypoint Centre for Mental Health Care, 2014)). Gesamthaft liegen heute 10 Replikationsstudien zur Validität des ODARA vor, die im englisch- oder deutschsprachigen Raum im peer-review Verfahren veröffentlicht worden sind. Tabelle 8 enthält eine Übersicht zu den Ergebnissen der Replikationsstudien des ODARA (Stand März 2014).

**Tabelle 8:** Validierungsstudien zum ODARA (Stand März 2014)

Autoren	Jahr	Land	N (Geschlecht)	Fragestellung	Rückfallkriterium	Reliabilität	AUC (95%-KI)	Andere/weitere Validitätsaspekte
Hilton et al. (2014)	2014	CA	30 (w)	Trennschärfe des ODARA bei Täterinnen von IPG	Anklage wegen IPG	ICC = .94	.72 (.50-.94)	Es liegt kein Zusammenhang des ODARA-Summenwertes mit der Schwere des Rückfalls vor.
Folkes et al. (2013)	2013	CA	1421 (m)	Zusammenhang zwischen Waffenverwendung oder -zugang beim Indexdelikt und dem Rückfallrisiko	polizeilich registrierte IPG oder Anklagen/Verurteilungen wegen IPG	k.A.	n.z.	Zusätzliche Informationen über die Verwendung von Waffen oder Zugang zu Schusswaffen verbessert das Risk-Assessment anhand des ODARA nicht.
Oliszowy et al. (2013)	2013	CA	40 (m)	Differenzierung zwischen Rückfälligkeit mit Filizid und Intimidid anhand des ODARA	Filizid, Intimidid	k.A.	n.z.	Das ODARA ist nicht diskriminativ valide bezüglich Filizid.
Rettenberger & Eher (2013)	2013	AT	66 (m)	Trennschärfe des ODARA bei sexuell motivierter Intimpartnergewalt	Verurteilung wegen IPG	k.A.	.71 (.58-.84)	
Eke et al. (2011)	2011	CA	13 (m)	Validierung des ODARA bzgl. (versuchter) Intimpartnerötung	(versuchte) Intimpartnerötung	r = .98	k.A.	92% der Täter, die mit einer (versuchten) Intimpartnerötung rückfällig wurden, fielen in die höchste Risikokategorie des ODARA.
Hilton et al. (2010)	2010	CA	150 (m)	Trennschärfe des ODARA bei inhaftierten Tätern häuslicher Gewalt	Anklage wegen IPG	ICC = .95	.64 (.54-.73)	

*Legende.* IPG = Intimpartnergewalt, k.A. = keine Angaben, n.z. = nicht zutreffend, Land = Land in dem die Untersuchung durchgeführt worden ist: CA = Kanada, AT = Österreich, N = Stichprobengröße, Fragestellung = Untersuchte Fragestellung bzw. Ziel der Untersuchung, Rückfallkriterium = Art der Rückfälligkeit (z.B. polizeilich registrierte Intimpartnergewalt oder Anklagen/Verurteilungen wegen Intimpartnergewalt), Reliabilität = Reliabilität des ODARA-Summenwertes (r = Pearson's Correlation, ICC = Intraclass Correlation Coefficient), AUC = Trennschärfe des ODARA (abgebildet über die Area under the Curve [AUC]), 95%-KI = 95%-Konfidenzintervall

**Tabelle 8 weitergeführt:** Validierungsstudien zum ODARA (Stand März 2014)

Autoren	Jahr	Land	N (Geschlecht)	Fragestellung	Rückfallkriterium	Reliabilität	AUC (95%-KI)	Andere/weitere Validitätsaspekte
Hilton & Harris (2009)	2009	CA	309 (m)	Trennschärfe des ODARA bzgl. erneuter IPG bei einem Vergleich zwischen einschlägig rückfälligen und Tätern, die mit keinem Gewaltdelikt irgendeiner Art rückfällig wurden	polizeilich registrierte IPG oder Verurteilungen wegen IPG	$r > .90$	.74 (.68-.80)	
			391 (m)	Trennschärfe des ODARA bzgl. erneuter IPG unter Berücksichtigung aller Täter			.67 (.61-.73)	
Hilton et al. (2008)	2008	CA	346 (m)	Trennschärfe des ODARA im Rahmen der Entwicklung des DVRAG	polizeilich registrierte IPG oder Verurteilungen wegen IPG	$r \geq .90$	.65 (.59-.71)	
Hilton et al. (2008)	2008	CA	108 Opfer (w)	Konstruktvalidität des ODARA bei Befragung einer hospitalisierten Opferpopulation	n.z.	$r > .99$	n.z.	Das ODARA ist zur Anwendung auf Grundlage von Opferbefragungen geeignet.
Hilton et al. (2007)	2007	CA	522 (m)	Der Effekt von Verhaftungen auf das IPG-Rückfallrisiko	polizeilich registrierte IPG	$r \geq .90$	n.z.	Verhaftungen weisen keine inkrementelle Validität gegenüber dem ODARA auf. Verhaftungen korrelieren mit der Schwere des Index-Übergriffs und der Beurteilung des Rückfallrisikos vor der Verhaftung.

*Legende.* IPG = Intimpartnergewalt, k.A. = keine Angaben, n.z. = nicht zutreffend, Land = Land in dem die Untersuchung durchgeführt worden ist: CA = Kanada, AT = Österreich, N = Stichprobengröße, Fragestellung = Untersuchte Fragestellung bzw. Ziel der Untersuchung, Rückfallkriterium = Art der Rückfälligkeit (z.B. polizeilich registrierte Intimpartnergewalt oder Anklagen/Verurteilungen wegen Intimpartnergewalt), Reliabilität = Reliabilität des ODARA-Summenwertes ( $r$  = Pearson's Correlation, ICC = Intraclass Correlation Coefficient), AUC = Trennschärfe des ODARA (abgebildet über die Area under the Curve [AUC]), 95%-KI = 95%-Konfidenzintervall

**Kriteriumsvalidität des ODARA**

Die Kriteriumsvalidität eines Risk-Assessment Instruments setzt sich aus zwei Aspekten zusammen: Der Trennschärfe und der Kalibrierung (Endrass et al., 2009). Dabei gibt die Trennschärfe die Fähigkeit eines Instruments an, zwischen rückfälligen und nicht rückfälligen Tätern zu diskriminieren. Die Trennschärfe kann dabei als eine Funktion der Spezifität und Sensitivität des Instruments über alle möglichen cut-off-Werte des Entscheidungskriteriums (z.B. ODARA-Summenwert) hinweg betrachtet werden. Sie wird häufig über die Area Under the Curve [*AUC*] ausgewiesen. Demgegenüber stellt Kalibrierung ein Maß für die Übereinstimmung erwarteter und beobachteter Rückfallraten innerhalb verschiedener Risikokategorien dar (Endrass et al., 2009; Rossegger et al., 2014). In der Entwicklungsstichprobe erreichte das ODARA eine *AUC* von .77 (Hilton et al., 2004).

Sechs der zehn Validierungsstudien des ODARA untersuchten die Trennschärfe des Instrumentes. Mit einer Trennschärfe zwischen  $AUC=.64$  (Hilton, Harris, Popham, et al., 2010) und  $AUC=.74$  (Hilton & Harris, 2009) kann auf Grundlage bisheriger Ergebnisse von einer moderaten bis guten Trennschärfe des ODARA gesprochen werden (siehe Tabelle 3; Rice & Harris, 2005). Nur eine der Validierungsstudien (Rettenberger, 2013) wurde unabhängig von den Autoren des ODARA durchgeführt und nur drei der Validierungsstudien (Hilton, Harris, Popham, et al., 2010; Hilton, Harris, Rice, et al., 2008; Hilton et al., 2014) können als „echte“ Replikationen charakterisiert werden, da nur diese sich streng am Studienprotokoll der Entwicklungsstudie orientieren, d.h. bezüglich Art der Stichprobe, Länge des Beobachtungszeitraumes, Art des Rückfalls etc. mit den Charakteristika der Entwicklungsstichprobe übereinstimmen und somit für vergleichende Analysen geeignet sind (Rossegger, Gerth, Seewald, et al., 2013).

Rettenberger & Eher (2013) wiesen für eine Stichprobe von 66 in Österreich inhaftierten Straftätern, die ein sexuell motiviertes Delikt gegenüber ihrer (ehemaligen) Intimpartnerin

begangen haben, eine *AUC* von .71 aus. Eine etwas höhere Trennschärfe des ODARA wiesen Hilton et al. (2009) für eine Stichprobe von 309 Tätern bezüglich erneuter Intimpartnergewalt auf (*AUC* = .74). Moderate Trennschärfen (*AUC*=.64 bzw. *AUC*=.65) zeigten sich in zwei weiteren kanadischen Replikationsstudien, die 150 inhaftierte bzw. 346 polizeilich registrierte Täter untersuchten (Hilton, Harris, Popham, et al., 2010; Hilton, Harris, Rice, et al., 2008).

Die Fragestellungen der anderen vier Studien (siehe Tabelle 8) gingen Fragen der inkrementellen Validität (Folkes et al., 2013; Hilton, Harris, & Rice, 2007), der Konstruktvalidität (Hilton, Harris, & Holder, 2008) und der diskriminanten Validität nach (Olszowy et al., 2013). Folkes et al. (2013) griffen eine am ODARA geäußerte Kritik auf und prüften, ob die zusätzliche Berücksichtigung zweier Informationen – Zugang zu Schusswaffen und Einsatz von Waffen im Index-Übergriff – zu einer Verbesserung der Trennschärfe des ODARA führt. Die ergänzten Kriterien erhöhten die Validität der Risikobeurteilung jedoch nicht (Folkes et al., 2013).

In der Entwicklungsstichprobe korrelierten der ODARA-Summenwert und die ODARA-Risikokategorie auch mit der Schwere des Rückfalls, die über den Verletzungsgrad des Opfers operationalisiert wurde (Hilton et al., 2004). Eke et al. (2011) konnten diesen Befund indirekt replizieren, indem sie in einer retrospektiven Untersuchung an Männern, die ein (versuchtes) Tötungsdelikt an ihrer (ehemaligen) Partnerin begangen haben, das ODARA für vorangegangene Vorfälle häuslicher Gewalt anwendeten. Zwölf von dreizehn Tätern wurden der höchsten ODARA-Kategorie zugeordnet.

In der aktuellsten Validierungsstudie erhoben Hilton et al. (2014) das ODARA bei einer Stichprobe von 30 Frauen, die gegenüber ihrem Intimpartner gewalttätig wurden (Hilton et al., 2014). Zwar wies das ODARA eine gute Trennschärfe auf (*AUC*=.72), allerdings war das Konfidenzintervall sehr groß (95%-Konfidenzintervall=0.50-0.94) und der ODARA-

Summenwert korrelierte positiv aber nicht signifikant mit erneuter polizeilich registrierter Intimpartnergewalt. Grund hierfür könnte die sehr kleine Stichprobe der Untersuchung sein.

### **Kalibrierung und Grenzwerte**

Ob die in den Normtabellen aufgeführten erwarteten Rückfallraten pro Risikokategorie den in anderen Stichproben erzielten beobachtbaren Rückfallraten entsprechen, wurde für den ODARA bislang nicht empirisch untersucht. Über die Kalibrierung des ODARA kann entsprechend noch keine Aussage getroffen werden.

Damit einhergehend ist auch die Überprüfung des Grenzwertes von Risikokategorie 3 zur Indikation einer vertieften Risikoabklärung noch offen. Eine Aussage über Rückfallwahrscheinlichkeiten pro Risikokategorie und einem damit einhergehenden Grenzwert von Risikokategorie 3 zur Indikation einer spezifischeren Risikoabklärung kann daher unter der aktuellen Befundlage nicht zuverlässig getroffen werden.

### **2.3.6. Zusammenfassung und Schlussfolgerung für die Praxis**

Das ODARA ist ein Risk-Assessment Instrument für Risikobeurteilungen in Fällen von Intimpartnergewalt. Als „front-line“-Verfahren ist es für einen breiten Anwenderkreis geeignet, zeichnet sich durch eine ökonomische Anwendung aus und ist daher als Ausgangspunkt – d.h. erste Stufe – einer Risikoanalyse geeignet. Nach gegenwärtigem Stand der Forschung weist es die durchschnittlich höchste Trennschärfe zur Beurteilung des Risikos erneuter Intimpartnergewalt im Vergleich zu anderen Verfahren, deren Ergebnisse im Rahmen von peer-reviewed veröffentlichten Untersuchungen vorliegen, auf. Das ODARA zeichnet sich als sensitives Screening-Instrument zur Triage von Niedrigrisiko-Tätern und Hochrisiko-Tätern aus, bei denen eine weiterführende Abklärung notwendig ist. Die Überprüfung der durch die Autoren des Instruments ausgewiesenen Risikonormen steht noch aus, weshalb eine Zuordnung spezifischer Risikoraten zu Risikokategorien bisher nur unter

Vorbehalt getroffen werden sollte. Für die Anwendung des Instruments im deutschen Sprachraum steht mit der vorliegenden Publikation eine validierte und autorisierte Übersetzung inklusive konkreter Hinweise zur praktischen Anwendung gemäß ODARA-Manual zur Verfügung.

### **2.3.7. Take Home Message**

Das ODARA ist ein aktuarisches Risk-Assessment Instrument zur Beurteilung des Rückfallrisikos von Intimpartnergewalt. Mit einer moderaten bis guten Fähigkeit zwischen nicht rückfälligen und rückfälligen Tätern zu diskriminieren, weist es gegenüber anderen Risk-Assessment Verfahren die durchschnittlich höchste Trennschärfe auf. Das ODARA ist ökonomisch und auf Grundlage weniger Informationen anzuwenden, wobei es als sensitives Screening-Instrument vor allem zur Triagierung und gegebenenfalls Empfehlung weiterführender und vertiefender Abklärungen geeignet ist.

### **2.3.8. Anhang: Deutsche Übersetzung des ODARA**

#### **Vorbemerkung zur Übersetzung**

Die Übersetzung des ODARA ins Deutsche basiert auf der 2010 publizierten und überarbeiteten Version des Instruments (Hilton, Harris, & Rice, 2010). Das englische Original des ODARA wurde erst ins Deutsche und die Deutsche Version dann ins Englische „rückübersetzt“. Die Übersetzer hatten keine Kenntnis vom Originaltext. Unterschiede zwischen dem englischen Original und der englischen Rückübersetzung wurden mit N. Zoe Hilton diskutiert und der Text entsprechend adaptiert.

Unser Dank gilt den Entwicklern des ODARA, N. Zoe Hilton, Grant T. Harris, Marnie E. Rice, Carol Lang, Catherine A. Cormier und Kathryn J. Lines für die Autorisierung der Übersetzung des ODARA in die deutsche Sprache sowie N. Zoe Hilton, Carol Lang und Elke Ham für die Unterstützung im Prozess der Übersetzung. Weiterer Dank gilt Jay P. Singh und

Mitarbeiterinnen des Worldwide Translation Services (WTT) für die Rückübersetzung der deutschen Version ins Englische sowie Martin Rettenberger für den kollegialen Austausch über die von ihm, Kathrin Gaunersdorfer und Reinhard Eher (2009) verfasste Übersetzung des ODARA. Diese frühere deutsche Übersetzung basiert auf der ersten Version des ODARA aus dem Jahr 2005 (Mental Health Centre Penetanguishene, 2005) und weist gegenüber dem englischen Original eine leicht modifizierte Operationalisierung einzelner Items auf.

### **Anwendungsbereich des ODARA – Definition des Index-Übergriffs**

Der Index-Übergriff sollte für jedes Assessment der zeitnahste polizeilich bekannte Vorfall sein, bei dem der Mann gewalttätig gegenüber einer Lebenspartnerin wurde. Der Index-Übergriff erfordert physischen Kontakt mit dem Opfer oder eine glaubhafte Todesdrohung mit einer Waffe in der Hand in ihrer Anwesenheit. Eine Lebenspartnerin ist definiert als eine Frau, mit der der Täter verheiratet ist oder war, in einer Lebensgemeinschaft lebt oder lebte oder für irgendeine Zeitperiode zusammenlebt oder zusammenlebte.

Das ODARA kann ebenso bei einem Fall angewendet werden, bei dem ein Übergriff gegen eine Frau vorliegt, mit der der Mann zwar keine Lebenspartnerschaft führte, aber in einer intimen Beziehung ist oder war.

Damit ein Vorfall häuslicher Gewalt als Index-Übergriff gelten kann, ist es nicht notwendig, dass deswegen Anklage erhoben wurde. Falls beim zeitnahsten Vorfall nicht eindeutig Gewalt gegen eine Partnerin involviert war, kann ein früherer Vorfall, der die Einschlusskriterien erfüllt, als Index-Übergriff verwendet werden.

Beurteilende sollten jeden Ereignisbericht der Polizei als einen separaten Vorfall zählen. Bei einem Cluster von Übergriffen, die der Polizei zum selben Zeitpunkt berichtet werden, ist der zeitnahste infrage kommende Übergriff, der mindestens 24 Stunden nach einem früheren Übergriff stattfand, der Index-Übergriff. Mehrere infrage kommende Übergriffe innerhalb

von 24 Stunden zählen als ein Index-Übergriff. Diese Cluster-Regel gilt auch, um frühere Vorfälle häuslicher und nicht-häuslicher Gewalt zu identifizieren.

### **Definition von Gewalt zur Erhebung von Gewalthandlungen**

Jede der folgenden Handlungen wird in die Definition von Gewalt eingeschlossen: hielt sie fest, warf etwas nach ihr, das für sie schmerzhaft sein könnte, verdrehte ihren Arm oder zog an ihren Haaren, stieß oder schubste sie, packte sie (schließt ein, an ihr zu zerren und sie hinter sich herzuführen), ohrfeigte sie (schließt ein, sie zu schlagen), fügte ihr andere geringfügige Gewalt zu (z.B. schüttelte sie), schlug sie mit der Faust oder etwas, das für sie schmerzhaft sein könnte, würgte sie (schließt ein, sie am Nacken oder Hals zu packen, sie in den Schwitzkasten zu nehmen), schlug sie heftig gegen eine Wand, „schlug sie zusammen“, verbrannte oder verbrühte sie absichtlich, trat sie, verwendete ein Messer oder eine Schusswaffe gegen sie (d.h. tatsächlicher oder versuchter Kontakt mit dem Körper des Opfers, schließt ein, eine Schusswaffe zu entladen, während mit dieser auf das Opfer gezeigt wird oder die Androhung physischer Schädigung mit einer Waffe in der Hand), und fügte ihr andere schwere Gewalt zu (z.B. hob sie auf und stieß sie weg, versetzte ihr einen Kopfstoß, stieß sie die Treppen hinunter, biss sie).

Zusätzlich gilt jede Form von Druck als Gewalt, mit der das Opfer gegen ihren Willen zu einem sexuellen Kontakt gezwungen wurde. Dieselben Kriterien für Gewalt gelten für die Kodierung von Vorfällen früherer häuslicher oder nicht-häuslicher Gewalt.

### **Kriteriumsvariable – Definition von Rückfälligkeit**

Vorfälle, die sich zeitlich nach dem Index-Übergriff ereigneten und bei denen eine Gewalthandlung gemäß Definition unter Abschnitt ‚Definition von Gewalt zur Erhebung von Gewalthandlungen‘ stattfand, erfüllen die Definition von Rückfälligkeit mit häuslicher Gewalt.

**ODARA-Items*****Item 1: Früherer häuslicher Vorfall***

Ein häuslicher Vorfall ist definiert als ein Vorfall, bei dem der zu beurteilende Mann seine aktuelle oder frühere Lebenspartnerin und/oder deren Kinder tätlich angegriffen hat und der Übergriff in einem Polizeibericht oder im Vorstrafenregister verzeichnet ist.

Alle der folgenden Kriterien müssen vorliegen:

- Eine Gewalthandlung (wie oben definiert), die vom zu beurteilenden Mann verübt wurde
- und ein Vorfall, der zu einem separaten Anlass vor dem Index-Übergriff stattfand
- und Beteiligung der Polizei oder ein anschließender Bericht an die Polizei.

Falls eines dieser Kriterien fehlt, bewerten Sie das Item mit „0“, auch wenn jedes andere Kriterium vorliegt. Alle Kriterien müssen erfüllt sein.

Mindestens ein Kriterium der folgenden muss vorliegen:

- Ein Opfer, das eine aktuelle oder frühere Lebensgefährtin des zu beurteilenden Mannes ist
- und/oder ein Opfer, das das Kind der aktuellen oder einer früheren Lebensgefährtin des Mannes ist.

Eines dieser Kriterien muss vorliegen, um das Item mit „1“ zu werten. Es müssen nicht beide vorliegen. Falls keines dieser Kriterien erfüllt ist, bewerten Sie das Item mit „0“.

Ein früherer Übergriff, der vom Opfer erst zum Zeitpunkt des Index-Übergriffs berichtet wird, kann als früherer häuslicher Vorfall gewertet werden, wenn

- er sich mindestens 24 Stunden vor dem Index-Überfall ereignete und
- er im Polizeibericht als eigenständiger Vorfall zu einem spezifischen Datum dokumentiert wird.

Beziehen Sie nicht ein:

- Den Index-Übergriff,
- Vorfälle, die nur Haustiere oder Eigentum involvieren
- oder Vorfälle, die nur Fremde, Bekannte, Freunde, Eltern, Geschwister, andere Familienmitglieder oder Polizeibeamte involvieren.

***Item 2: Früherer nicht-häuslicher Vorfall***

Ein nicht-häuslicher Vorfall ist definiert als ein Vorfall, bei dem der zu beurteilende Mann irgendeine andere Person als seine aktuelle oder frühere Lebenspartnerin oder deren Kinder tätlich angriff und der Übergriff in einem Polizeibericht oder im Vorstrafenregister verzeichnet ist.

Alle der folgenden Kriterien müssen vorliegen:

- Eine Gewalthandlung (wie oben unter 3. definiert), die von dem zu beurteilenden Mann verübt wurde
- und ein Vorfall, der sich zu einem separaten Anlass vor dem Index-Übergriff ereignete
- und Beteiligung der Polizei oder ein anschließender Bericht an die Polizei
- und ein Opfer, das irgendeine andere Person als die aktuelle oder frühere Lebenspartnerin des zu beurteilenden Mannes oder deren Kinder ist.

Falls eines dieser Kriterien fehlt, bewerten Sie das Item mit „0“, auch wenn jedes andere Kriterium vorliegt. Alle Kriterien müssen erfüllt sein.

Beziehen Sie nicht ein:

- Den Index-Übergriff,
- Vorfälle, die nur Haustiere oder Eigentum involvieren
- oder Vorfälle, die nur eine aktuelle oder frühere Lebenspartnerin des zu beurteilenden Mannes oder ihre Kinder involvieren.

***Item 3: Frühere Freiheitsstrafe von 30 Tagen oder mehr***

Eine Freiheitsstrafe beinhaltet nur die rechtskräftige Entscheidung die das Gericht für eine Straftat gefällt hat. Es ist nicht notwendig, dass der Mann die gesamte Strafe verbüßt hat, um das Item mit „1“ zu bewerten.

Alle der folgenden Kriterien müssen vorliegen:

- Eine Strafe, die vor dem Index-Übergriff gefällt wurde
- und eine Strafe, die zu einer Inhaftierung führte
- und eine Strafe von 30 Tagen oder mehr
- und ein tatsächlicher Eintritt in eine Jugendanstalt, ein Gefängnis, oder eine Bundes- oder Landesjustizvollzugsanstalt.

Falls eines dieser Kriterien fehlt, bewerten Sie das Item mit „0“, auch wenn jedes andere Kriterium vorliegt. Alle Kriterien müssen erfüllt sein.

Die Strafe muss nicht wegen eines häuslichen Vorfalls verbüßt worden sein; Strafen für einen nicht-häuslichen Vorfall oder jede andere strafrechtliche Verurteilung sollten einbezogen werden;

Strafen, die für irgendeinen Zeitraum durchgehend oder mit Unterbrechungen verbüßt wurden sollten genauso einbezogen werden wie Strafen, die nicht vollständig (aber zumindest zum Teil) in Haft verbüßt wurden, solange die Gesamtanzahl der Tage, die vom Gericht festgesetzt wurde, mindestens 30 betrug.

Beziehen Sie nicht ein:

- Die für den Index-Übergriff verhängte Strafe,
- die in Polizeigewahrsam verbrachte Zeit,
- die in Untersuchungshaft verbrachte Zeit außer das Urteil bezieht diese Zeit als Teil der Haftstrafe mit ein (was in manchen Rechtssystemen eine gesetzliche Vorgabe sein kann).

***Item 4: Versagen bei früherer bedingter Entlassung***

Bedingte Entlassung beinhaltet Entlassung als Verwaltungsvorgang, Entlassung gegen Kautions, Aufschiebung des Vollzugs, bedingte Entlassung mit Bewährungshilfe, Bewährungsstrafe, Entlassung aus der Untersuchungshaft vor der Hauptverhandlung, Strafaussetzung oder irgendeinen anderen Anlass, bei dem der Mann frei in der Gesellschaft unter Supervision oder anderen vom Strafgericht angeordneten Auflagen war; es beinhaltet außerdem Kontaktverbote, Schutzmaßnahmen oder einstweilige Verfügungen, die von einem Straf- oder Zivilgericht verhängt wurden.

Einige Beispiele für das Versagen bei bedingter Entlassung sind: Begehen einer neuen Straftat, Nichterscheinen vor Gericht, Nichtwahrnehmen eines Termins bei einem Bewährungshelfer, Trinken von Alkohol trotz Verbot, Besitz von Waffen trotz Verbot, Erscheinen am Haus oder Arbeitsplatz einer Person trotz Verbot, Kontaktieren einer Person trotz Verbot.

Mindestens ein Kriterium der folgenden muss vorliegen:

- Irgendein aktenkundiger Verstoß gegen eine bedingte Entlassung, unabhängig davon, ob er dafür verhaftet oder angeklagt wurde oder nicht
- und/oder ein Verstoß, der sich zum Zeitpunkt des Index-Übergriffs oder zu einem separaten Anlass vor dem Index-Übergriff ereignet hat
- und/oder eine Anklage für eine Straftat, die während einer vom Strafgericht angeordneten bedingten Entlassung begangen wurde.

Eines dieser Kriterien muss vorliegen, um das Item mit „1“ zu werten. Es müssen nicht beide vorliegen. Falls keines dieser Kriterien erfüllt ist, bewerten Sie das Item mit „0“.

Beziehen Sie nicht ein:

- Bedingte Entlassungen, die er befolgte
- oder Versagen bei bedingter Entlassung, die sich bei einem separaten Anlass nach dem Index-Übergriff ereignete.

***Item 5: Drohung beim Index-Übergriff zu verletzen oder zu töten***

Um dieses Item zu werten, muss sich eine Drohung eindeutig auf die physische Verletzung einer Person beziehen. Normalerweise liegt ein Anhaltspunkt in Form einer spezifischen verbalen Drohung, eine Person physisch zu schädigen, vor.

Im Vorstrafenregister wäre ein Hinweis auf eine Drohung, die die Kriterien dieses Items erfüllt, in Form einer Anklage für Drohen, Androhung von körperlicher Verletzung oder Tod, Aussprechen einer Drohung oder Störung des öffentlichen Friedens durch Androhung von Straftaten zu finden.

Hinweise auf eine physische Geste, die im Allgemeinen als Drohung, eine Person physisch zu schädigen, wahrgenommen wird, kann ebenso zur Wertung dieses Items verwendet werden.

**Alle der folgenden Kriterien müssen vorliegen:**

- Eine Drohung, irgendeine Person physisch zu schädigen
- und eine Drohung während des Index-Übergriffs unabhängig davon, ob sie ausgeführt wurde oder nicht.

Beide dieser Kriterien müssen erfüllt sein.

**Mindestens ein Kriterium der folgenden muss vorliegen:**

- Eine Drohung, die gegenüber irgendeiner Person, einschließlich Polizeibeamten, ausgesprochen wurde
- und/oder eine Anklage für Drohen, ausgesprochene Drohungen oder Störung des öffentlichen Friedens durch Androhung von Straftaten.

Eines dieser Kriterien muss vorliegen, um das Item mit „1“ zu werten. Es müssen nicht beide vorliegen. Falls keines dieser Kriterien erfüllt ist, bewerten Sie das Item mit „0“.

Beziehen Sie nicht ein:

- Androhungen von emotionaler Schädigung, finanzieller Schädigung, rechtlichen Schritten oder eines Sorgerechtsstreits,
- Drohungen gegen Haustiere oder Eigentum,
- Drohungen, sich selbst zu verletzen oder zu töten
- oder Drohungen, die zu einem separaten Anlass als dem Index-Übergriff ausgesprochen wurden.

Verdeckte Handlungen oder Unterlassungshandlungen, die im Allgemeinen nicht als Drohung wahrgenommen werden, können nicht verwendet werden, um das Item zu werten, auch wenn das Opfer den Eindruck hat, dass eine Schädigung droht. Dasselbe gilt für weiteres Verhalten, hinter dem die Absicht steht eine Person zu verängstigen, ohne dass eine offenkundige Drohung vorliegt.

Ein Übergriff mit einer Waffe ist kein ausreichender Hinweis für das Vorliegen des Items.

Stalking ist kein ausreichender Hinweis für das Vorliegen des Items.

***Item 6: Einsperren der Partnerin beim Index-Übergriff***

Einsperren ist definiert als jede Handlung, die darauf abzielt, physisch zu verhindern, dass das Opfer den Ort des Vorfalls verlässt. In den meisten Fällen würde Einsperren auf Grundlage von Hinweisen gewertet werden, wonach der Täter das Opfer in einem Raum gefangen hielt und den Ausgang versperrte. Im Vorstrafenregister wäre ein Hinweis zur Wertung des Items in Form einer Anklage wegen Freiheitsberaubung oder Geiselnahme im Anlassdelikt zu finden, vorausgesetzt, das Opfer ist die Lebenspartnerin.

In manchen Fällen können auch andere Handlungen des Täters, die dem Opfer die Möglichkeit nehmen, den Ort des Vorfalls zu verlassen, zur Wertung des Items verwendet werden. Diese weniger üblichen Beispiele des Einsperrens beinhalten Festhalten oder Festbinden des Opfers, als sie versuchte zu fliehen; absichtliches Stehen zwischen dem

Opfer und dem Fluchtweg, als sie versuchte zu fliehen; Ausziehen oder Zerstören der Kleider des Opfers, als sie versuchte nach draußen zu fliehen und gewaltsame Wegnahme von Fahrzeugschlüsseln oder Beschädigen eines Fahrzeugs, während sie versucht darin zu fliehen.

Alle der folgenden Kriterien müssen vorliegen:

- Einsperren der Partnerin, die das Opfer des Index-Übergriffs ist
- und eine Handlung, mit der der zu beurteilende Mann physisch das Opfer daran hinderte oder zu hindern versuchte, den Ort zu verlassen
- und Einsperren während des Index-Übergriffs unabhängig davon, ob das Opfer den Ort letztlich verließ.

Falls eines dieser Kriterien fehlt, bewerten Sie das Item mit „0“, auch wenn jedes andere Kriterium vorliegt. Alle Kriterien müssen erfüllt sein.

Beziehen Sie nicht ein:

- Drohungen des Täters, das Opfer zu verletzen, falls sie geht,
- Herausreißen des Telefons oder Durchschneiden der Telefonleitungen
- oder Einsperren bei einem vom Index-Übergriff separaten Anlass.

Hält der Täter das Opfer während eines Übergriffs fest, reicht dies für die Wertung des Items nicht aus.

### ***Item 7: Besorgnis des Opfers***

Dieses Item erfasst die Vorhersage des Opfers bezüglich künftiger tätlicher Angriffe gegen sie selbst oder ihre Kinder. Jede Aussage des Opfers, die auf Bedenken, Angst, Besorgnis oder Gewissheit bezüglich eines möglichen zukünftigen Vorfalls häuslicher Gewalt hinweist, wird als Beispiel für die Besorgnis des Opfers betrachtet.

Alle der folgenden Kriterien müssen vorliegen:

- Eine Aussage, der Partnerin, die das Opfer des Index-Übergriffs ist

- und eine Aussage in den ersten Berichten des Opfers gegenüber der Polizei oder gegenüber Opferberatungsstellen, falls die Information nicht für die Polizei verfügbar war, auch wenn sie anschließend keine Besorgnis mehr äußerte
- und eine Aussage, die auf Bedenken, Angst, Besorgnis oder die Gewissheit bezüglich eines möglichen zukünftigen Vorfalls häuslicher Gewalt hinweist.

Falls eines dieser Kriterien fehlt, bewerten Sie das Item mit „0“, auch wenn jedes andere Kriterium vorliegt. Alle Kriterien müssen erfüllt sein.

Mindestens ein Kriterium der folgenden muss vorliegen:

- Eine Aussage bezüglich möglicher zukünftiger Übergriffe gegen die Partnerin, die das Opfer des Index-Übergriffs ist
- und/oder eine Aussage bezüglich möglicher zukünftiger Übergriffe gegen ihre Kinder.

Eines dieser Kriterien muss vorliegen, um das Item mit „1“ zu werten. Es müssen nicht beide vorliegen. Falls keines dieser Kriterien erfüllt ist, bewerten Sie das Item mit „0“.

Beziehen Sie nicht ein:

- Die Angst des Opfers um ihre Sicherheit während des Index-Übergriffs,
- Aussagen des Opfers, die bei einem separaten Anlass vor dem Index-Übergriff gemacht wurden
- oder Rückschlüsse auf das Vorliegen einer Besorgnis, aus vom Opfer ergriffenen Schutzmaßnahmen.

***Item 8: Mehr als ein Kind***

Addieren Sie die Anzahl der Kinder, die der Täter hat plus jedes weitere Kind, das das Opfer hat. Die Gesamtanzahl von Kindern muss grösser als eins sein, um dieses Item mit einer „1“ zu werten.

Beziehen Sie ein:

- leibliche oder adoptierte Kinder des zu beurteilenden Mannes,
- leibliche oder adoptierte Kinder der Partnerin, die das Opfer des Index-Übergriffs ist,
- minderjährige oder erwachsene Kinder und
- Kinder, die beim Opfer leben oder anderswo leben.

Beziehen Sie nicht ein:

- Kinder, die zum Zeitpunkt des Index-Übergriffs ungeboren waren,
- Kinder, die vor dem Index-Übergriff verstorben sind,
- Kinder eines früheren Partners, die weder leibliche noch adoptierte Kinder des zu begutachtenden Mannes oder Opfers sind.

***Item 9: Leibliches Kind des Opfers von einem früheren Partner***

Um dieses Item mit „1“ zu werten, muss das Opfer des Index-Übergriffs ein leibliches Kind haben, dessen Vater nicht der Täter ist. Nur ein Kind ist erforderlich, um dieses Item mit „1“ zu werten.

Beziehen Sie ein:

- Leibliche Kinder der Partnerin, die das Opfers des Index-Übergriffs ist,
- Minderjährige oder erwachsene Kinder
- und Kinder, die beim Opfer oder anderswo leben.

Beziehen Sie nicht ein:

- Kinder, die vom Opfer adoptiert wurden,
- Kinder, die nicht die leiblichen Kinder des Opfers sind
- oder Kinder, die vor dem Index-Übergriff verstorben sind.

***Item 10: Gewalt gegen andere***

Alle der folgenden Kriterien müssen vorliegen:

- Eine Gewalthandlung (wie oben unter 3. definiert), die von dem zu beurteilenden Mann verübt wurde
- und ein Vorfall, der sich zu einem separaten Anlass vor dem Index-Übergriff ereignete
- und ein Opfer, das irgendeine andere Person als die aktuelle oder frühere Lebenspartnerin des zu beurteilenden Mannes oder deren Kinder ist.

Falls eines dieser Kriterien fehlt, bewerten Sie das Item mit „0“, auch wenn jedes andere Kriterium vorliegt. Alle Kriterien müssen erfüllt sein.

Eine Beteiligung der Polizei ist nicht erforderlich und die Information, die zum Wert des Items verwendet wird, kann aus anderen Quellen als justiziellen Dokumenten stammen. Die zur Wertung des Items erforderlichen Kriterien sind eine Teilmenge der Kriterien des Items „Früherer nicht-häuslicher Vorfall“. Wird „Früherer nicht-häuslicher Vorfall“ mit „1“ gewertet, dann wird „Gewalt gegen andere“ automatisch mit „1“ bewertet.

Beziehen Sie ein:

Eine Gewalthandlung gegenüber einer Partnerin, die nie mit dem zu beurteilenden Mann zusammengelebt hat.

Unspezifische Informationen über eine Gewalthandlung, die einen spezifischen Vorfall nicht konkret benennen, reichen für die Wertung des Items nicht aus.

***Item 11: Substanzmissbrauch***

Um dieses Item mit „1“ zu werten, muss mehr als ein Element von Substanzmissbrauch vorliegen. Alkoholkonsum beim Index-Übergriff ist für sich genommen nicht ausreichend. Die Elemente sind in absteigender Reihenfolge von „am häufigsten“ zu „am seltensten“ aufgelistet – entsprechend der Forschung zur Entwicklung des ODARA.

Mindestens zwei der folgenden Kriterien müssen vorliegen:

- Er konsumierte Alkohol unmittelbar vor oder während des Index-Übergriffs.

- Er konsumierte Drogen unmittelbar vor oder während des Index-Übergriffs.
- Er missbrauchte Drogen und/oder Alkohol in den Tagen oder Wochen vor dem Index-Übergriff (z.B. Alkoholintoxikation, häufiger Alkoholkonsum, Konsum von illegalen Drogen, Missbrauch von Medikamenten).
- Er steigerte seinen Konsum/Missbrauch von Drogen und/oder Alkohol merklich in den Tagen oder Wochen vor dem Index-Übergriff (ohne zum normalen Konsumverhalten vor dem Index-Übergriff zurückzukehren).
- In der Zeit vor dem Index-Übergriff ist er wütender und gewalttätiger gewesen, wenn er Drogen und/oder Alkohol konsumiert hatte.
- Er konsumierte Alkohol vor oder während einer Straftat, die zeitlich vor dem Index-Übergriff stattfand.
- Sein Alkoholkonsum vor dem Index-Übergriff, aber seit dem Alter von 18 führte zu Problemen oder Beeinträchtigungen in seinem Leben.
- Sein Drogenkonsum vor dem Index-Übergriff aber seit dem Alter von 18 führte zu Problemen oder Beeinträchtigungen in seinem Leben.

Beliebige zwei dieser Kriterien müssen vorliegen, um das Item mit „1“ werten zu können.

Nicht alle Kriterien müssen vorliegen. Wenn nur eines dieser Kriterien vorliegt, dann wird das Item mit „0“ gewertet.

***Item 12: Übergriff auf das Opfer während der Schwangerschaft***

Informationen zum Werten des Items können aus anderen Quellen als justiziellen Dokumenten stammen und der Vorfall muss der Polizei nicht bekannt sein.

Alle der folgenden Kriterien müssen vorliegen:

- Eine Gewalthandlung (wie oben definiert), die von dem zu beurteilenden Mann verübt wurde
- und ein Vorfall gegen die Partnerin, die das Opfer des Index-Übergriffs ist

- und die Schwangerschaft des Opfers zum Zeitpunkt des Übergriffs.

Falls eines dieser Kriterien fehlt, bewerten Sie das Item mit „0“, auch wenn jedes andere Kriterium vorliegt. Alle Kriterien müssen erfüllt sein. Der Index-Übergriff kann für die Wertung dieses Items verwendet werden.

Beziehen Sie nicht ein:

- Vorfälle, bei denen nur Haustiere oder Eigentum involviert sind,
- Vorfälle, die sich ereigneten, als das Opfer nicht schwanger war.

***Item 13: Barrieren der Opferunterstützung***

Dieses Item erfasst die Lebensverhältnisse des Opfers zum Zeitpunkt des Index-Übergriffs und sollte auf Grundlage von Informationen gewertet werden, die so zeitnah wie möglich zum Anlassdelikt dokumentiert wurden. Die Kriterien sind in absteigender Reihenfolge von „am häufigsten“ zu „am seltensten“ aufgelistet – entsprechend der Forschung zur Entwicklung des ODARA.

Mindestens ein Kriterium der folgenden muss vorliegen:

- Das Opfer des Index-Übergriffs hat ein oder mehrere Kinder im Alter von 18 Jahren oder jünger, die bei ihr wohnen und für die sie sorgt.
- und/oder sie hat kein Telefon (d.h. kein Mobiltelefon und kein Festnetztelefon zu Hause)
- und/oder sie verfügt über keine Transportmittel (d.h. kein Zugang zu einem Fahrzeug und keine öffentlichen Verkehrsmittel in der Umgebung ihres Wohnortes und kein Geld für ein Taxi)
- und/oder sie ist geografisch isoliert (d.h., sie lebt in einer ländlichen Gegend mit niemandem, der in der Nähe wohnt)
- und/oder sie konsumierte Alkohol oder Drogen kurz vor oder während des Index-Übergriffs oder hat in der Vergangenheit Alkohol- oder Drogen missbraucht (z.B.

Alkoholintoxikation, häufiger Alkoholkonsum, Konsum von illegalen Drogen, Missbrauch von Medikamenten).

Irgendeines dieser Kriterien muss vorliegen, um das Item mit „1“ zu werten. Nicht alle Kriterien müssen vorliegen. Wenn keines dieser Kriterien erfüllt ist, dann wird das Item mit „0“ gewertet.

## **2.4. Assessing the discrimination and calibration of the Ontario Domestic Assault Risk**

### **Assessment in Switzerland**

#### **2.4.1. Abstract**

Intimate partner violence (IPV) is a major public health issue. Worldwide, almost one in three women is affected. Police involvement in IPV cases has substantially increased and in the course of these changes, several front-line instruments have been developed to structure police risk assessment. One of those is the Ontario Domestic Assault Risk Assessment (ODARA). To investigate its validity in a Swiss police setting a total cohort of male IPV offenders was retrospectively assessed for a fixed time at risk of five years. ODARA scores were significantly correlated with police-registered IPV recidivism, but discrimination and calibration proved deficient compared to previous research. Several reasons for those deviations, such as level of intervention and the dynamic nature of IPV, are discussed.

#### **2.4.2. Introduction**

Intimate partner violence (IPV) is a major public health issue. Representative surveys reveal that worldwide, almost one out of three women (30%) is victimized at least once during their lifetime by their intimate partner (World Health Organisation, 2013). In the U.S., in the European Union (EU) and Switzerland, more than every third (36%, Black et al., 2011) and every fifth (22%, European union agency for fundamental rights, 2014; 21%, Gillioz, De Puy, & Ducret, 1997) woman, respectively, is abused by a violent (ex-)intimate partner. Several representative surveys have claimed that men are equally or even slightly more frequently affected by intimate partner violence than women (J. Archer, 2002; Cho, 2012; Straus, 2009). Nonetheless these figures are not reflected in correctional samples, where male offenders are clearly overrepresented; the large majority of all police-registered intimate partner offenders are male (e.g. U.S.: 87% Melton & Sillito, 2012; Switzerland: 81% Swiss

Federal Statistical Office, 2014). In addition, the consequences of IPV are more severe for female victims as they are more often injured (J. Archer, 2000; Straus, 2009) and more frequently need post-assault treatment (Greenfeld et al., 1998). Additionally, female victims significantly more often manifest severe symptoms of psychiatric disorders, including symptoms of depression, post-traumatic stress disorder, and anxiety (Swan et al., 2008; Tjaden & Thoennes, 2000).

### **Risk Management**

In most cases of IPV, the assault is not a one-time phenomenon. In England and Wales for example, three quarters of victims have been abused repeatedly, and almost half were victimized twice within one year (Office for National Statistics, 2013). Regarding sexual violence, the latest European survey revealed that more than half of those women affected have experienced IPV several times within the same relationship (European union agency for fundamental rights, 2014).

There is evidence that the chain of violence can be interrupted by punitive interventions. Data show that involving the police has “a very strong deterrent effect” on re-victimization rates (Felson, Ackerman, & Gallagher, 2005, p. 563). Furthermore, recent studies indicate that police interventions – often in the form of protection orders – are associated with a reduced risk of recidivism (Belfrage & Strand, 2012; Belfrage et al., 2012; Logan & Walker, 2009; Storey, Kropp, Hart, Belfrage, & Strand, 2014). However, the appropriateness and effectiveness of an intervention strongly depends on valid risk assessment strategies (Belfrage et al., 2012; Bonta & Andrews, 2007). In accordance with the Risk-Needs-Responsivity model (Bonta & Andrews, 2007), Belfrage et al. (2012) showed that effective risk management is required to match the intensity of risk management strategies with the risk level. However, heuristics as well as stereotypical beliefs, such as ethnicity of the offender, marital status of the victim, and victim’s responsibility, strongly influence police

involvement in cases of domestic violence and can undermine valid risk assessment when decision-making is not reasonably guided (ref. Holland-Davis & Davis, 2014; van den Heuvel, Alison, & Power, 2014; Weller, Hope, & Sheridan, 2012). Additionally, compared to other fields of risk assessment, front-line IPV risk assessment involves special needs and circumstances. Decisions about further case procedures are urgent, and the basis of information is small, often missing information on the offender's childhood or crucial psychiatric/psychological attributes (Hilton et al., 2004).

### **IPV Risk Assessment Instruments**

There is a general scientific consensus about the inferiority of unstructured-clinical decision-making strategies in comparison to other assessment approaches (Grove et al., 2000; Meehl, 1954). Accordingly, several IPV risk assessment instruments have been developed on the basis of empirically identified risk factors over the last three decades. Those risk assessment instruments differ with regard to several aspects. Firstly, the definition of the outcome criterion is crucial. Whereas most instruments address general IPV, such as the *Ontario Domestic Assault Risk Assessment* (ODARA; Hilton et al., 2004), the *Spousal Assault Risk Assessment* (SARA; Kropp & Hart, 2000) and the *Domestic Violence Supplementary Report* (DVSR; Ontario Ministry of the Solicitor General, 2000), other tools are specifically designed to estimate the risk of intimate partner homicide (IPH), including the *Danger Assessment* (DA; Campbell et al., 2009) and the *Domestic Violence Method* (DV-MOSAIC; De Becker, 2000). Secondly, IPV instruments differ regarding the basis of information that is used for the assessment. That is, some instruments depend on victims' interviews (e.g., DA), whereas others strongly rely on police and criminal files, such as the ODARA and B-SAFER (Kropp, Hart, & Belfrage, 2005). Thirdly, some instruments require assessors to have specific professional qualifications, such as psychiatric/psychological skills (e.g., SARA, DVRAG). Others were specifically developed for use in front-line settings, and their application does

not rely on such skills (e.g., DVSR, ODARA, B-SAFER). Finally, IPV instruments are characterized by different levels of standardization as they can be categorized into structured-clinical (i.e., structured professional judgment [SPJ] guidelines like the SARA), mechanical, such as, *Kingston Screening Instrument for Domestic Violence* (K-SID; Gelles & Tolman, 1998), or actuarial (e.g., ODARA, DVRAG) instruments. In contrast to unstructured-clinical assessments, all of the above examples are characterized by some degree of standardized and pre-defined judgment guidelines (Helmus & Bourgon, 2011). SPJ guidelines are characterized by an established evidence-based set of putative risk factors that guide the evaluator through the risk assessment process, but the combination and weighting of factors and the final risk estimation are left to the clinical expertise of the evaluator (Hart & Logan, 2011). Mechanical instruments combine ratings on risk factors into a total risk score by applying a standardized algorithm. Finally, actuarial instruments are empirically developed on a sample-based level to provide risk rates for different intervals of total risk scores (categories), thus enabling comparisons between individuals and a representative group of offenders with similar characteristics (Helmus & Bourgon, 2011).

### **Validity of IPV Risk Assessment Instruments**

There is a vivid debate about what defines an appropriate risk assessment approach. Whereas mechanical and actuarial instruments are less biased by subjective and potentially inconsistent risk estimations, SPJ guidelines allow individually based and risk management-focused assessments (Hart & Logan, 2011; Quinsey et al., 2006). In the field of general violent risk assessment, the majority of publications point toward a superiority of actuarial instruments concerning discrimination between recidivists and non-recidivists (ref. Falzer, 2013; Guy, 2008). However, the actual differences are marginal, and the robustness of this trend is impeded by the lack of comparability and quality often found in primary studies (Falzer, 2013; Singh et al., 2011). More specifically, reviews of IPV research have shown

that there is still a general lack of primary validation studies for most of the currently available instruments (Guo & Harstall, 2008; Messing & Thaller, 2013), and independent research remains scarce (Bowen, 2011; Messing & Thaller, 2013; Northcott, 2012).

However, the preliminary findings of those reviews investigating discrimination on the basis of receiver operating curve statistics (ROC) cautiously suggest that on average, actuarial IPV instruments tend to outperform IPV instruments of other approaches. Recent publications report higher mean discrimination coefficients for actuarial tools (area under the curve [ $AUC$ ] = .62-.70) compared to structured-clinical ones ( $AUC$  = .54-.63; Bowen, 2011; Kilvinger et al., 2012; Messing & Thaller, 2013).

In their meta-analysis, Messing and Thaller (2013) included all IPV risk assessment instruments for which more than one prospective validation had been published. This applied for only five instruments – the ODARA, SARA, DA, DVSI, and K-SID. The ODARA was found to be the most valid IPV risk assessment instrument, with a mean moderate effect size of  $d = .61$  (corresponding to an average  $AUC$  of .67) that was significantly larger than those of the other four IPV risk assessment instruments. In the development sample, the ODARA reached an  $AUC$  of .77 (Hilton et al., 2004), and across all following peer-reviewed ODARA validation studies that had been published up to June 2014, the ODARA showed a moderate ( $AUC = .64$ ; Hilton, Harris, Popham, et al., 2010) to good ( $AUC = .74$ ; Hilton & Harris, 2009) ability to discriminate between recidivists and non-recidivists (see Table 9).

Although previous ODARA validation studies are promising, independent and European peer-reviewed replication studies are still missing, except one Austrian evaluation in a sample of 66 high-risk sexual domestic violence offenders released from the prison system (Rettenberger & Eher, 2013, see Table 9). In a mean follow-up of 4.6 years, the ODARA showed good reliability and acceptable discrimination ( $AUC = .71$ ). Because this investigation applied a substantially different methodological approach than the

developmental study (i.e., it included only sexually motivated and convicted IPV offenders and defined recidivism as new IPV convictions rather than police registered IPV), it cannot be considered as a valid replication study (Rossegger, Gerth, Seewald, et al., 2013). However, all studies with more valid replication were conducted in Canadian samples with the primary involvement of the author of the ODARA (see Table 9). Thus to date, independent and valid ODARA replication studies are lacking.

### **The Present Study**

In 2007, the Swiss Canton of Zurich issued the “Law on Protection against Violence (LPV),” which was established to protect people who “*suffer from violence or which are threatened with violence by a person with which the victim has a partnership or is married to.*” (Kantonsrat, 2006, p. 1). In order to allow for efficient and immediate reactions to cases of domestic violence, the LPV assigns police officers with the legal competence to issue protection orders without a court decision within the first 48 hours after a domestic violence incident had occurred. Protection orders include no-contact, eviction and rayon (i.e., the designation of off-limit areas) orders, which are often also issued alongside with each other. In high-risk situations, a domestic violence offender can be taken into a 24-hour custody by the means of the police. The protection orders are issued and implemented by the police, irrespective of the victim’s consent. In addition, the state prosecutor might follow up on the domestic violence offense as a criminal offense (Kantonsrat, 2006).

In the 2.5 years after the LPV came into force, more than 2000 cases of domestic violence were registered, most of them concerning male-to-female IPV (Endrass, Rossegger, & Urbaniok, 2012). Until that point, no standardized assessment tools were implemented. In 2011, when a man – who had previously been registered by the police for several assaults and death threats toward his wife and children – shot his wife and her social worker, authorities reacted with the establishment of an interdisciplinary committee to bring together expert

opinions and improve communication structures. Furthermore, a special force unit was founded within the police system, and the implementation of a front-line standardized risk assessment instrument was required to reliably screen IPV offenders. With reference to the literature review mentioned above (also see Table 9), the ODARA was considered.

Thus, it was the present study's aim to evaluate the ODARA's discrimination and calibration in a Swiss police setting. Particular attention was paid to the design to strongly match the demographic and design characteristics between the current and development studies.

DISCRIMINATION AND CALIBRATION OF THE ODARA

**Table 9:** Previous validation studies investigating the discrimination of the ODARA (June 2014)

Previous studies		Current study						
Year of publication	2004	2008	2009	2010	2013	2014	-	
Country	Canada	Canada	Canada	Canada	Austria	Canada	Switzerland	
Sample size	589	346	391	150	66	30	186	
Sex	Male	Male	Male	Male	Male	Female	Male	
Type of IPV index assault	Police registered	Police registered	Police registered	Incarceration	Sexually motivated	Police registered	Police registered	
Mean LoFU (years)	4.9	5.1	5.0	5.1 (time at risk)	4.6	9.0	5.0 (time at risk)	
IPV recidivism criteria	Police registered/convicted	Police registered/convicted	Police registered/convicted	Charges for IPV	IPV convictions	Charges for IPV	Police registered IPV	
Recidivism rate	30%	41%	27%	27%	21%	23%	32%	
<i>AUC (CI)</i>	.77 (.73-.81)	.65 (.59-.71)	.67 (.61-.73) <sup>b</sup>	.74 (.68-.80) <sup>c</sup>	.64 (.54-.73)	.71 (.58-.84)	.72 (.50-.94)	.63 (.55-.72)

*Note.* *AUC* = area under the curve; *CI* = confidence interval; *IPV* = intimate partner violence; *NR* = not reported; *LoFU* = length of follow-up

<sup>a</sup>ODARA development study.

<sup>b</sup>Discrimination performance taking the whole sample into account by assigning ambiguous offenders into the non-recidivist group.

<sup>c</sup>Discrimination performance while comparing offenders who recidivated unambiguously with a new IPV incident with those who did not violently recidivate.

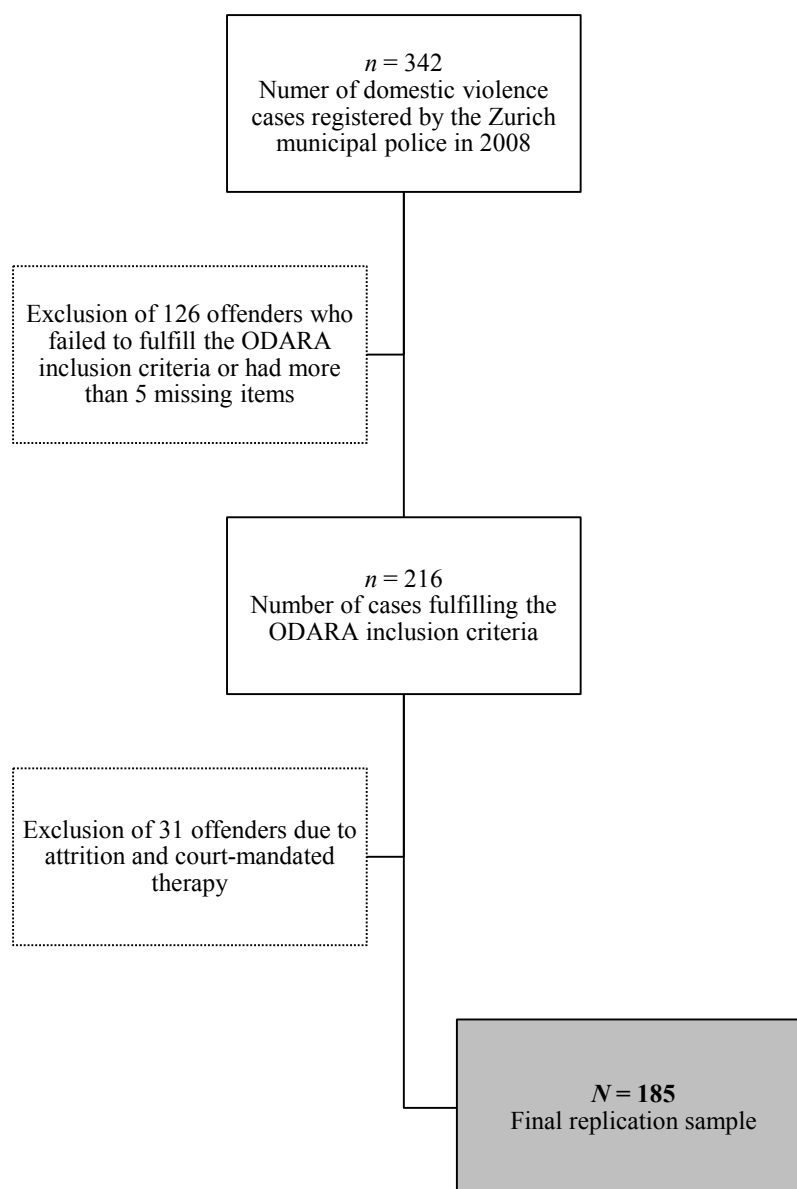
### 2.4.3. Method

#### Participants

The sample was drawn from a total cohort of domestic violence offenders registered by the municipal police of Zurich in 2008 ( $n = 342$ ). Cases that did not match the ODARA's definition of the index assault (i.e., an adult male offender who had physically assaulted his female (ex-)intimate partner or made a credible threat of death with a weapon in hand against her) and those with more than five missing items were excluded.

According to the ODARA norms published in 2010 (Hilton, Harris, & Rice) the remaining 216 cases were followed for a fixed time of five years. Controlling for attrition (incarceration, death or deportation [ $n = 30$ ]) and excluding those of the remaining sample who received court-mandated therapy after the index assault ( $n = 1$ ), resulted in a final sample size of  $N = 185$  (see Figure 11).

According to the Swiss laws, the present study required approval by the municipal data protection authority but not the cantonal ethics committee. As of March 2013, the study design was approved in the current form.



**Figure 11.** Flow chart depicting the process of fulfilling the ODARA's including criteria and applying a fixed time at risk of five years.

### Procedure

**Data collection.** The application of the ODARA was based on two different sources: cantonal police files and criminal records. Police files provided information on the index assault, the offender's criminal history and socio-economic characteristics, and victim and witness reports if available. As the police files only gather cantonal information, criminal

records added information on nationwide entries about previous charges, convictions, and corresponding judicial information.

Recidivism was assessed based on police files of the Canton of Zurich and was defined as IPV (including threats and physical [sexual] assaults) registered by the police that occurred at least 24 hours after the index assault and within a 5-year time of risk. Recidivism was blindly scored for the ODARA assessment.

Data collection was conducted by four psychologists; three had at least a master's degree and one held a bachelor's degree. Raters were thoroughly introduced into the reading of police files by police officers of the municipal police of Zurich before the data collection process started. Data collection was conducted in the police station, and the raters were continuously assisted by police officers whenever questions arose during the process of data collection. All raters were trained in applying the ODARA and had successfully assessed five training cases. Questions arising during the assessment process were either discussed within the team or forwarded to and answered by the ODARA's author, N. Zoe Hilton. The items were then rated accordingly.

**Measure: The Ontario Domestic Assault Risk Assessment (ODARA).** The original ODARA manual printed in 2005 (Mental Health Centre Penetanguishene) was slightly adapted in 2010 (Hilton, Harris, & Rice). The construction of the ODARA was based on a literature survey on predictors of IPV recidivism and a multivariable regression analysis in a representative sample of 589 Canadian IPV offenders in the province of Ontario. Only variables that were assumed to be easily collected by police officers during front-line involvement in IPV cases were included. The 13 dichotomous items of the ODARA (see Table 10) are static in nature and mainly include information on previous domestic and non-domestic violence, characteristics of the index assault, substance abuse problems, and the

victim's attributes and risk estimation (Hilton, Harris, & Rice, 2010). The instrument's inter-rater reliability is strong (interclass correlation [ $ICC$ ] = .95).

DISCRIMINATION AND CALIBRATION OF THE ODARA

**Table 10.** Correlation between ODARA items and police-registered IPV recidivism in the Zurich sample and internal consistency analysis of the ODARA scale (n = 185)

ODARA item	%	$\phi$	Item poling	Item-rest correlation	Average intertem correlation	alpha
1 Has a prior domestic assault (against a partner or child)	67.4	<b>.18*</b>	+	.05	.07	.49
2 Has a prior nondomestic assault (against anyone other than a partner or child)	26.0	.10	+	.46	.04	.35
3 Has a prior sentence to a term of 30 days or more	6.5	.06	+	.27	.06	.41
4 Has a prior failure on conditional release including bail, parole, probation, no-contact order	14.6	<b>.21**</b>	+	.28	.05	.41
5 Threatened to harm or kill anyone during index offense	58.5	.03	-	.06	.07	.48
6 Unlawful confinement of victim during index offense	24.0	.03	-	.03	.07	.48
7 Victim fears repetition of violence	82.5	.06	+	-.06	.08	.50
8 Victim and/or offender have more than one child altogether	49.7	.04	+	.13	.07	.46
9 Offender is in a stepfather role in this relationship	25.1	<b>.17*</b>	+	.16	.06	.44
10 Offender is violent outside the home (to people other than a partner or child)	30.3	.08	+	.46	.04	.35
11 Offender has more than one indicator of substance abuse problem	66.4	.17	+	.22	.06	.43
12 Offender has ever assaulted victim when she was pregnant	16.1	.10	-	.01	.07	.49
13 Victim faces at least one barrier to support	69.6	.02	+	.07	.07	.47
Test Scale					.06	.46

\* $p \leq .05$ , \*\* $p < .01$

Scorings on each of the 13 items are added to a total sum score that can be transferred to one of seven risk bins. These were shown to be positively correlated with the occurrence, frequency, and severity of police-registered IPV recidivism in a representative Canadian IPV offender sample with a mean follow-up of five years. Recidivism rates are displayed for each of the seven risk bins. Altogether missing information is allowed for up to five items, prorating is necessary in the case of missing information. Prorating values can be withdrawn from a corresponding table in the ODARA's official manual (Hilton, Harris, & Rice, 2010).

### **Statistical Analysis**

Discrimination analyses were calculated by generating receiver-operating curves (ROCs) that plot the function of the true- and false-positive rates for all possible values of a decision criterion (i.e., ODARA sum score or ODARA risk bin). Its corresponding measure of effect size is displayed by the *AUC*. ROC analyses were calculated for 1- and 5-year times at risk.

Because the *AUC* coefficient solely describes the ability of the instrument to discriminate between recidivists and non-recidivists, it is only one part of a relevant performance measure, as it does not take into account any calibration measures to assess the distribution of recidivism rates across different risk bins or scores. Thus, calibration analyses were conducted using Sander's decomposition of the Brier score (Schmid & Griffith, 2005). The Brier score is known as an overall performance measure that includes both discrimination and calibration (Rufibach, 2010; Steyerberg, Vickers, Cook, Gerds, & Gonen, 2010; Vergouwe, Steyerberg, Eijkemans, & Habbema, 2005). By applying Sanders' decomposition of the modified Brier score, the calibration property of the overall performance measure is analyzed by using its first term (Rogers, 1992; Schmid & Griffith, 2005; Spiegelhalter, 1986).

For a bin-specific investigation of rate differences between the current study and previously published ODARA norms, we calculated Bayesian credible intervals by using the

Jeffreys' prior for the Beta distribution (Breslow, 1990; Carlin & Louis, 2009; Edwards, Lindman, & Savag, 1963; Kass & Wasserman, 1996).

Additional analyses were conducted to gain a more comprehensive understanding of the current findings. We examined the instrument's structure as we expected that either the ODARA would reflect a composition of rather independent correlates of IPV or that the ODARA would act as a scale showing a high internal consistency. Tetrachoric correlations were calculated between each ODARA item and IPV recidivism by applying the phi-coefficient, and item homogeneity was tested by calculating Cronbach's  $\alpha$ .

STATA/IC 13.1 for Windows (StataCorp, 2013) was used for all analyses, and two-tailed tests were calculated with a standard significance threshold of  $\alpha = .05$ .

#### **2.4.4. Results**

##### **Sample Characteristics**

The sample consisted of 185 male offenders with a mean age of 38.1 ( $SD = 11.1$ ); the youngest and oldest offenders at the time of the index assaults were 19 and 74 years old, respectively. Two-thirds ( $n = 118$ , 63.8%) of the sample were foreign nationals.

Index assaults as documented by the police included minor assaults ( $n = 109$ , 58.9%) and threatening ( $n = 97$ , 52.4%) in more than half of the cases, bodily injury in 44.9% ( $n = 83$ ), and coercion in 28.1% ( $n = 52$ ). Sexual coercion/rape was reported in nine cases (4.9%), and endangering someone's life was registered in two cases (1.1%). False imprisonment ( $n = 5$ , 2.7%) and extortion ( $n = 3$ , 1.6%) were rare assault types.

In almost half of the cases (43.5%), the offender was under the influence of alcohol at the time of or just before the index assault. In every tenth (9.4%) assault, a weapon was involved (defined as any object suited to cause injury and used purposely against the victim). Almost every fourth (23.6%) offender had stalked his victim prior to the index assault, and in more

than half (58.5%) of cases, threats to harm or kill anyone were uttered during the index assault.

The police issued for personal protection orders against almost all offenders (98.9%), including no-contact, rayon, and eviction orders in 97.8%, 70.0%, and 66.1% of cases, respectively. More than one-fourth (27.6%) of offenders were incarcerated following the index assault.

### **ODARA Scores**

The mean ODARA sum score was 5.1 ( $SD = 2.0$ , range = 0 to 10), which was significantly higher than the development study's mean sum score of 2.9 ( $t[772] = 12.4$ ,  $p = .000$ ; Hilton, Harris, & Rice, 2010). The mean ODARA risk bin was 5.5 ( $SD = 1.4$ , range = 1 to 7).

### **IPV Recidivism**

Within a fixed time at risk of five years, nearly one-third ( $n = 59$ , 31.9%) of the cohort recidivated with a police-registered threat or physical (including sexual) assault toward an (ex-) intimate partner. Mostly offenders recidivated with a minor assault ( $n = 36$ , 19.5%) or threatening ( $n = 34$ , 18.4%). Overall, 8.1% ( $n = 15$ ) were registered with a new offense of bodily injury and 7.0% ( $n = 13$ ) with coercion. Two (1.1%) offenders endangered their (ex-) intimate partner's life, and one (0.5%) offender was reported to the police because of rape.

### **Performance Measures**

In general, significant positive relationships were detected between the ODARA sum score and recidivism as well as between the ODARA risk bin and recidivism ( $r = 0.23$ ,  $p = .002$  and  $r = 0.20$ ,  $p = .005$ , respectively). That is, a higher ODARA score or risk bin was associated with a higher rate of recidivism, and recidivists and non-recidivists differed significantly with regard to their mean ODARA risk bin ( $t[183] = -2.826$ ,  $p = .003$ ).

Nonetheless the recidivism rate in the highest risk bin did not exceed 46%, suggesting that too many non-recidivists were categorized in the highest risk bins of the ODARA (see Table 11). In accordance with this finding, the discrimination performance of the ODARA was low, with an *AUC* of .63 (95% confidence interval [*CI*] = .55-.71, *p* = .004) for its risk bins. The overall Brier score reached *BS* = .25, which equates to a prediction performance as accurate as chance (Schmid & Griffith, 2005). Additionally, comparing the recidivism rates of the current study with those published in Canada, the first term of Sander’s decomposition of the Brier Score was 0.4, which corresponds to an average error of prediction of 20.6% per risk bin.

A bin-specific investigation of the calibration showed that the original recidivism rates of bins 6 (53%) and 7 (74%) were not covered, and risk bin 5 (39%) was just covered by the corresponding Bayesian credible intervals of the Zurich sample (0.23-0.46, 0.32-0.59 and 0.08-0.40, respectively; see Table 11 and Figure 12).

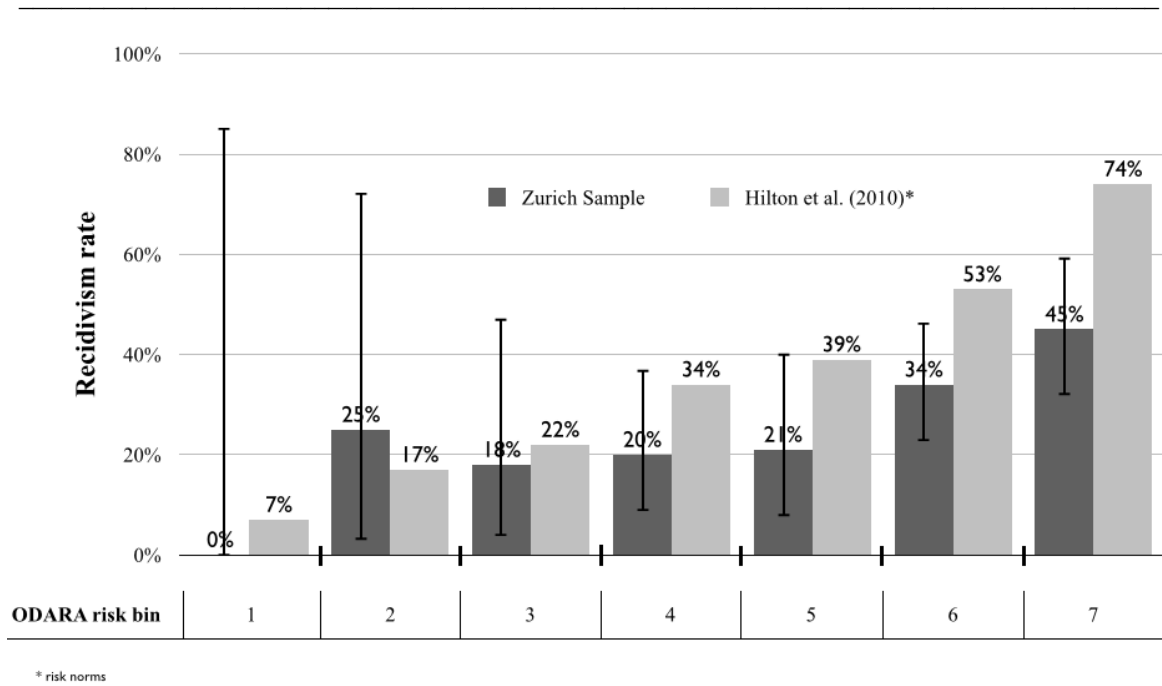
**Table 11:** Recidivism rates for the ODARA after a 5-year time at risk (n = 185) and comparisons with normative risk rates.

ODARA risk bin	Total risk score	Percentage of sample in each risk bin		Recidivism rate (95% Bayesian credible interval <sup>a</sup> )	
		Zurich sample	Hilton et al. (2010) <sup>b</sup>	Zurich sample <sup>c</sup>	Hilton et al. (2010) <sup>b</sup>
1	0	0.5%	9%	0.0% (0.00-0.85)	7%
2	1	2.2%	17%	25.0% (0.03-0.72)	17%
3	2	6.0%	21%	18.2% (0.04-0.47)	22%
4	3	16.2%	20%	20.0% (0.09-0.37)	34%
5	4	13.0%	13%	20.8% (0.08-0.40)	39%
6	5-6	33.5%	14%	33.9% (0.23-0.46)	53%
7	7-13	28.7%	6%	45.3% (0.32-0.59)	74%

<sup>a</sup>Bayesian credible intervals were calculated using the Jeffreys’ prior for the Beta distribution

<sup>b</sup>ODARA development study with an average follow-up of five years.

<sup>c</sup>Recidivism base rate = 31.9%; *AUC* ODARA risk bin = .63 (95% *CI* = .55-.71)



**Figure 12.** Observed IPV recidivism rates within a time at risk of five years surrounded by Bayesian credible intervals calculated by using the Jeffreys’ prior for the Beta distribution. The observed rates were compared to the normative data (Hilton, Harris, & Rice, 2010).

**Scale structure**

By conducting correlation analyses using the *phi* coefficient, we found that only 3 of the 13 ODARA items significantly correlated with the outcome criterion at a significance level of  $p < .05$ . These were item 1 (“Has a prior domestic assault against a partner or child,”  $\phi = .18, p = .015$ ), item 4 (“Has a prior failure on conditional release including bail, parole, probation, no-contact order,”  $\phi = .21, p = .004$ ) and item 9 (“Offender is in stepfather role in this relationship,”  $\phi = .16, p = .024$ ). See Table 10.

Furthermore, correlation coefficients close to zero ( $\phi < .05$ ) were detected for four items (i.e., item 13 “Victim faces at least one barrier to support,”  $\phi = .02$ ; item 5 “Threatened to harm or kill anyone during index offense,”  $\phi = .03$ ; item 6 “Unlawful confinement of victim during index offense,”  $\phi = .03$ ; and item 8 “Victim and/or offender have more than one child altogether,”  $\phi = .04$ ). An internal consistency analysis revealed a Cronbach’s  $\alpha$  of .46, item-rest correlations ranging from -.06 to .46, and an average inter-item correlation of .06. More over, three items (item 5, 6 and 12) entered the analysis conversely poled (see Table 10).

### 2.4.5. Discussion

The aim of the present work was to conduct the first European replication study to closely match the development study characteristics of the ODARA. With an *AUC* of .63, the ODARA showed a significant but poor performance in discriminating between IPV recidivists and non-recidivists. Accordingly, recidivism rates across the ODARA's risk bins were substantially different from those published by the authors of the instrument (20.6% in average per risk bin). Similar findings have recently been published for other actuarial risk assessment instruments in Switzerland, suggesting that the Canadian normative data are inappropriate for use in Switzerland, and the establishment of jurisdiction-specific norms seems advisable (Rossegger et al., 2014; Rossegger, Gerth, Singh, et al., 2013). In the current study, differences were most obvious in the two highest risk bins, where Bayesian credible intervals of the Zurich sample failed to cover the risk norms. As the differences do not reflect a general deviation of the samples' base rates, the reasons for such deviations may be due to several other sources.

Firstly, the samples may not be comparable with respect to their risk disposition and the kind or level of intervention strategies or other influencing contextual factors. The mean ODARA sum score in the present sample is significantly higher than that reported in the development study (i.e.,  $M = 5.1$  and  $M = 2.9$ , respectively), and almost two-thirds (62%) of the sample were assigned into the two highest risk bins. Following the ODARA model, this suggests that the Zurich sample is a high-risk sample. However, even the subjects in the highest risk bins had a moderate recidivism rate, which is for obvious reasons inconsistent with the notion of a high-risk population. The relatively moderate rate of recidivism is not limited to the current study. Hilton, Harris, Popham, et al. (2010) reported a similar outcome in a Canadian sample of incarcerated IPV offenders who were transferred to a domestic violence treatment program. With a mean ODARA sum score of 5.8, only 27% offenders

recidivated within a mean follow-up of five years, and a ROC analysis revealed an *AUC* of .64. While all offenders in this sample were incarcerated and were even enrolled in a treatment program (Hilton, Harris, Popham, et al., 2010), only half of the offenders included in the development study got arrested, not even one-third (31%) was held in custody, and a no-contact order was issued for at most every tenth individual (9%; personal communication Hilton, August 19, 2014). Other intervention strategies of the police included advising the victim to seek legal help, mediation, seizing weapons, or generally providing information (ref. Hilton, Harris, & Rice, 2007). Obviously the risk disposition of the incarcerated sample was estimated to be high as they resulted in incarcerations and treatment recommendations. Similarly, all of the offenders in the current Zurich study were taken or ordered to the police department for hearings. For almost all offenders ( $n = 183$ , 99%) protection orders of some kind were issued; 97% of offenders had to stay distant from his partner for at least two weeks, 76% were arrested after the index offense, and more than one-fourth (28%) were incarcerated for an average of 17.7 days ( $SD = 14.1$ , range = 1 to 62). The moderate rate of recidivism could be the consequence of a self-canceling effect. However, research on risk management mediating the association between risk assessment and actual recidivism remains scarce, and the significant effects of the few studies in the field (e.g. Belfrage et al., 2012) have yet to be replicated (Storey et al., 2014).

Secondly, reasons for a reduction in IPV recidivism may also be found in the highly dynamic nature of IPV. Study findings suggest that relationship-specific factors play a significant role in an offender's persistence or desistance from IPV (e.g. Walker, Bowen, & Brown, 2013). Accordingly, Whitaker, Le, and Niolon (2010) found that desistance from violence across relationships was better explained by factors relating to the dynamic nature of the relationship (e.g., relationship duration, educational differences between partners, and partner using violence) rather than individual perpetrator factors such as drug use or peer

violence. Also, a much higher IPV desistance rate is reported *across* relationships than *within the same* relationship (ref. Whitaker et al., 2010), and several studies have shown that changes in IPV perpetration are associated with relationship termination (e.g. Shortt et al., 2012; Timmons Fritz & Slep, 2009). Consequently, dyadic interactional factors are also hypothesized to stimulate persistent IPV. In summary, the performance of an IPV risk assessment may most likely be impeded by taking only static risk factors into account (Dutton, 2012), especially when acute and short-term risk is assessed and reasonable risk-tailored management strategies need to be implemented (de Ruiter & Nicholls, 2011).

Thirdly, the disproportion between the assessed level of risk and the prevalence of recidivism may be a result of the victim's reporting behavior. Although awareness of IPV has increased and important policy changes have been achieved, it is estimated that only one third (33%) of victimizations is reported in the EU (European union agency for fundamental rights, 2014). Assaults reported to the police are often the most severe ones or the far end of a series of assaults (European union agency for fundamental rights, 2014). Interviews reveal that victims often place less faith in the system, as they fear reprisal and retaliation by the offender (Rennison & Welchans, 2000; Tjaden & Thoennes, 2000), feel embarrassed (Birdsey & Snowball, 2013; European union agency for fundamental rights, 2014) or do not consider the police to understand or be able to effectively resolve their conflicts (i.e., implementing appropriate interventions to reduce the risk of further assaults; Birdsey & Snowball, 2013). Thus, the "true" recidivism rate is undetermined as substantial number of re-offenses may not be detected on the level of police crime statistics.

Fourthly, the measure of the ODARA may not be transferrable to different cultural and jurisdictional contexts. In the Zurich sample, only 3 of the 13 ODARA items were significantly correlated with IPV recidivism. Low levels of correlation with the criterion variable would not be critical if the ODARA was a scale with a high level of internal

consistency (Amelang & Schmidt-Atzert, 2006). However, the scale is an inappropriate measurement if acceptable internal consistency is not reached and significant correlations between its items and the criterion are not found (Lienert & Raatz, 1998).

#### **2.4.6. Limitations**

Although we attempted to strictly adhere to the development study protocol, several limitations must be noted. Firstly, police files were not always sufficient to fully score the ODARA. Sometimes, important issues were not broached or clarified. Besides the fact that missing information is allowed, two items frequently could not be scored: “Victim fears repetition of violence” (16.8% missing information) and “Offender has more than one indicator of substance abuse problem” (38.9% missing information). There is a considerable body of evidence indicating that both factors are associated with IPV. The contradicting results of the current study might be a function of the high level of missing information for these items. This source of error might be eliminated in a prospective analysis. Thirdly, the outcome criterion was defined as a new police-registered IPV incident. However, as we were bound to the police-specific data protection regulations, we only had access to police entries carried out within the Canton of Zurich. If an offender moved to another Swiss Canton, new IPV incidents were not detected. Nonetheless, testing the ODARA’s discrimination validity by additionally including new violent or sexual charges and convictions that were centrally reported for all offenses committed in Switzerland did not improve the discrimination of the ODARA.

#### **2.4.7. Conclusion**

In the current study, the ODARA was significantly associated with recidivism in a Swiss sample of IPV offenders. However, discrimination was much lower than in the developmental sample, and calibration was especially poor in the high-risk categories. On the

one hand, these deviations may be due to effects of interventions and/or contextual factors that interfere with an originally assessed risk disposition and thus do not reflect an actual lack of assessment quality by the ODARA. On the other hand, societal and jurisdictional differences between the country in which the ODARA was developed (Canada) and the site of the current replication (Switzerland) may play an important role; actuarial instruments are sample-based risk models and may be inappropriate for other contexts. Future research is needed to clarify these different sources of disruption and develop improved measures to screen for IPV recidivism in the Canton of Zurich.

## **2.5. Assessing the Risk of Severe Intimate Partner Violence: Validating the DyRiAS in Switzerland**

### **2.5.1. Abstract**

The aim of the present study was to investigate the of the Dynamic Risk Analysis System's (DyRiAS) performance in assessing the likelihood of lethal and potentially lethal intimate partner violence (IPV) in the Canton of Zurich, Switzerland. Police records were used to retrospectively administer the DyRiAS for 146 IPV offenders processed by the municipal police of Zurich in 2008. The sample was subsequently followed for between six months to five years. The ability of the six DyRiAS risk categories to discriminate between recidivists and non-recidivists was investigated using correlational and receiver operating characteristic curve analyses. DyRiAS assessments were not found to produce significant associations with lethal or potentially lethal IPV. No non-recidivists were assigned to the lowest risk categories of the instrument and none of the offenders of the highest risk category recidivated, despite intense police interventions have been implemented for all offenders at highest risk. On the basis of the current study, the ability of the DyRiAS to assess the risk of severe IPV could not be demonstrated. Nonetheless, further research is necessary to replicate these findings in larger samples and using prospective study designs.

### **2.5.2. Introduction**

Approximately four out of every 10 female homicide victims are murdered by their current or former intimate partner (38%; World Health Organisation, 2013). Systematic review evidence suggests that this statistic is particularly representative of high-income countries (Stöckl et al., 2013) and that female homicide victims are predominantly victims of such crimes. In the United States, 39% of female homicide victims are killed by their partners whereas this applies to only 3% of all American male homicides (Catalano, 2007). Similarly,

the rate of intimate partner homicide (IPH) in Canada has remained four times higher for women than for men over the past thirty years (Northcott, 2012). Outside of North America in England and Wales, 65% of all IPH victims are female (Dixon & Graham-Kevan, 2011). And this trend extends to Central Europe, where 41% of all police-registered female homicides in Germany (German Ministry of the Interior, 2014) and 58% in Switzerland are attributed to partners (Zoder, 2008).

Given that more than half of women killed in Switzerland are murdered by their partner, there is considerable interest on behalf of policymakers in the reliable and valid assessment of IPH risk (e.g. Schweizer Fernsehen, 2011). Of particular interest is the assessment of individuals reported to the police for domestic disturbances, as cases of IPH are frequently preceded by less severe forms of violence (65%-85%; Bailey et al., 1997; Browne, Williams, & Dutton, 1998; Dixon & Graham-Kevan, 2011; Moracco, Runyan, & Butts, 1998; Roehl, O'Sullivan, Webster, & Campbell, 2005).

### **Intimate Partner Homicide Risk Assessment**

There is a large evidence base suggesting that risk assessments made using structured methods outperform unstructured clinical judgments (Ægisdóttir et al., 2006; Grove et al., 2000; Hanson & Morton-Bourgon, 2009; Meehl, 1954). Despite a considerable number of instruments having been developed for the purpose of assessing intimate partner violence (IPV) risk and despite the publication of several articles on the use of those instruments for the prediction of femicide (e.g. the Ontario Domestic Assault Risk Assessment in Eke et al., 2011), we identified only one validated risk assessment instrument that has been developed focusing on IPH: the Danger Assessment (DA; Campbell et al., 2009). The DA is a 20-item instrument that is administered following an interview with an IPV victim. Specific domains captured by the items on the instrument include threatening and sexual violence,

characteristics of the relationship between the victim and offender, socio-demographic characteristics of the offender, and family status of the victim (Campbell et al., 2009).

The DA follows the “structured professional judgment” approach to structured risk assessment, which allows the administering clinician to combine risk and protective factors on the instrument as they see fit to come to a final risk judgment. This results in DA assessments being more vulnerable to cognitive biases that may reduce accuracy and reliability rates. In frontline assessment settings with large caseloads and a lack of clinically-trained staff in such as police departments, “mechanical” recidivism risk assessment instruments – instruments that follow inflexible rules to assign individuals into categories of low to high risk of recidivism – may be more practically useful (Hilton, Harris, & Rice, 2010).

### **The Dynamic Risk Analysis System**

An alternative to the more clinically-based DA has recently been proposed in the German-speaking region of Europe to provide a mechanical method of assessing the likelihood of severe, such as potentially lethal and lethal, incidents of IPV: The Dynamic Risk Analysis System (DyRiAS; Hoffmann & Glaz-Ocik, 2012). The DyRiAS was developed to assess the immediate (days to months) likelihood of severe IPV. The instrument contains static as well as dynamic items, which allows raters to monitor escalations in relational conflicts. The instrument is used as a Web-based application and its items are weighted relative to their importance in the prediction of IPH as determined by a literature review. Interdependencies between risk and protective factors for IPH are considered by incorporating a hierarchical decision making process (Hoffmann & Glaz-Ocik, 2012). Based on the item scoring on the DyRiAS, individuals are assigned to a specific risk category.

The DyRiAS is currently used in women’s support organizations and police departments in Germany, Austria, and Switzerland (Hoffmann: personal communication, July

25, 2013). A systematic literature search conducted using PubMed, PsycINFO, MEDLINE and Google Scholar identified only one previous publication examining the validity of the measure (Hoffmann & Glaz-Ocik, 2012). This retrospective validation study evaluated the concurrent validity of the DyRiAS using 61 cases of attempted IPH in Germany. The authors found that 82% ( $n = 50$ ) of the offenders were assigned to the two highest risk categories on the instrument. These findings may have been confounded, however, by the fact that assessors were not blinded to offenders' offenses when scoring the DyRiAS. Hence, additional research is needed to establish the validity of the instrument, especially research in applied settings.

### **The Present Study**

The objective of the present study was to investigate the discriminative validity of the DyRiAS in the prediction of both short-term as well as long-term IPV recidivism risk in the Canton of Zurich, Switzerland. Specifically, we aimed to conduct a cross-validation study in a police setting using a total cohort of male to female domestic violence cases in which an intimate partner was the perpetrator.

### **2.5.3. Methods**

#### **Participants**

In the present study, all cases of domestic violence processed by the municipal police of Zurich between January 1, 2008 and December 31, 2008 were eligible for inclusion ( $N = 342$ ). Of these, only adult men who either physically assaulted their female partner or issued them death threats with a weapon were included ( $N = 216$ ). An additional 39 offenders were excluded due to a lack of adequate information to score the DyRiAS and three who had no time at risk as they had never been released from prison or a forensic psychiatry within a follow up of five years ( $N = 174$ ). As the DyRiAS was designed to monitor risk over time, its

discrimination performance was measured over four lengths of time at risk: three months, six months, one year and five years. Offenders, who died, were deported to their home country or intermediately incarcerated and thus did not reach the necessary time at risk, were additionally dropped from the total sample. The four sub samples which served for discrimination analyses were the following: 3-months sample ( $n = 168$ ), 6-months sample ( $n = 167$ ), 1-year sample ( $n = 166$ ) and 5-years sample ( $n = 146$ ).

### **Procedure**

**Measures.** The DyRiAS consists of 39 dichotomous items which are answered via a Web-based application. The item content includes both static and dynamic risk factors, capturing situational, behavioral, and cognitive-emotional domains. To score the DyRiAS, information must be available for at least 55% of the instrument's items. Missing items are not prorated, but rather scored as "0". Based on the instrument's standardized weighting system, an individual is assigned into one of six risk categories (Category 0 = "No Risk"; Category 5 = "High Risk"). The hierarchical weighting of the items is automatically produced by a Web-based algorithm.

**Recidivism.** Consistent with the DyRiAS manual, recidivism was defined as an incident of severe IPV, operationalized as bodily harm, endangering the life of another, or (attempted) homicide. Only incidents registered by the police where the victim was a current or former female partner and which occurred after the index assault of IPV were considered acts of recidivism.

**Data collection.** DyRiAS assessments were conducted retrospectively by four psychologists trained in the use of the instrument and blind to offender outcomes. Scoring was based on police files, which included self-reports of the offender, the victim, and collaterals on the index incident as well as both medical evidence related to the index incident and information on the offender's previous contacts with the police. Data was collected at the

station of the municipal police of Zurich, where police staff instructed the raters in how to read their files. Uncertainty about item ratings were discussed with and clarified by an author of the DyRiAS, Dr. Jens Hoffmann.

### **Ethics approval**

In accordance with Swiss law, the present study did not need approval by the cantonal ethics committee but rather by the municipal data protection authority. Such approval was provided in March 2013.

### **Statistical Analysis**

The ability of DyRiAS to discriminate between recidivists and non-recidivists was investigated using correlational and receiver operating characteristic (ROC) curve analyses at a time at risk of three months, six months, one year, and five years in the community. All analyses were two-tailed and used a standard significance threshold of  $\alpha = .05$  in STATA/IC 13.0 for Windows (StataCorp, 2013).

## **2.5.4. Results**

### **Characteristics of the Total Sample**

At the time of the index incident, IPV offenders were an average age of 37.5 years ( $SD = 11.1$ ). The majority of offenders were not Swiss ( $n = 113, 64.9\%$ ) and approximately half of them were employed ( $n = 94, 54.0\%$ ). In about three-quarters of the cases, the index incident occurred in the victim's home ( $n = 135, 77.6\%$ ) and resulted in physical injury to the victim ( $n = 120, 69.0\%$ ). According to police files, the index incident involved threats in 91 cases (52.3%), minor assaults in 100 cases (57.5%), bodily harm in 81 cases (46.6%), and coercion in 56 cases (32.2%). At the time of the index incident, approximately half of the offenders were in a partnership with their victim ( $n = 88, 50.6\%$ ). Almost three-quarters (71.3%,  $n =$

124) of the offenders had previously been registered with an IPV assault in police files, and in four out of five cases the victim was repeatedly abused by the offender ( $n = 142$ , 81.6%).

### **DyRiAS Risk Category Distribution**

For no cases could all 39 DyRiAS items be scored based on file information, with an average of 36.5% ( $SD = 5.8\%$ ) of items unscored. While none of the offenders were assigned to risk categories 0 and 1, the mode as well as mean was Category 3, with only 4.0% of the sample being classified to the Category 5.

### **Rates of Recidivism**

No offender recidivated with a lethal offence. The recidivism rate of severe non-lethal IPV was 0.6% ( $n = 1$ ) within three months, 2.4% ( $n = 4$ ) within six months, 3.0% ( $n = 5$ ) within one year, and 8.9% ( $n = 13$ ) within five years of time at risk. Recidivism rates for each of the five DyRiAS risk category and within each of the four periods of time at risk are displayed in Table 12. Notably, no offender in Category 5 recidivated.

**Table 12.** DyRiAS risk category distribution and recidivism rates for IPV offenders at 3 months (n = 168), 6 months (n = 167), 1 year (n = 166) and 5 years (n = 146) time at risk.

Risk category	Percentage of sample in each risk category					Recidivism rate				
	Time at risk of 3 months	Time at risk of 6 months	Time at risk of 1 year	Time at risk of 5 years	Time at risk of 3 months	Time at risk of 6 months	Time at risk of 1 year	Time at risk of 5 years	Time at risk of 1 year	Time at risk of 5 years
0	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
1	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
2	26.8% (n = 45)	26.3% (n = 44)	26.5% (n = 44)	24.0% (n = 35)	0.0% (n = 0)	2.3% (n = 1)	2.3% (n = 1)	5.7% (n = 2)	2.3% (n = 1)	5.7% (n = 2)
3	45.8% (n = 77)	46.1% (n = 77)	46.4% (n = 77)	49.3% (n = 72)	0.0% (n = 0)	0.0% (n = 0)	1.3% (n = 1)	5.6% (n = 4)	1.3% (n = 1)	5.6% (n = 4)
4	23.8% (n = 40)	24.0% (n = 40)	24.1% (n = 40)	24.0% (n = 35)	2.5% (n = 1)	7.5% (n = 3)	7.5% (n = 3)	20.0% (n = 7)	7.5% (n = 3)	20.0% (n = 7)
5	3.6% (n = 6)	3.6% (n = 6)	3.0% (n = 5)	2.7% (n = 4)	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)
Overall recidivism base rate					0.6%	2.4%	3.0%	8.9%	3.0%	8.9%

Note. N.A. = Not applicable.

### **Discrimination Performance**

Severe IPV correlated positively but non-significantly with the DyRiAS risk categories across all subsamples (three months:  $r = .09$ ,  $p = .24$ ; six months:  $r = .09$ ,  $p = .26$ ; one year:  $r = .08$ ,  $p = .30$ ; five years:  $r = .13$ ,  $p = .11$ ). Similarly, ROC curve analyses revealed non-significant discrimination of recidivists and non-recidivists by the DyRiAS at three months ( $AUC = 0.85$ , 95%  $CI = 0.00-1.00$ ,  $p = .25$ ), six-months ( $AUC = 0.67$ , 95%  $CI = 0.32-1.00$ ,  $p = .26$ ), one year ( $AUC = 0.64$ , 95%  $CI = 0.36-0.92$ ,  $p = 0.30$ ), and five years follow-up ( $AUC = 0.64$ , 95%  $CI = 0.48 - 0.80$ ,  $p = 0.11$ ). *Post hoc* power analyses for discrimination analyses revealed that only the five year subsample was sufficiently large to detect effects in case “true” effects existed with a probability of 80% at a significance level of  $\alpha = .05$ .

### **Specificity Analyses**

Specificity analyses were conducted by considering differences in police interventions ordered after index assaults for IPV in those offenders followed-up for three months. For almost all offenders ( $n = 166$ , 98.8%), personal protective orders such as no-contact ( $n = 162$ , 97.0%), eviction ( $n = 109$ , 65.3%), and rayon orders (i.e., the designation of off-limit areas,  $n = 117$ , 70.1%) were issued. 77.4% ( $n = 130$ ) offenders were arrested and 69.0% ( $n = 116$ ) were held in custody. All of the offenders assigned to the highest risk category ( $n = 6$ ) did not recidivate. However, they were all arrested, held in custody and transferred to the federal prosecutor subsequent to the index assault. Of the offenders in the second highest DyRiAS category who did not receive such intensive interventions, also none (100%,  $n = 5$ ) recidivated (Table 13).

**Table 13.** Level of intervention for offenders of the 3-months subsample being assigned to the high-risk DyRiAS categories, which were issued by the police subsequent to the index assault

Level of intervention	DyRiAS category 4 (n = 40)		DyRiAS category 5 (n = 6)	
	non-recidivists	recidivists	non-recidivists	recidivists
no arrest + protective order	100% (n = 5)	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)
arrest + protective order	100% (n = 1)	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)
arrest + custody <sup>a</sup>	100% (n = 1)	0.0% (n = 0)	0.0% (n = 0)	0.0% (n = 0)
arrest + custody <sup>a</sup> + protective order	97.0% (n = 32)	3.0% (n = 1)	100% (n = 6)	0.0% (n = 0)

<sup>a</sup>custody while one's case is being transferred to the federal prosecutor

### 2.5.5. Discussion

The present study investigated the ability of a recently developed risk assessment instrument – the DyRiAS – to discriminate between IPV offenders who went on to commit acts of potentially lethal or lethal IPV and those who did not. A total cohort of IPV offenders processed by the municipal police of Zurich was followed for between three months to five years. Randomly selected recidivists were found to be classified into higher DyRiAS risk categories more often than not, but no significant evidence of discrimination was found.

Although these discrimination findings have important implications for practitioners who use the DyRiAS, discrimination represents only part of the criterion validity. The other part is the calibration (Rossegger et al., 2014; Singh, 2013) and measures the fit between the expected and observed rates of recidivism (Schmid & Griffith, 2005). However, normative recidivism rates for the six DyRiAS risk categories have not been published, thus, the evaluation of the instrument's calibration is currently limited to a descriptive analysis of how recidivists are distributed across the instrument's six risk categories. As no recidivists were in the highest risk category and no non-recidivists were in the lowest risk category, the calibration of the DyRiAS in the Swiss intimate partner offender population appears to be wanting.

In accordance with Risk-Needs-Responsivity Principles (Bonta & Andrews, 2007), IPV offenders judged to be at higher risk of recidivism can benefit from intensive police interventions (cf. Belfrage et al., 2012). Thus, our finding that no offenders in the highest DyRiAS risk category recidivated could be attributed to effective intervention. However, the same argument cannot be made for explaining why there were no recidivists in the next highest risk category under those offenders not receiving such intensive interventions at all. As it is the DyRiAS' aim to display the immediate risk of lethal or potentially lethal IPV, higher rates of recidivism are expected in "high risk" categories when offenders received only a comparably low level of intervention.

#### **2.5.6. Limitations**

There were several limitations of the present study beyond the small sample size that limited the statistical power needed null hypothesis significance testing. First, the current study was a retrospective, file-based study which has implications for the quality and completeness of the data collected. In 18% of the cases that were processed by the police in 2008, the DyRiAS was unable to be administered due to missing information. In the remainder, our descriptive analyses suggested that the average offender had 37% of necessary information missing, meaning that more than a third of the information needed to administer the DyRiAS seems to be not normally collected in routine police practice. Hence, there is a discrepancy at this time between the information collected by the police and the information that is necessary to administer the DyRiAS. This may be important, as previous research on mechanical instruments has found that discrimination increases with fewer missing items (e.g. G. T. Harris & Rice, 2003). It should be the aim of a prospective study to test whether missing information can be made accessible through elaborated witness interviews or if the findings of the present study reflect a limitation of the practical usefulness of the DyRiAS in the Canton of Zurich.

Second, the practical utility of the DyRiAS was not assessed. Thus, it may be beneficial to conduct a qualitative investigation such as a survey or set of interviews with police to examine the perceived usefulness of the DyRiAS in not only risk assessment but also the development and monitoring of risk management plans. A recent international survey of psychiatrists, psychologists, and nurses suggests that practitioners rate violence risk assessment instruments differently in their utility in such tasks (Singh et al., 2014).

Third, lethal or potentially lethal recidivism was defined as subsequent bodily harm, endangering the lives of another, and (attempted) homicide. As bodily harm may not reflect potentially lethal acts in all cases this may have led to an overestimation of recidivism. However, excluding such incidents would also underestimate the rate of severe violent recidivism. Relatedly, outcome information was only able to be collected within the Canton of Zurich, possibly resulting in an underestimation of recidivism.

### **2.5.7. Conclusion**

The current study is the first validation study of the DyRiAS, which (to our knowledge) is the only mechanical instrument developed for the assessment of potentially lethal and lethal intimate partner violence risk. Keeping in mind the limitations of the current study, the DyRiAS was found to moderately, but non-significantly discriminate between recidivists and non-recidivists of severe forms of IPV within time at risk periods of three months to five years. On the basis of the current study, the ability of the DyRiAS to assess the risk of severe IPV could not be demonstrated. Nonetheless, further research is necessary to replicate these findings in larger samples and using a prospective study design.

## **2.6. Identifikation von Hoch-Risiko-Drohungen**

### **2.6.1. Zusammenfassung**

Drohungen kündigen (schwere) Gewaltdelikte an. Wenngleich empirische Untersuchungen auf einen Zusammenhang zwischen Drohungen und Gewalttätigkeit hinweisen, ist der prädiktive Wert einer Drohung für die Begehung von Gewaltdelikten vergleichsweise gering. Um die Ausführungsgefahr von Drohungen beurteilen zu können, müssen weitere risikorelevante Faktoren einbezogen werden. Nachfolgend wird ein Modell vorgeschlagen, welches es erlaubt, Hoch-Risiko-Drohungen von ungefährlichen Drohungen zu unterscheiden.

### **2.6.2. Drohungen**

#### **Erscheinungsform von Drohungen**

Drohungen drücken den Wunsch oder die Absicht aus, einer Person Schaden zuzufügen und deren körperliche Unversehrtheit zu verletzen (Meloy, 2001). Charakteristika und Erscheinungsform von Drohungen können unterschiedlich sein und von mehrdeutigen, bedrohlich interpretierbaren Aussagen bis hin zu sehr spezifischen und direkten Konfrontationen reichen. O'Toole (2000) schlug für eine Klassifikation von Drohungen folgende Einteilung anhand formaler und inhaltlicher Kriterien vor:

- Direkte Drohung: Eindeutige und unmissverständliche Ankündigung einer spezifischen Gewalthandlung gegen ein bestimmtes Ziel.
- Indirekte Drohung: Weniger konkretisierte Gewaltandrohung, Details darüber oder über das angestrebte Ziel werden entweder nicht spezifiziert oder ausgelassen.
- Maskierte Drohung: Aussagen, die Gewalthandlungen andeuten, können aber müssen nicht als Drohung interpretiert werden.

- Konditionale Drohung: Gewalthandlungen sind an das Eintreten spezifischer Wünsche und Bedingungen gekoppelt.

Ein weiteres Kriterium zur Klassifikation von Drohungen ist das Ausmaß der in der Drohung enthaltenen verbalen Aggression. Die Overt Aggression Scale (OAS; Silver & Yudofsky, 1991) erlaubt die Einordnung der Drohung auf einer vierstufigen Skala: *Stufe 1*: Laute Geräusche und wütendes Rufen, *Stufe 2*: Geringfügige Beleidigungen, *Stufe 3*: Ausstoßen von böartig erscheinendem Fluchen und milde Gewaltandrohung gegenüber Dritten oder der eigenen Person und *Stufe 4*: Explizite und unmissverständliche Gewaltandrohung gegen andere oder sich selbst, oder Ersuchen um Hilfe zur Selbstkontrolle.

### **Prävalenz und Kontext der Drohung**

**Drohungen.** Länderübergreifend konnte aufgezeigt werden, dass die meisten zur Anzeige gebrachten Drohungen gegenüber Personen aus dem näheren Umfeld ausgesprochen werden (Bundesamt für Statistik Schweiz, 2010; Daffern & Howells, 2002; Warren et al., 2011). So findet z.B. in der Schweiz jede dritte polizeilich registrierte Drohung (31%) im häuslichen Kontext statt (Bundesamt für Statistik Schweiz, 2010). Abgesehen vom häuslichen Bereich werden Drohungen häufig im Kontext von Behandlungen psychischer Störungen oder im Rahmen von Auseinandersetzungen mit Behördenmitgliedern beziehungsweise Staatsangestellten bekannt. (Adams, Hazelwood, Pitre, Bedard, & Landry, 2009; Bundesamt für Statistik Schweiz, 2010; Calhoun, 1998; Hatch-Maillette, Scalora, Bader, & Bornstein, 2007). Calhoun (1998) untersuchte in den USA über 3000 Drohungsfälle gegenüber Justizmitarbeitern zwischen 1980 und 1993. Demnach waren Richter, unter den Justizangestellten, die am häufigsten von der Drohung betroffene Berufsgruppe. Drohungen, die sich gegen Justizmitarbeiter richten, erfolgen mehrheitlich im Zusammenhang mit einem Gerichtsprozess, in den der Täter persönlich involviert ist.

Verschiedene Untersuchungen haben aufgezeigt, dass Mitarbeiter psychiatrischer Krankenhäuser häufig Drohungen ausgesetzt sind. Bei einer Befragung von Mitarbeitern eines forensisch-psychiatrischen Krankenhauses in den USA mit Langzeit- und akutpsychiatrischen Patienten, gaben vier von fünf Mitarbeitern (83%) an, während der Berufsausübung mindestens einmal bedroht worden zu sein, wobei mehr als die Hälfte der Betroffenen von einem Vorfall innerhalb der letzten 12 Monate berichtete (Hatch-Maillette et al., 2007). Im Kontext des Behandlungssettings haben Berufserfahrung und Geschlecht einen Einfluss auf die Wahrscheinlichkeit, Bedrohungssituationen ausgesetzt zu sein: Junge und unerfahrene Psychiater und Psychologen werden häufiger bedroht als ihre älteren und erfahreneren Kollegen (Carmel & Hunter, 1991; Nijman, Bowers, Oud, & Jansen, 2005). Und Frauen werden im Vergleich zu Männern signifikant häufiger indirekt bedroht und sexuell belästigt (siehe Daffern & Howells, 2002; Hatch-Maillette et al., 2007; Nijman et al., 2005). Des Weiteren spielt die Kontaktintensität der Mitarbeiter eine wichtige Rolle: Je enger und fortdauernder der Kontakt zum Patienten, desto höher ist die Wahrscheinlichkeit, von diesem bedroht oder angegriffen zu werden (Binder & McNiel, 1994; Daffern & Howells, 2002; Hatch-Maillette et al., 2007).

**Todesdrohungen.** Todesdrohungen kündigen ein Tötungsdelikt an und stellen somit eine besonders schwere Form von Drohungen dar. Warren et al. (2011) untersuchten in Australien 144 Personen, deren Gefährlichkeit zwischen 2002 und 2005 wegen einer Todesdrohung ambulant in einem klinischen Setting abgeklärt wurde. Die Autoren konnten zeigen, dass auch für Todesdrohungen gilt, dass die ins Hellfeld der Justiz gelangenden Todesdrohungen überwiegend gegen Personen gerichtet sind, die dem Drohenden gut bekannt sind (siehe Abbildung 13). So waren Familienmitglieder, Freunde und Arbeitskollegen (39%) sowie (ehemalige oder aktuelle) Beziehungspartner (29%) am häufigsten betroffen.. Todesdrohungen wurden in 10% der Fälle an Mitarbeiter psychiatrischer

Kliniken, in 8% an Mitarbeiter der Justiz und jeweils 1% an Kinderschutzbeauftragte, prominente und unbekannte Personen adressiert. Weitere 11% richteten sich z.B. an Firmen oder Opfer früherer Delikte (Warren et al., 2011). Die Todesdrohungen wurden in jedem zweiten Fall im direkten Kontakt (51%) und dabei fast immer unter Verwendung einer Waffe (90%) ausgesprochen – wobei es sich bei der Waffe nur in wenigen Fällen um eine Schusswaffe handelte (Warren et al., 2011).

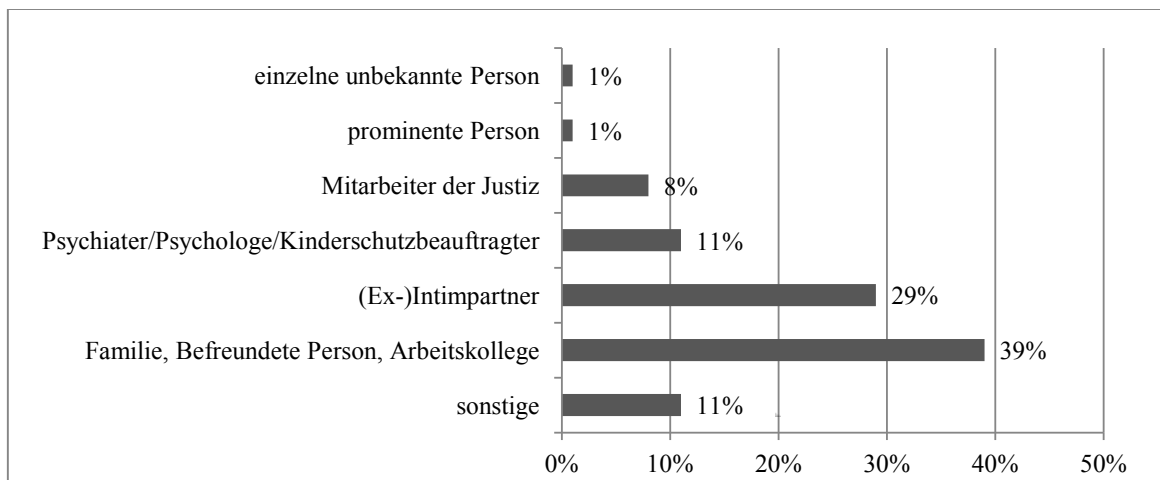


Abbildung 13: Zielpersonen von Todesdrohungen (Warren et al., 2011)

### 2.6.3. Drohungen und schwere Gewaltdelikte

Drohungsdelikte haben nicht nur einen die betroffene Person verängstigenden und erschreckenden Charakter, sondern kündigen auch die Begehung von (schweren) Gewaltdelikten an. Kommt es zu einer Drohung, stellt sich somit die Frage, wie hoch die Ausführungsgefahr einzuschätzen ist und welchen prädiktiven Wert die Drohung für sich genommen aufweist.

Insgesamt legen Untersuchungen an Straftätern die Annahme eines Zusammenhangs zwischen Drohungen und Gewalthandlungen nahe. So konnte wiederholt gezeigt werden, dass Tötungsdelikten häufig eine Drohung vorausgeht. Dies gilt gleichermaßen für Personen, die das Tötungsdelikt in einem psychotischen Zustand (Nitschke, Osterheider, & Mokros, 2011), im Rahmen intimer Beziehungen (Belfrage & Rying, 2004) oder als Attentat auf

Politiker begingen (Vossekuil, Reddy, & Fein, 2000). Ebenso waren häufig auch Suizidandrohungen beziehungsweise suizidales Verhalten im Vorfeld schwerer Gewaltdelikte zu beobachten (Fein & Vossekuil, 1999). In einzelnen Studien konnte überdies ein prädiktiver Wert einer Drohung für spätere Gewalthandlungen aufgezeigt werden: In (ehemaligen) Paarbeziehungen ausgesprochene Todesdrohungen korrelieren mit nachfolgend gewalttätigen Handlungen. Dieser Befund findet darin Niederschlag, dass das Vorliegen einer Todesdrohung eins von dreizehn Items des Ontario Domestic Assault Risk Assessment (ODARA; Hilton et al., 2004), einem mechanischen Risk-Assessment Instrument zur Beurteilung des Risikos häuslicher Gewalt, darstellt.

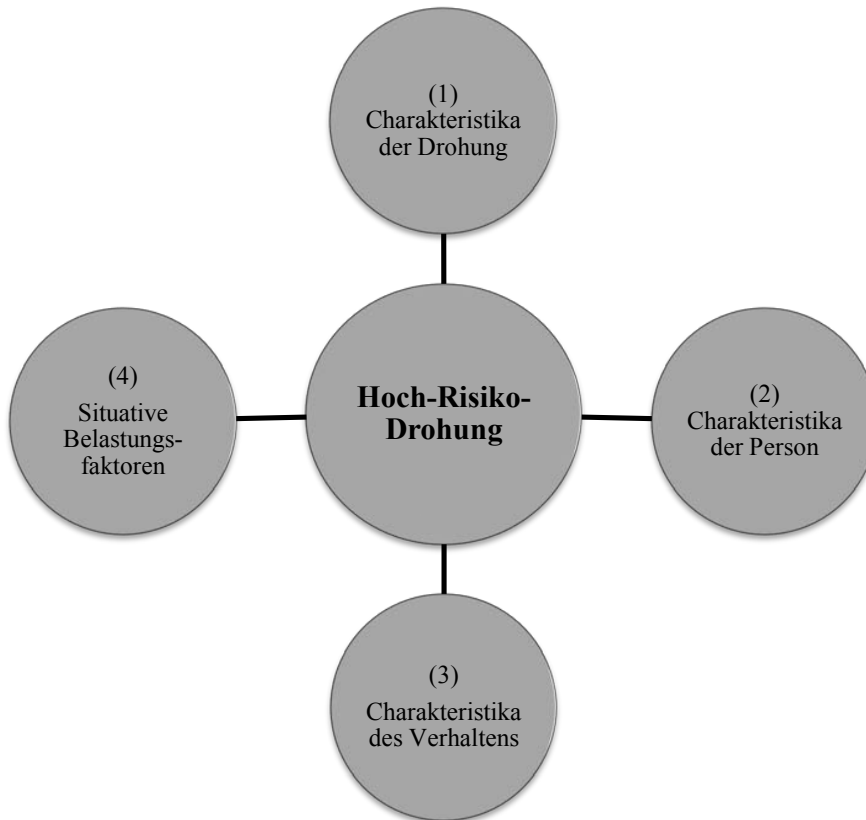
Dennoch muss davon ausgegangen werden, dass Drohungen allein kein hinreichendes Kriterium darstellen, um von einem hohen Gewaltisiko ausgehen zu können. So verlangt auch das ODARA, dass neben der Todesdrohung mindestens ein weiteres kritisches Merkmal erfüllt sein muss, damit von einem abklärungsbedürftigen Gewaltisiko im Rahmen einer bedrohlichen Situation ausgegangen werden kann.

Die wenigen empirischen Studien über den weiteren Verlauf von Personen, die gedroht hatten, zeigen auf, dass nur ein Teil der Drohungen in gewalttätigen Handlungen mündet. Als wie hoch dieser Anteil anzunehmen ist, scheint von dem Vorliegen weiterer Merkmale der Drohung und der drohenden Person abzuhängen (Urbaniok, Rossegger, Steinfeld, & Endrass, 2006). Calhoun (1998) wies in seiner Untersuchung von bedrohten Justizmitarbeitern aus, dass auf die Drohung lediglich in 4% der Fälle eine gewalttätige Handlung folgte. Zu weitaus höheren Prävalenzangaben kamen Warren et al. (2011; 2008) in ihrer prospektiv angelegten Untersuchung von Personen, die Todesdrohungen ausgesprochen hatten: Auf durchschnittlich eine von drei Drohungen folgte in den nachfolgenden Jahren eine Gewalthandlung (Warren et al., 2011; Warren et al., 2008). Knapp ein Viertel (23%), der durch Warren et al. (2011) beschriebenen Personen, die Todesdrohungen ausgesprochen hatten, beging innerhalb von

zwölf Monaten mindestens ein Gewaltdelikt – davon 18% ein Schweres. Bei jedem Zehnten richtete sich das Gewaltdelikt gegen das zuvor bedrohte Opfer. Zu schweren Gewaltdelikten im Sinne einer (versuchten) Tötung kam es bei 1%. Die gleiche Forschungsgruppe (Warren et al., 2008) berichtet, bei vergleichbarem Studiendesign, für einen Beobachtungszeitraum von zehn Jahren eine 44%igen Gewaltdeliktsrate, wovon es sich in 3% um Tötungsdelikte handelte.

#### **2.6.4. Hoch-Risiko-Drohungen**

Einerseits legen empirische Untersuchungen einen Zusammenhang zwischen Drohungen und Gewaltdelikten nahe, andererseits stellen sie für sich genommen keinen hinreichenden Prädiktor für die Beurteilung des Gewaltpotenzials einer Person dar. Es ist also weder die Schlussfolgerung zulässig, dass Drohungen harmlos sind, noch darf allein aus dem Vorliegen einer Drohung auf ein hohes Gewaltisiko der drohenden Person geschlossen werden. Um zwischen risikorelevanten (Hoch-Risiko-Drohungen) und nicht risikorelevanten Drohungen unterscheiden zu können, müssen weitere Parameter definiert und berücksichtigt werden. Hoch-Risiko-Drohungen sind durch bestimmte Charakteristika der Drohung, Persönlichkeitsmerkmale, Verhaltensmerkmale und/oder möglichen situativen Belastungsfaktoren der drohenden Person gekennzeichnet (siehe Abbildung 14).



*Abbildung 14:* Hoch-Risiko-Drohung

### **Charakteristika der Drohung**

Unter Berücksichtigung von Erscheinungsform und Inhalt der Drohung stellte O'Toole (2000) Kriterien auf, anhand derer sich Drohungen in eine von drei Risikokategorien (niedrig, moderat und hoch) überführen lassen (siehe Abbildung 15). Die Nennung spezifischer Details zum Motiv und den Mitteln der Tatumsetzung (z.B. Waffeneinsatz) zu Opfer und Tatort sind ein Hinweis auf eine besonders risikohafte Drohung (Hoch-Risiko-Drohung) – allerdings nur, wenn die Angaben plausibel erscheinen und sich auf realistische und nachvollziehbare Umstände (z.B. eines tatsächlich erreichbaren Opfers) beziehen.

Ein Mangel an Details, eine Häufung von unwichtigen Informationen oder unlogischen Inhalten der Drohung weisen auf ein geringes mit der Drohung verbundenes Gewaltisiko hin. Der emotionale Gehalt einer Drohung sollte bei Beurteilung des Risikopotenzials der Drohung nicht berücksichtigt werden. Da dieser zwar Schlüsse auf den emotionalen Zustand

der drohenden Person zum Zeitpunkt der Drohung zulässt, nicht jedoch auf eine grundsätzliche Stimmungslage der Person, die Rückschlüsse auf das allgemeine Ausführungspotential ermöglichen würde (O'Toole, 2000).

niedriges Risiko	moderates Risiko	hohes Risiko
<ul style="list-style-type: none"> <li>•Die Drohung ...</li> <li>•... ist vage und indirekt formuliert.</li> <li>•... ist inkonsistent, wenig plausibel, nicht detailliert.</li> <li>•Das Drohungsszenario ist nicht realistisch.</li> </ul>	<ul style="list-style-type: none"> <li>•Die Drohung ist ...</li> <li>•... konkret und direkt formuliert.</li> <li>•... enthält detailliertere Angaben zum Vorgehen bei der angedrohten Tat.</li> <li>•... enthält Hinweise auf Ort und Zeitpunkt der Tat.</li> <li>•... enthält mehrdeutige Anspielungen zu Vorbereitungsmaßnahmen (z.B. Anspielung auf einen Film oder grundsätzliche Möglichkeit, sich Waffen zu besorgen).</li> <li>•... enthält eine dezidierte Unterstreichung der Ernsthaftigkeit der Drohung.</li> </ul>	<ul style="list-style-type: none"> <li>•Die Drohung ...</li> <li>•... ist direkt, plausibel und detailliert formuliert.</li> <li>•... enthält Hinweise auf konkrete Vorbereitungsmaßnahmen (Besorgen einer Waffe etc.).</li> <li>•... enthält Hinweise auf eine Beobachtung (Ausspähung) des potentiellen Opfers.</li> <li>•... offenbart konkrete Kenntnisse über das Tatumfeld.</li> </ul>

**Abbildung 15:** Beurteilung des Risikopotenzials der Drohung auf einer 3-stufigen Skala (O'Toole 2000)

### Charakteristika der Person

Bei der Beurteilung der Ausführungsgefahr einer Drohung sollten immer auch Merkmale der drohenden Person berücksichtigt werden. Dabei kommt schweren psychiatrischen Erkrankungen sowie mit Gewalttätigkeit assoziierten früheren Handlungen eine besondere Bedeutung zu. Schwere psychische Erkrankungen – insbesondere Schizophrenien und (dissoziale) Persönlichkeitsstörungen – stellen einen Risikofaktor für die Begehung von Gewaltdelikten dar. Die Kombination aus dem Vorliegen einer schweren psychischen Erkrankung und einer Drohung muss als Risikofaktor für Gewalthandlungen bewertet werden (McNiel & Binder, 1989; Mullen et al., 2009).

Warren et al. (2011) gelang es, auf Grundlage der von ihnen analysierten Fälle von Todesdrohungen, ein empirisches Modell für auf die Drohung folgende Gewaltdelikte (innerhalb von zwölf Monaten) zu entwickeln ( $AUC = .76$ ): Gewaltdelikte waren signifikant

mit Substanzmittelmissbrauch, fehlender Behandlung bei psychisch Kranken, geringem Bildungsniveau (zehn Schulklassen oder weniger) und früheren Gewaltdelikten assoziiert. In einer früheren Studie der Autoren mit einem Beobachtungszeitraum von zehn Jahren wurde zudem auf den hohen prädiktiven Wert wahnhafter Erkrankungen für die Begehung von Tötungsdelikten hingewiesen (Warren et al., 2008). Beide Studien verdeutlichen, dass die Drohung im Kontext mehrerer Risikofaktoren interpretiert werden muss und für sich allein genommen keinen ausreichenden Prädiktor für Gewalthandlungen darstellt.

Folgende Personenmerkmale tragen zu einem erhöhten Ausführungsrisiko einer Drohung bei und weisen damit auf das Vorliegen einer Hoch-Risiko-Drohung hin:

- Wiederholte regelverletzende und dissoziale Verhaltensweisen und ein Mangel an Verankerung gesellschaftlicher Normen – bis hin zu einer dissozialen Persönlichkeitsstörung oder einer hohen Ausprägung psychopathischer Eigenschaften,
- wahnhaftes Erleben (insbesondere Verfolgungswahn) beispielsweise im Rahmen einer Schizophrenie oder eines wahnhaften Syndroms und die ausbleibende Behandlung psychischer Störungen,
- frühere oder aktuelle Gewalthandlungen (diese muss nicht zwingend zu einer Verurteilung geführt haben),
- aktueller oder früherer Einsatz von Waffen (während einer Drohung oder einer anderen strafbaren Handlung),
- Waffenaffinität, die sich im Sammeln von Waffen, dem Tragen von Waffen im Alltag, Verherrlichung von Waffen und Mitgliedschaften in Schützenvereinen darstellen kann,
- missbräuchlicher Konsum psychotroper Substanzen oder das Vorliegen einer Substanzabhängigkeit und/oder

- latente oder akute Suizidalität.

### **Warnverhalten**

Als dritter Bereich muss – neben den Charakteristika der Drohung und personalen Merkmalen – das aktuelle Verhalten der drohenden Person bei der Gefährlichkeitsbeurteilung berücksichtigt werden. Sogenanntes Warnverhalten spricht für einen hohen Beschäftigungsgrad mit den Drohungsinhalten. Die meisten Drohenden, die kurz davor stehen, das angedrohte Verhalten auszuführen, führen bestimmte Vorbereitungshandlungen durch – wie z.B. die gezielte und zeitintensive Planung oder die Kommunikation gegenüber anderen (Borum, Fein, Vossekuil, & Berglund, 1999). Warnverhalten zeigt sich jedoch nicht nur in Form von konkreten Vorbereitungshandlungen, sondern z.B. auch schon durch ein direktes – von Angesicht zu Angesicht – Aussprechen einer Drohung. So gibt es bei gegenüber Behörden ausgesprochenen Drohungen z.B. Hinweise darauf, dass schriftliche oder telefonische Drohungen seltener zu nachfolgenden Gewalthandlungen führen als im persönlichen, direkt verbalen Kontakt geäußerte Drohungen, die in Kombination mit auffälligem und verdächtigem Verhalten auftreten ((z.B. Vandalismus im unmittelbaren Umfeld des Opfers vgl. Calhoun, 2001).

Aufgrund ihrer hohen Relevanz zur Einschätzung des Ausführungsrisikos unternahmen Meloy et al. (2012) einen ersten Strukturierungsversuch verschiedener Warnverhalten zur Integration ins Drohungsassessment (vgl. Meloy et al., 2012, pp. 10-11):

- *Pathway warning behavior*: Jedes Verhalten, das auf eine intensive Beschäftigung mit Planung, Vorbereitung und Umsetzung einer Gewalttat hinweist.
- *Fixation warning behavior*: Jedes Verhalten, das auf eine zunehmend starke Wahrnehmungseinengung auf eine Person oder einen Konflikt hinweist.
- *Identification warning behavior*: Jedes Verhalten, das durch eine ungewöhnliche Anziehung gegenüber militärischen und kriegerischen Inhalten gekennzeichnet ist.

Die Person weist z.B. eine Waffenaffinität auf, identifiziert sich mit bekannten Attentätern oder sieht sich selbst als Agent einer wichtigen Mission.

- *Novel aggression warning behavior*: Eine erstmalige Gewalttat, die zeitlich an die Drohung geknüpft ist, sich aber nicht gegen die bedrohte Person richtet und für die drohende Person den Charakter eines «Testlaufs» hat, indem sie überprüft, ob sie zur Ausführung von Gewalttätigkeiten überhaupt in der Lage ist und welche Reaktionen zu erwarten sind.
- *Energy burst warning behavior*: Eine Zunahme von (harmlosen) Aktivitäten, die sich an das Opfer richten (z.B. Briefe schreiben, Kontaktaufnahme).
- *Leakage warning behavior*: Eine Mitteilung der drohenden Person an Dritte, dass sie die Absicht hat, einer anderen Person gezielt Schaden zuzufügen.
- *Last resort warning behavior*: Ein Verhalten, das die Annahme der drohenden Person widerspiegelt, dass eine Gewalttat ein «logischer nächster Schritt» oder ein «notwendiger nächster Schritt» sei und sich durch zunehmende Verzweiflung oder Not der drohenden Person ausdrückt. Die Person fühlt sich gefangen und sieht keine andere Möglichkeit als gewalttätig zu handeln, wobei die mit dem Gewaltdelikt verbundenen Konsequenzen als gerechtfertigt akzeptiert werden.

### **Akute Belastungsfaktoren**

Während der Fokus der Beurteilung des Rückfallrisikos bei Gewalt- und Sexualstraftätern eindeutig auf der Analyse personaler Merkmale liegt und situativen Umständen nur eine untergeordnete Rolle beigemessen wird, kommt diesen und daraus resultierenden akuten Belastungsfaktoren für die Wahrscheinlichkeit der Ausführung einer Drohung eine wichtige Rolle zu. Das bedeutet auch, dass sich die Beurteilung der Ausführungsgefahr einer Drohung immer auf einen begrenzten Zeitraum bezieht und in Abhängigkeit von der weiteren

Entwicklung aktualisiert werden muss (Meloy et al., 2012). Als akute, risikoe erhöhende Belastungsfaktoren können beispielsweise genannt werden:

- Es tritt die Situation ein, an deren Eintreten eine konditionale Drohung („wenn, dann...“) geknüpft war. Die drohende Person gerät unter Zugzwang.
- Bei einem jahrelangen Rechtsstreit steht das Urteil der letztmöglichen Instanz kurz bevor und ein ablehnender Entscheid ist wahrscheinlich. Legale Möglichkeiten die eigenen Interessen oder Forderungen durchzusetzen sind erschöpft.
- Es kommt zu einer Zuspitzung der sozialen Situation der drohenden Person: Beziehungen gehen auseinander, finanzielle Sorgen, Streit um das Sorgerecht für die Kinder etc.
- Aber auch das Absetzen von Medikamenten und die erneute Entwicklung einer akuten Psychose können eine solche Belastungssituation darstellen.

#### **2.6.5. Ausblick**

Eine Drohung an sich ist noch kein Prädiktor für anschließende Gewalthandlungen. Vielmehr spielen zur Beurteilung der Ausführungsgefahr neben Art und Inhalt der Drohung zusätzlich bestimmte Persönlichkeits-, Verhaltens- und situative Merkmale des Drohenden eine wichtige Rolle. Zur Einschätzung des Gefährlichkeitspotenzials muss ebenso in erster Linie die Plausibilität einer Drohung stehen sowie die Bewertung darüber, ob der Drohende überhaupt über die Mittel der Umsetzung (nötige Ressourcen, Zugang zu Tatwaffen, Ernsthaftigkeit des Vorsatzes und Motivation) verfügt (O'Toole, 2000).

Sind Auffälligkeiten auch nur in einem der vier Risikobereiche – Drohungsart, Persönlichkeit, Verhalten und Belastungsfaktoren des Drohenden – anzutreffen, ist eine genaue Abklärung der drohenden Person indiziert. Je mehr kritische Merkmale sich in diesen Bereichen als zutreffend erweisen, desto höher kann das Risiko für eine Ausführung der angedrohten bzw. einer äquivalenten Tat angenommen werden.

Drohungen sollten immer und in jedem Kontext zur Anzeige und damit ins Hellfeld gebracht werden. Nur so ist es möglich, die bisherigen Befunde des Risk-Assessments auszubauen und geeignete Managementstrategien zur Deeskalation von bedrohlichen Situationen zu entwickeln. Das vorliegende Hoch-Risiko-Drohungs-Modell (siehe Abbildung 14) ist eine unter Berücksichtigung aktueller wissenschaftlicher Erkenntnisse umfassende und strukturierte Grundlage zur Analyse und Bewertung von Drohungssituationen sowie der Umsetzungsgefahr von Gewaltdelikten.

---

### 3. Literaturverzeichnis

- Adams, S. J., Hazelwood, T. E., Pitre, N. L., Bedard, T. E., & Landry, S. D. (2009). Harassment of members of parliament and the legislative assemblies in Canada by individuals believed to be mentally disordered. *Journal of Forensic Psychiatry & Psychology*, 20(6), 801-814. doi: 10.1080/14789940903174063
- Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist*, 34(3), 341-382. doi: 10.1177/0011000005285875
- Agentur der Europäischen Union für Grundrechte. (2014). Gewalt gegen Frauen: Eine EU-weite Erhebung. Luxemburg.
- Amelang, M., & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4 ed.). Heidelberg: Springer Medizin.
- Andrews, D. A. (1989) Recidivism is predictable and can be influenced: Using risk assessments to reduce recidivism. *Forum on Corrections Research: Vol. 1* (pp. 11-18).
- Andrews, D. A., & Bonta, J. (2001). *The Level of Service Inventory-Revised. User's Manual*. Toronto: Multi-Health Systems
- Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5 ed.). New Providence, NJ: LexisNexis Matthew Bender.
- Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior*, 17, 19-52.

- Archer, J. (2000). Sex differences in aggression between heterosexual partners: A meta-analytic review. *Psychological Bulletin*, *126*(5), 651-680. doi: 10.1037//0033-2909.126.5.651
- Archer, J. (2002). Sex differences in physically aggressive acts between heterosexual partners: A meta-analytic review. *Aggression and Violent Behavior*, *7*(4), 313-351. doi: 10.1016/S1359-1789(01)00061-1
- Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment*, *87*(1), 84-94. doi: 10.1207/s15327752jpa8701\_07
- Bailey, J. E., Kellermann, A. L., Somes, G. W., Banton, J. G., Rivara, F. P., & Rushford, N. P. (1997). Risk factors for violent death of women in the home. *Archives of Internal Medicine*, *157*(7), 777-782. doi: 10.1001/archinte.1997.00440280101009
- Barbaree, H. E., Langton, C. M., Blanchard, R., & Cantor, J. M. (2009). Aging versus stable enduring traits as explanatory constructs in sex offender recidivism. *Criminal Justice and Behavior*, *36*(5), 443-465. doi: 10.1177/0093854809332283
- Barbaree, H. E., Langton, C. M., & Peacock, E. J. (2006). Different actuarial risk measures produce different risk rankings for sexual offenders. *Sex Abuse*, *18*, 423-440. doi: 10.1177/107906320601800408
- Bartosh, D. L., Garby, T., Lewis, D., & Gray, S. (2003). Differences in the predictive validity of actuarial risk assessments in relation to the sex offenders type. *International Journal of Offender Therapy and Comparative Criminology*, *47*(4), 422-438. doi: 10.1177/0306624X03253850
- Baxstrom v Herold, 383 US 107 C.F.R. (1966).

- Belfrage, H., & Rying, M. (2004). Characteristics of spousal homicide perpetrators: A study of all cases of spousal homicide in Sweden 1990–1999. *Criminal Behaviour and Mental Health, 14*(2), 121-133. doi: 10.1002/cbm.577
- Belfrage, H., & Strand, S. (2012). Measuring the outcome of structured spousal violence risk assessments using the B-SAFER: Risk in relation to recidivism and intervention. *Behavioral Sciences & the Law, 30*(4), 420-430. doi: 10.1002/bsl.2019
- Belfrage, H., Strand, S., Storey, J. E., Gibas, A. L., Kropp, P. R., & Hart, S. D. (2012). Assessment and management of risk for intimate partner violence by police officers using the Spousal Assault Risk Assessment guide. *Law and Human Behavior, 36*(1), 60-67. doi: 10.1037/h0093948
- Benda, B. B. (2005). Gender differences in life-course theory of recidivism: A survival analysis. *International Journal of Offender Therapy and Comparative Criminology, 49*(3), 325-342. doi: 10.1177/0306624X04271194
- Binder, R. L., & McNiel, D. E. (1994). Staff gender and risk of assault on doctors and nurses. *Journal of the American Academy of Psychiatry and the Law Online, 22*(4), 545-550.
- Birdsey, E., & Snowball, L. (2013). Reporting violence to police: A survey of victims attending domestic violence services *Crime and Justice Statistics*. Sydney.
- Black, M. C., Basile, K. C., Breiding, M. J., Smith, S. G., Walters, M. L., Merrick, M. T., . . . Stevens, M. R. (2011). The National Intimate Partner and Sexual Violence Survey (NISVS): 2010 Summary Report *Drug and Alcohol Dependence*. Atlanta, GA.
- Boetticher, A., Kröber, H.-L., Müller-Isberner, R., Böhm, K. M., Müller-Metz, R., & Wolf, T. (2007). Mindestanforderungen für Prognosegutachten. *Neue Zeitschrift für Strafrecht, 26*, 537-544.

- Bonta, J., & Andrews, D. A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation: Public Safety Canada.
- Bonta, J., Law, M., & Hanson, R. K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: a meta-analysis. *Psychological Bulletin*, *123*(2), 123-142. doi: 10.1037/0033-2909.123.2.123
- Borum, R., Fein, R., Vossekuil, B., & Berglund, J. (1999). Threat assessment: Defining an approach for evaluating risk of targeted violence. *Behavioral Sciences & the Law*, *17*(3), 323-337.
- Bowen, E. (2011). An overview of partner violence risk assessment and the potential role of female victim risk appraisals. *Aggression and Violent Behavior*, *16*(3), 214-226. doi: 10.1016/j.avb.2011.02.007
- Breslow, N. (1990). Biostatistics and Bayes. *Statistical Science*(5), 269-298.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1-3.
- Browne, A., Williams, K. R., & Dutton, D. C. (1998). Homicide between intimate partners. In M. D. Smith & M. Zah (Eds.), *Homicide: A sourcebook of social research* (pp. 149–164). Thousand Oaks, CA: Sage.
- Bundesamt für Statistik Schweiz. (2008). Tötungsdelikte in der Partnerschaft *Statistik der Schweiz*. Neuchâtel: Bundesamt für Statistik Schweiz.
- Bundesamt für Statistik Schweiz. (2010). Polizeiliche Kriminalstatistik. Neuchâtel: Sektion Kriminalität und Strafrecht.
- Bundesamt für Statistik Schweiz. (2014). Polizeiliche Kriminalstatistik *Statistik der Schweiz*. Schweiz.

- Bundesministerium des Innern. (2013). Polizeiliche Kriminalstatistik 2012, Bundesrepublik Deutschland. Berlin.
- Bundestag Bundesrepublik Deutschland Gesetz zum zivilrechtlichen Schutz vor Gewalttaten und Nachstellungen (Gewaltschutzgesetz - GewSchG) (2001).
- Calhoun, F. S. (1998). Hunters and howlers: Threats and violence against federal judicial officials in the United States, 1980-1993: US Department of Justice US Marshals Service Arlington.
- Calhoun, F. S. (2001). Violence toward judicial officials. *The ANNALS of the American Academy of Political and Social Science*, 576(1), 54-68. doi: 10.1177/000271620157600105
- Campbell, J. C., Webster, D. W., & Glass, N. (2009). The Danger Assessment: Validation of a lethality risk assessment instrument for intimate partner femicide. *Journal of Interpersonal Violence*, 24, 653-674. doi: 10.1177/0886260508317180
- Cantonal Court of Zurich. (2012). Decree of June 5th, 2012. from [http://www.gerichte-zh.ch/fileadmin/user\\_upload/entscheide/oeffentlich/UG070045-O2.pdf](http://www.gerichte-zh.ch/fileadmin/user_upload/entscheide/oeffentlich/UG070045-O2.pdf)
- Carlin, B. P., & Louis, T. A. (2009). Approaches to statistical inferences *Bayesian methods for data analysis* (3 ed.). Boca Raton, FL: CRC Press.
- Carmel, H., & Hunter, M. (1991). Psychiatrists injured by patient attack. *Journal of the American Academy of Psychiatry and the Law Online*, 19(3), 309-315.
- Catalano, S. (2007). *Intimate partner violence in the United States*. Washington: Department of Justice Statistics.

- Cho, H. (2012). Examining gender differences in the nature and context of intimate partner violence. *Journal of Interpersonal Violence, 27*(13), 2665-2684. doi: 10.1177/0886260512436391
- Daffern, M., & Howells, K. (2002). Psychiatric inpatient aggression: A review of structural and functional assessment approaches. *Aggression and Violent Behavior, 7*(5), 477-497. doi: 10.1016/s1359-1789(01)00073-8
- De Becker, G. (2000). Domestic Violence Method–DV-MOSAIC. Retrieved from <http://www.mosaicmethod.com>
- de Ruiter, C., & Nicholls, T. L. (2011). Protective factors in forensic mental health: A new frontier. *International Journal of Forensic Mental Health, 10*(3), 160-170. doi: 10.1080/14999013.2011.600602
- Dennis, J. A., Khan, O., Ferriter, M., Huband, N., Powney, M. J., & Duggan, C. (2012). Psychological interventions for adults who have sexually offended or are at risk of offending (Review): Issue 12. Art. No.: CD007507. DOI: 10.1002/14651858.CD007507.pub2.
- Diamond, B. L. (1974). The psychiatric prediction of dangerousness. *University of Pennsylvania Law Review, 123*(2), 439-452.
- Dixon, L., & Graham-Kevan, N. (2011). Until death do they part: Preventing intimate partner homicide. *The Psychologist, 24*(11), 820-823.
- Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20V3: Assessing risk of violence – User guide*. Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.
- Dutton, D. G. (2012). The prevention of intimate partner violence. *Prevention Science, 13*(4), 395-397. doi: 10.1007/s11121-012-0306-1

- Edwards, W., Lindman, H., & Savag, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193-242. doi: 10.1037/h0044139
- Eher, R., Matthes, A., & Rettenberger, M. (2012). Der STABLE-2007 – ein Instrument zur Erfassung des stabil-dynamischen Rückfallrisikos bei Sexualstraftätern. In J. Endrass, A. Rossegger, F. Urbaniok & B. Borchard (Eds.), *Interventionen bei Gewalt- und Sexualstraftätern: Risk-Management, Methoden und Konzepte der forensischen Therapie*. Berlin: Medizinisch wissenschaftliche Verlagsgesellschaft.
- Eher, R., Matthes, A., Schilling, F., Haubner-MacLean, T., & Rettenberger, M. (2012). Dynamic risk assessment in sexual offenders using STABLE-2000 and the STABLE-2007: An investigation of predictive and incremental validity. *Sex Abuse*, *24*(5), 5-28. doi: 10.1177/1079063211403164
- Eher, R., Rettenberger, M., Schilling, F., & Pfäfflin, F. (2008a). Validität oder praktischer Nutzen? Prückfallvorhersagen mittels Static-99 und SORAG. Eine prospektive Rückfallstudie an 275 Sexualstraftätern [Validity or benefit? The prediction of relapses with Static-99 and SORAG. A prospective study with 275 sex offenders]. *Recht und Psychiatrie*, *26*(2), 79-88.
- Eher, R., Rettenberger, M., Schilling, F., & Pfäfflin, F. (2008b). Failure of Static-99 and SORAG to predict relevant reoffense categories in relevant sexual offender subtypes: A prospective study. *Sexual Offender Treatment*, *3*(1), 1-14.
- Eke, A. W., Hilton, N. Z., Harris, G. T., Rice, M. E., & Houghton, R. E. (2011). Intimate partner homicide: Risk assessment and prospects for prediction. *Journal of Family and Violence*, *26*, 211-216. doi: 10.1007/s10896-010-9356-y
- Endrass, J., Rossegger, A., Frischknecht, A., Noll, T., & Urbaniok, F. (2008). Using the Violence Risk Appraisal Guide (VRAG) to predict in-prison aggressive behavior in a

- Swiss offender population. *International Journal of Offender Therapy and Comparative Criminology*, 52(1), 81-89. doi: 10.1177/0306624X07301643
- Endrass, J., Rossegger, A., & Urbaniok, F. (2012). Häusliche Gewalt im Kanton Zürich - Evaluation der polizeilichen Schutzmassnahmen im Kanton Zürich gemäss kantonalem Gewaltschutzgesetz für den Zeitraum der Inkraftsetzung des Gesetzes vom 1. April 2007 – 31. Dezember 2009. Zürich: Psychiatrisch-Psychologischer Dienst des Kanton Zürich.
- Endrass, J., Urbaniok, F., Held, L., Vetter, S., & Rossegger, A. (2009). Accuracy of the Static-99 in predicting recidivism in Switzerland. *International Journal of Offender Therapy and Comparative Criminology*, 53(4), 482-490. doi: 10.1177/0306624X07312952
- European union agency for fundamental rights. (2014). Gewalt gegen Frauen: Eine EU-weite Erhebung [Violence against women: an EU-wide survey] Luxemburg.
- Falzer, P. R. (2013). Valuing structured professional judgment: Predictive validity, decision-making, and the clinical-actuarial conflict. *Behavioral Sciences and the Law*, 31, 40-54. doi: 10.1002/bsl.2043
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874. doi: 10.1016/j.patrec.2005.10.010
- Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. *British Medical Journal*, 345, 1-12. doi: 10.1136/bmj.e4692

- Fein, R. A., & Vossekuil, B. (1999). Assassination in the United States: An operational study of recent assassins, attackers, and near-lethal approachers. *Journal of Forensic Sciences, 44*(2), 321-333.
- Felson, R. B., Ackerman, J. M., & Gallagher, C. A. (2005). Police intervention and the repeat of domestic assault. *Criminology, 43*(3), 563-588. doi: 10.1111/j.0011-1348.2005.00017.x
- Fleiss, J. L., Levin, B., & Paik, M. C. (2003). The measurement of interrater agreement. In D. J. Balding, N. A. C. Cressie, N. I. Fisher, I. M. Johnstone, J. B. Kadane, L. M. Ryan, D. W. Scott, A. F. M. Smith & J. L. Teugels (Eds.), *Statistical methods for rates and proportions* (3 ed., pp. 598-626). New York: John Wiley & Sons.
- Folkes, S. E. F., Hilton, N. Z., & Harris, G. T. (2013). Weapon use increases the severity of domestic violence but neither weapon use nor firearm access increases the risk or severity of recidivism. *Journal of Interpersonal Violence, 28*, 1143-1156. doi: 10.1177/0886260512468232
- Gelles, R., & Tolman, R. M. (1998). *The Kingston Screening Instrument for Domestic Violence (KSID)*. Unpublished risk instrument. University of Rhode Island. Providence.
- Generalversammlung der Vereinten Nationen. (1993). Resolution 48/104.
- German Ministry of the Interior. (2014). Polizeiliche Kriminalstatistik 2013, Bundesrepublik Deutschland [Police Crime Statistics 2013, Germany]. Wiesbaden.
- Gerth, J., & Graber, C. (2012). Identifikation von Hoch-Risiko-Drohungen. In J. Endrass, A. Rossegger, F. Urbaniok & B. Borchard (Eds.), *Interventionen bei Gewalt- und Sexualstraftätern: Risk-Management, Methoden und Konzepte der forensischen Therapie* (pp. 393-401). Berlin: Medizinisch wissenschaftliche Verlagsgesellschaft.

- 
- Gillioz, L., De Puy, J., & Ducret, V. (1997). *Domination et violence envers la femme dans le couple*. Payot Lausanne.
- Gloor, D., & Meier, H. (2004). Repräsentativbefragung bei Patientinnen der Maternité Inselhof Triemli, Klinik für Geburtshilfe und Gynäkologie, Zürich *Frauen, Gesundheit und Gewalt im sozialen Nahraum*: Büro für die Gleichstellung von Frau und Mann der Stadt Zürich und Maternité Inselhof Triemli.
- Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin*, 73(6), 422-432.
- Grann, M., & Pallvik, A. (2002). An empirical investigation of written risk communication in forensic psychiatric evaluations. *Psychology, Crime and Law*, 8, 113-130. doi: 10.1080/10683160290000923
- Greenfeld, L. A., Rand, M. R., Craven, D., Klaus, P. A., Perkins, C. A., Ringel, C., . . . Fox, J. A. (1998). Violence by intimates: Analysis of data on crimes by current or former spouses, boyfriends, and girlfriends. In U.S. Department of Justice (Ed.), *Bureau of Justice Statistics Factbook*.
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, 2(2), 293-323.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological assessment*, 12(1), 19-30. doi: 10.1037//1040-3590.12.1.19
- Guo, B., & Harstall, C. (2008). Spousal violence against women: Preventing recurrence *IHE Report*. Alberta, Canada.

- Guy, L. S. (2008). *Performance indicators of the structured professional Judgment approach for assessing risk for violence to others: A meta-analytic survey*. (PhD), Simon Fraser University, Burnaby, BC, Canada.
- Guy, L. S., Packer, I. K., & Warnken, W. (2012). Assessing risk of violence using Structured Professional Judgment Guidelines. *Journal of Forensic Psychology Practice, 12*(3), 270-283. doi: 10.1080/15228932.2012.674471
- Hanson, R. K. (2005). Twenty years of progress in violence risk assessment. *Journal of Interpersonal Violence, 20*, 212-217. doi: 10.1177/0886260504267740
- Hanson, R. K. (2012). Static-99/R FAQ. Retrieved July 2012, from <http://www.static2099.org/pdffdocs/faq.pdf>.
- Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). A meta-analysis of the effectiveness of treatment for sexual offenders: Risk, need, and responsivity: Public Safety Canada.
- Hanson, R. K., Harris, G. T., Scott, T. L., & Helmus, L. (2007). Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project.
- Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological assessment, 21*(1), 1-21. doi: 10.1037/a0014421
- Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24*(1), 119-136.
- Hare, R. D. (2003). *Manual for the revised Psychopathy Checklist*. Toronto ON: Multi-Health Systems

- Harris, A., Phenix, A., Thornton, D., & Hanson, R. K. (2003). *Static 99: Coding rules revised 2003*. Canada: Solicitor General
- Harris, G. T., & Rice, M. E. (2003). Actuarial assessment of risk among sex offenders. *Annals of the New York Academy of Sciences, 989*, 198-210.
- Harris, G. T., & Rice, M. E. (2007). Characterizing the value of actuarial violence risk assessments. *Criminal Justice and Behavior, 34*(12), 1638-1658. doi: 10.1177/0093854807307029
- Harris, G. T., & Rice, M. E. (2013). Bayes and base rates: What is an informative prior for actuarial violence risk assessment? *Behavioral Sciences and the Law, 31*, 103-124. doi: 10.1002/bsl.2048
- Harris, G. T., Rice, M. E., & Camilleri, J. A. (2004). Applying a forensic actuarial assessment (the Violence Risk Appraisal Guide) to nonforensic patients. *Journal of Interpersonal Violence, 19*(9), 1063-1074. doi: 10.1177/0886260504268004
- Harris, G. T., Rice, M. E., & Cormier, C. A. (2002). Prospective replication of the Violence Risk Appraisal Guide in predicting violent recidivism among forensic patients. *Law and Human Behavior, 26*, 377-394.
- Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumière, M. L., Boer, D., & Lang, C. (2003). A multisite comparison of actuarial risk instruments for sex offenders. *Psychological assessment, 15*(3), 413-425. doi: 10.1037/1040-3590.15.3.413
- Hart, S. D., & Logan, C. (2011). Formulation of violence risk using evidence-based Assessments: The structured professional judgment approach. In P. Sturmey & M. McMurrin (Eds.), *Forensic case formulation*. Chichester, UK: Wiley-Blackwell.
- Hastings, M. E., Krishnan, S., Tangney, J. P., & Stuewig, J. (2011). Predictive and incremental validity of the Violence Risk Appraisal Guide scores with male and

- female jail inmates. *Psychological assessment*, 23(1), 174-183. doi: 10.1037/a0021290
- Hatch-Maillette, M. A., Scalora, M. J., Bader, S. M., & Bornstein, B. H. (2007). A gender-based incidence study of workplace violence in psychiatric and forensic settings. *Violence and Victims*, 22(4), 449-462. doi: 10.1891/088667007781553982
- Heilbrun, K., Douglas, K. S., & Yasuhara, K. (2009). Controversies in violence risk assessment. In J. K. Skeem, K. S. Douglas & S. O. Lilienfeld (Eds.), *Psychological science in the courtroom: Controversies and consensus* (pp. 333-356). New York, NY: Guilford.
- Helmus, L., & Bourgon, G. (2011). Taking stock of 15 years of research on the Spousal Assault Risk Assessment guide (SARA): A critical review. *International Journal of Forensic Mental Health*, 10(1), 64-75. doi: 10.1080/14999013.2010.551709
- Hilton, N. Z., Carter, A. M., Harris, G. T., & Sharpe, A. J. B. (2008). Does using nonnumerical terms to describe risk aid violence risk communication? *Journal of Interpersonal Violence*, 23(2), 171-188. doi: 10.1177/0886260507309337
- Hilton, N. Z., & Harris, G. T. (2009). How nonrecidivism affects predictive accuracy: Evidence from a cross-validation of the Ontario Domestic Assault Risk Assessment. *Journal of Interpersonal Violence*, 24, 326-337. doi: 10.1177/0886260508316478
- Hilton, N. Z., Harris, G. T., & Holder, N. (2008). Actuarial assessment of violence risk in hospital-based partner assault clinics. *Canadian Journal of Nursing Research*, 40(4), 56-70.
- Hilton, N. Z., Harris, G. T., Popham, S., & Lang, C. (2010). Risk assessment among incarcerated male domestic violence offenders. *Criminal Justice and Behavior*, 37(8), 815-832. doi: 10.1177/0093854810368937

- Hilton, N. Z., Harris, G. T., Rawson, K., & Beach, C. A. (2005). Communicating violence risk information to forensic decision makers. *Criminal Justice and Behavior, 32*(97-116). doi: 10.1177/0093854804270630
- Hilton, N. Z., Harris, G. T., & Rice, M. E. (2006). Sixty-six years of research on the clinical versus actuarial prediction of violence. *The Counseling Psychologist, 34*(3), 400-409. doi: 10.1177/0011000005285877
- Hilton, N. Z., Harris, G. T., & Rice, M. E. (2007). The effect of arrest on wife assault recidivism: Controlling for pre-arrest risk. *Criminal Justice and Behavior, 34*, 1334-1344. doi: 10.1177/0093854807300757
- Hilton, N. Z., Harris, G. T., & Rice, M. E. (2010). *Risk assessment for domestically violent men*. Washington DC: American Psychological Association
- Hilton, N. Z., Harris, G. T., Rice, M. E., Eke, A. W., & Lowe-Wetmore, T. (2007). Training front-line users in the Ontario Domestic Assault Risk Assessment (ODARA): A tool for police domestic investigations. *The Canadian Journal of Police & Security Services, 5*(1/2), 92-96.
- Hilton, N. Z., Harris, G. T., Rice, M. E., Houghton, R. E., & Eke, A. W. (2008). An indepth actuarial assessment for wife assault recidivism: The Domestic Violence Risk Appraisal Guide. *Law and Human Behavior, 32*, 150–163. doi: 10.1007/s10979-007-9088-6
- Hilton, N. Z., Harris, G. T., Rice, M. E., Lang, C., Cormier, C. A., & Lines, K. J. (2004). A brief actuarial assessment for the prediction of wife assault recidivism: The Ontario domestic assault risk assessment. *Psychological assessment, 16*(3), 267-275. doi: 10.1037/1040-3590.16.3.267

- Hilton, N. Z., Popham, S., Lang, C., & Harris, G. T. (2014). Preliminary validation of the ODARA for female intimate partner violence offenders. *Partner Abuse, 5*(2), 189-203. doi: 10.1891/1946-6560.5.2.189
- Hilton, N. Z., & Simmons, J. L. (2001). The influence of actuarial risk assessment in clinical judgments and tribunal decisions about mentally disordered offenders in maximum security. *Law and Human Behavior, 25*(4), 393-408.
- Hoffmann, J., & Glaz-Ocik, J. (2012). DyRiAS-Intimpartner: Konstruktion eines online gestützten Analyse-Instrumentes zur Risikoeinschätzung von tödlicher Gewalt gegen aktuelle oder frühere Intimpartnerinnen [DyRiAS intimate partner: Constructing of a Web-based instrument for assessing the risk of lethal violence against a current or former female intimate partner]. *Polizei & Wissenschaft, 2*, 45-57.
- Hoffmann, J., & Roshdi, K. (2013). School shootings in Germany: Research, prevention through risk assessment and threat assessment. In N. Böckler, T. Seeger, P. Sitzer & W. Heitmeyer (Eds.), *School shootings: International research, case studies, and concepts for prevention* (pp. 363-378). New York: Springer.
- Holland-Davis, L., & Davis, J. (2014). Victim arrest in intimate partner violence incidents: A multilevel test of black's theory of law. *The Journal of Public and Professional Sociology, 6*(1), 8.
- Holtzworth-Munroe, A., Meehan, J. C., Herron, K., Rehman, U., & Stuart, G. L. (2003). Do subtypes of maritally violent men continue to differ over time? *Journal of Consulting and Clinical Psychology, 71*(4), 728-740. doi: 10.1037/0022-006X.71.4.728
- Jackson, R. L., & Hess, D. T. (2007). Evaluation for civil commitment of sex offenders: A survey of experts. *Sexual Abuse: A Journal of Research and Treatment, 19*, 409-448. doi: 10.1007/s11194007-9062-3

- Kantonspolizei Zürich. (2014). Gewaltschutz. Retrieved 19.10.2014, from <http://www.kapo.zh.ch/internet/sicherheitsdirektion/kapo/de/fach/gewaltschutz.html>
- Law on Protection against Violence (LPV) [Gewaltschutzgesetz], 351 C.F.R. (2006).
- Kantonsrat Kanton Zürich Schweiz Law on Protection against Violence (LPV) [Gewaltschutzgesetz] (2006).
- Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*(91), 1343-1370.
- Kilvinger, F., Rossegger, A., Urbaniok, F., & Endrass, J. (2012). Risikokalkulation bei häuslicher Gewalt. *Fortschritte der Neurologie - Psychiatrie*, 80(6), 312-319. doi: 10.1055/ s-0031-1273200
- Kozol, H. L., Boucher, R. J., & Garofalo, R. F. (1972). The diagnosis and treatment of dangerousness. *Crime & Delinquency*, 18(4), 371-392. doi: 10.1177/001112877201800407
- Kröner, C., Stadtland, C., Eidt, M., & Nedopil, N. (2007). The validity of the Violence Risk Appraisal Guide (VRAG) in predicting criminal recidivism. *Criminal Behavior and Mental Health*, 17, 89-100. doi: 10.1002/cbm.644
- Kroner, D. G., & Mills, J. F. (2001). The accuracy of five risk appraisal instruments in predicting institutional misconduct and new convictions. *Criminal Justice and Behavior*, 28(4), 471-489. doi: 10.1177/009385480102800405
- Kropp, P. R., & Hart, S. D. (2000). The Spousal Assault Risk Assessment (SARA) guide: reliability and validity in adult male offenders. *Law and Human Behavior*, 24(1), 101-118. doi: 10.1023/A:1005430904495

- 
- Kropp, P. R., Hart, S. D., & Belfrage, H. (2005). *Brief spousal assault form for the evaluation of risk (B-SAFER). User manual*. Vancouver, BC: ProActive ReSolutions.
- Kropp, P. R., Hart, S. D., & Lyon, D. (2008). *Guidelines for stalking assessment and management (SAM)*. Vancouver, BC: ProActive ReSolutions.
- Kropp, P. R., Hart, S. D., Webster, C. D., & Eaves, D. (1998). *Spousal Assault Risk Assessment: User's Guide*. Toronto: Multi-Health Systems, Inc.
- Kropp, P. R., Hart, S. D., Webster, C. W., & Eaves, D. (1995). *Manual for the Spousal Assault Risk Assessment Guide* (2 ed.). Vancouver, BC: British Columbia Institute on Family Violence.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.
- Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L., & Hansen, K. T. (2007). Actuarial assessment of risk for reoffense among adult sex offenders: Evaluating the predictive accuracy of the Static-2002 and five other instruments. *Criminal Justice and Behavior*, 34, 37-59. doi: 10.1177/0093854806291157
- Latessa, E., Listwan, S., & Koetzle, D. (2013). *What works (and doesn't) in reducing recidivism*. Waltham, MA: Elsevier Anderson Publishing.
- Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6 ed.). Weinheim: Beltz, Psychologie Verlags Union.
- Lindsay, W. R., Todd, H. E., Taylor, J., Steptoe, L., Mooney, P., O'Brien, G., . . . Smith, A. H. W. (2008). Risk assessment in offenders with intellectual disability: A comparison across three levels of security. *International Journal of Offender Therapy and Comparative Criminology*, 52, 90-111. doi: 10.1177/0306624X07308111

- Lipsey, M. W., & Cullen, F. T. (2007). The effectiveness of correctional rehabilitation: A review of systematic reviews. *The Annual Review of Law and Social Science*, 3, 297–320. doi: 10.1146/annurev.lawsocsci.3.081806.112833
- Logan, T. K., & Walker, R. (2009). Civil protective order effectiveness: justice or just a piece of paper? *Violence and Victims*, 25(3), 332-348.
- Looman, J. (2006). Comparison of Two Risk Assessment Instruments for Sexual Offenders. *Sexual Abuse: A Journal of Research and Treatment*, 18(2). doi: 10.1007/s11194-006-9013-4
- Loza, W., Villeneuve, D. B., & Loza-Fanous, A. (2002). Predictive validity of the Violence Risk Appraisal Guide: A tool for assessing violent offender's recidivism. *International Journal of Law and Psychiatry*, 25(1), 85-92. doi: 10.1016/S0160-2527(01)00092-9
- Lurigio, A. J., & Taxman, F. S. (2013). Forensic assessment of risk in criminal justice. In J. B. Helfgott (Ed.), *Criminal Psychology* (Vol. 3, pp. 3-19). Santa Barbara, CA: Praeger.
- McNiel, D. E., & Binder, R. L. (1989). Relationship between preadmission threats and later violent behavior by acute psychiatric inpatients. *Hospital & Community Psychiatry*, 40(6), 605-608.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence* (P. E. Meehl Ed.). Minneapolis: University of Minnesota.
- Meloy, J. R. (2001). Threats, stalking, and criminal harassment. In G. F. Pinard & L. Pagani (Eds.), *Threats, stalking, and criminal harassment* (pp. 238-257). New York: Cambridge University Press.

- Meloy, J. R., Hart, S. D., & Hoffmann, J. (2014). Threat assessment and threat management. In J. R. Meloy & J. Hoffmann (Eds.), *International Handbook of Threat Assessment* (pp. 3-17). Oxford: University Press.
- Meloy, J. R., Hoffmann, J., Guldemann, A., & James, D. (2012). The role of warning behaviors in threat assessment: An exploration and suggested typology. *Behavioral Sciences & the Law*, *30*(3), 256-279. doi: 10.1002/bsl.999
- Meloy, J. R., White, S. G., & Hart, S. D. (2013). Workplace assessment of targeted violence risk: the development and reliability of the WAVR-21. *Journal of Forensic Sciences*, *58*(5), 1353-1358. doi: 10.1111/1556-4029.12196
- Melton, H. C., & Sillito, C. L. (2012). The role of gender in officially reported intimate partner abuse. *Journal of Interpersonal Violence*, *27*(6), 1090-1111. doi: 10.1177/0886260511424498
- Mental Health Centre Penetanguishene. (2005). *Ontario Domestic Assault Risk Assessment: General Scoring Criteria*. Penetanguishene, ON Canada.
- Messing, J. T., & Thaller, J. (2013). The average predictive validity of intimate partner violence risk assessment instruments. *Journal of Interpersonal Violence*, *28*(7), 1537-1558. doi: 10.1177/0886260512468250
- Monahan, J. (1984). The prediction of violent behavior: Toward a second generation of theory and policy *American Journal of Psychiatry*, *141*(1), 10-15.
- Monahan, J. (1996). Violence prediction: The past twenty and the next twenty years. *Criminal Justice and Behavior*, *23*, 107-120. doi: 10.1177/0093854896023001008
- Moracco, K. E., Runyan, C. W., & Butts, J. (1998). Femicide in North Carolina. *Homicide Studies*, *2*, 422-446.

- 
- Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology, 62*(4), 783-792.
- Mossman, D. (2006). Another look at interpreting risk categories. *Sexual Abuse: A Journal of Research and Treatment, 18*(1), 41-63. doi: 10.1007/s11194-006-9001-8
- Mossman, D. (2013). Evaluating risk assessments using receiver operating characteristic analysis: Rationale, advantages, insights, and limitations. *Behavioral Sciences and the Law, 31*, 23-39. doi: 10.1002/bsl.2050
- Mullen, P. E., James, D. V., Meloy, J. R., Pathé, M. T., Farnham, F. R., Preston, L., . . . Berman, J. (2009). The fixated and the pursuit of public figures. *Journal of Forensic Psychiatry & Psychology, 20*(1), 33-47. doi: 10.1080/14789940802197074
- Nationalrat Republik Österreich Bundesgesetz zum Schutz vor Gewalt in der Familie (GeSchG) (1997).
- Nijman, H., Bowers, L., Oud, N., & Jansen, G. (2005). Psychiatric nurses' experiences with inpatient aggression. *Aggressive Behavior, 31*(3), 217-227. doi: 10.1002/ab.20038
- Nitschke, J., Osterheider, M., & Mokros, A. (2011). Schizophreniforme Erkrankungen, Psychose und Tötungsdelikte: Die Bedeutung sozialtherapeutischer Maßnahmen zur Prävention von Delikten. *Psychiatrische Praxis, 38*(2), 82-86. doi: 10.1055/s-0030-1248603
- Northcott, M. (2012). Intimate partner violence risk assessment tools: A review: Canada Department of Justice, Research Statistics Division.
- Nunes, K. L., Firestone, P., Bradford, J. M., Greenberg, D. M., & Broom, I. (2002). A comparison of modified versions of the Static-99 and the Sex Offender Risk Appraisal Guide. *Sexual Abuse: A Journal of Research and Treatment, 14*, 253-269.

- O'Toole, M. E. (2000). *The school shooter: A threat assessment perspective*. Quantico, VA: National Center for the Analysis of Violent Crime, Federal Bureau of Investigation.
- Office for National Statistics. (2013). *Focus on: Violent Crime and Sexual Offences England and Wales, 2011/12 Statistical Bulletin*.
- Olszowy, L., Jaffe, P. G., Campbell, M., Hazel, L., & Hamilton, A. (2013). Effectiveness of risk assessment tools in differentiating child homicides from other domestic homicide cases. *Journal of Child Custody, 10*(2), 185-206. doi: 10.1080/15379418.2013.796267
- Ontario Ministry of the Solicitor General. (2000). *A guide to the domestic violence supplementary report form*. Toronto, Canada.
- Pham, T. H., & Ducro, C. (2008). Risk assessment in social defence: Preliminary factorial analysis of the Sex Offender Recidivism Appraisal Guide (SORAG) and the Static-99. *Annales Médico-Psychologiques, 166*, 575-579. doi: 10.1016/j.amp.2008.06.001
- Phenix, A., Hanson, R. K., Harris, A. J. R., & Thornton, D. (2012). Static-99 and related risk assessment research. Retrieved July 2012, from <http://www.static99.org>
- Quinsey, V. L., & Ambtman, R. (1979). Variables affecting psychiatrists' and teachers' assessments of the dangerousness of mentally ill offenders. *Journal of Consulting and Clinical Psychology, 47*(2), 353-362.
- Quinsey, V. L., Book, A., & Skilling, T. A. (2004). A follow-up of deinstitutionalized men with intellectual disabilities and histories of antisocial behaviour. *Journal of Applied Research in Intellectual Disabilities, 17*(4), 243-253. doi: 10.1111/j.1468-3148.2004.00216.x
- Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2 ed.). Washington, DC: American Psychological Association.

- 
- Rabkin, J. G. (1979). Criminal behavior of discharged mental patients: A critical appraisal of the research. *Psychological Bulletin*, 86(1), 1-27.
- Rennison, C. M., & Welchans, S. (2000). Intimate partner violence. In Bureau of Justice Statistics (Ed.), *Special Report*. U.S. Department of Justice.
- Rettenberger, M., & Eher, R. (2013). Actuarial risk assessment in sexually motivated intimate partner violence. *Law and Human Behavior*, 37(2), 75-86. doi: 10.1037/b0000001
- Rettenberger, M., Gaunersdorfer, K., & Eher, R. (2009). *Deutsche Übersetzung und Adaption des Ontario Domestic Assault Risk Assessment*. Wien, Österreich: Verein für forensische Forschung und Weiterbildung.
- Rettenberger, M., Matthes, A., Boer, D. P., & Eher, R. (2009). Prospective actuarial risk assessment: A comparison of five risk assessment instruments in different sexual offender subtypes. *International Journal of Offender Therapy and Comparative Criminology*, 54(2), 169-186. doi: 10.1177/0306624x08328755
- Rice, M. E., & Harris, G. T. (2002). Men who molest their sexually immature daughters: Is a special explanation required? *Journal of Abnormal Psychology*, 111(2), 329-339. doi: 10.1037//0021-843X.111.2.329
- Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior*, 29(5), 615-620. doi: 10.1007/s10979-005-6832-7
- Roehl, J., O'Sullivan, C., Webster, D. W., & Campbell, J. (2005). Intimate partner violence risk assessment validation study: Final report: U.S. Department of Justice.
- Rogers, W. (1992). Brier score decomposition. *Stata Technical Bulletin*, 20-22.

- Rossegger, A., Endrass, J., & Gerth, J. (2012). Einführung ins Risk-Assessment. In J. Endrass, A. Rossegger, F. Urbaniok & B. Borchard (Eds.), *Interventionen bei Gewalt- und Sexualstraftätern: Risk-Management, Methoden und Konzepte der forensischen Therapie* (pp. 91-97). Berlin: Medizinisch wissenschaftliche Verlagsgesellschaft.
- Rossegger, A., Endrass, J., Gerth, J., & Singh, J. P. (2014). Replicating the Violence Risk Appraisal Guide: A total forensic cohort study. *PLoS One*, *9*(3), 1-8. doi: 10.1371/journal.pone.0091845
- Rossegger, A., Gerth, J., Seewald, K., Urbaniok, F., Singh, J. P., & Endrass, J. (2013). Current obstacles in replicating risk assessment findings: A systematic review of commonly used actuarial instrument. *Behavioral Sciences & the Law*, *31*(1), 154-164. doi: 10.1002/bsl.2044
- Rossegger, A., Gerth, J., Singh, J. P., & Endrass, J. (2013). Examining the predictive validity of the SORAG in Switzerland. *Sexual Offender Treatment*, *8*(2).
- Rossegger, A., Gerth, J., Urbaniok, F., Laubacher, A., & Endrass, J. (2010). The Sex Offender Risk Appraisal Guide (SORAG). Validity and authorised German translation. *Fortschritte der Neurologie-Psychiatrie*, *78*(11), 658-667. doi: 10.1055/s-0029-1245688
- Rufibach, K. (2010). Use of Brier score to assess binary predictions. *Journal of Clinical Epidemiology*(63), 938-942. doi: 10.1016/j.jclinepi.2009.11.009
- Sampson, R. J., & Laub, J. H. (2003). Life-course desisters? Trajectories of crime among delinquent boys followed to age 70. *Criminology*, *41*(3), 301-340. doi: 10.1111/j.1745-9125.2003.tb00997.x
- Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, *2*(2), 191-201.

- Schmid, C. H., & Griffith, J. L. (2005). Multivariate classification rules: Calibration and discrimination. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (2 ed., Vol. 5): John Wiley & Sons.
- Schweizer Fernsehen. (2011). Schiesserei in Pfäffikon: Täter gesteht [Shooting in Pfäffikon: Offender confesses]. Retrieved 22/10/2013, 2013, from <http://www.tagesschau.sf.tv/Nachrichten/Archiv/2011/08/16/Vermischtes/Schiesserei-in-Pfaeffikon-Taeter-gesteht>
- Seto, M. C., Harris, G. T., Rice, M. E., & Barbaree, H. E. (2004). The screening scale for pedophilic interests predicts recidivism among adult sex offenders with child victims. *Archives of Sexual Behavior, 33*, 455-466.
- Shortt, J. W., Capaldi, D. M., Kim, H. K., Kerr, D. C. R., Owen, L. D., & Feingold, A. (2012). Stability of intimate partner violence by men across 12 years in young adulthood: Effects of relationship transitions. *Prevention Science, 13*(4), 360-369. doi: 10.1007/s11121-011-0202-0
- Silver, J. M., & Yudofsky, S. C. (1991). The overt aggression scale: Overview and guiding principles. *The Journal of Neuropsychiatry and Clinical Neurosciences, 3*(2), 22-29.
- Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law, 31*(1), 8-22. doi: 10.1002/bsl.2052
- Singh, J. P., Desmarais, S., & Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioral Sciences and the Law, 31*, 55-73. doi: 10.1002/bsl.2053
- Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., . . . Otto, R. K. (2014). International perspectives on the practical application of

- 
- violence risk assessment: A global survey of 44 countries. *International Journal of Forensic Mental Health*, 13(3), 193-206. doi: 10.1080/14999013.2014.922141
- Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: a systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychological Review*, 31(3), 499-513. doi: 10.1016/j.cpr.2010.11.009
- Skeem, J. L., & Monahan, J. (2011). Current Directions in Violence Risk Assessment. *Public Law and Legal Theory Research Paper, No. 2011-13*, 1-16.
- Snowden, R. J., Gray, N. S., & Taylor, J. (2010). Risk assessment for future violence in individuals from an ethnic minority group. *International Journal of Forensic Mental Health*, 9, 118-123. doi: 10.1080/14999013.2010.501845
- Spiegelhalter, D. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5, 421-433.
- StataCorp. (2012). Stata statistical software: Release 12. College Station: TX: StataCorp LP.
- StataCorp. (2013). Stata statistical software: Release 13: College Station: TX: StataCorp LP.
- Steadman, H. J., & Cocozza, J. J. (1974). *Careers of the criminally insane - Excessive social control of deviance*. New York: Lexington Books.
- Steadman, H. J., & Cocozza, J. J. (1978). Psychiatry, dangerousness and the repetitively violent offender. *The Journal of Criminal Law & Criminology*, 69(2), 226-231.
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., & Gonen, M. (2010). Assessing the performance of prediction models - a framework for traditional and novel measures. *Epidemiology*, 21, 128-138. doi: 10.1097/EDE.0b013e3181c30fb2

- Stöckl, H., K., D., Rotstein, A., Abrahams, N., Campbell, J., Watts, C., & Garcia Moreno, C. (2013). The global prevalence of intimate partner homicide: A systematic review. *The Lancet*, 382(9895), 859-865. doi: 10.1016/S0140-6736(13)61030-2
- Storey, J. E., Kropp, P. R., Hart, S. D., Belfrage, H., & Strand, S. (2014). Assessment and management of risk for intimate partner violence by police officers using the brief spousal assault form for the evaluation of risk. *Criminal Justice and Behavior*, 41(2), 256-271. doi: 10.1177/0093854813503960
- Storey, J. E., Watt, K. A., Jackson, K. J., & Hart, S. D. (2012). Utilization and implications of the Static-99 in practice. *Sexual Abuse*, 24(3), 289-302. doi: 10.1177/1079063211423943
- Straus, M. A. (2009). Gender symmetry in partner violence: Evidence and implications for prevention and treatment. In D. J. Whitaker & J. R. Lutzker (Eds.), *Preventing partner violence: Research and evidence-based intervention strategies* (pp. 245-271). Washington, DC: American Psychological Association.
- Swan, S. C., Gambone, L. J., Caldwell, J. E., Sullivan, T. P., & Snow, D. L. (2008). A review of research on women's use of violence with male intimate partners. *Violence and Victims*, 23(3), 301-314. doi: 10.1891/0886-6708.23.3.301
- Swanson, J. W. (2008). Preventing the unpredicted: Managing violence risk in mental health care. *Psychiatric Services*, 59(2), 191-193. doi: 10.1176/appi.ps.59.2.191
- Swiss Federal Statistical Office. (2014). Polizeiliche Kriminalstatistik (PKS) - Jahresbericht 2013 [Police Crime Statistics - Annual report 2013] *Statistik der Schweiz*. Neuchâtel.
- Timmons Fritz, P. A., & Slep, A. M. S. (2009). Stability of physical and psychological adolescent dating aggression across time and partners. *Journal of Clinical Child & Adolescent Psychology*, 38(3), 303-314. doi: 10.1080/15374410902851671

- Tjaden, P., & Thoennes, N. (2000). Prevalence and consequences of male-to-female and female-to-male intimate partner violence as measured by the National Violence Against Women Survey. *Violence Against Women, 6*(2), 142-161.
- Tutty, L., Wyllie, K., Abbott, P., Mackenzie, J., Ursel, E. J., & Koshan, J. M. (2008). The justice response to domestic violence: A literature review. Canada.
- Urbaniok, F. (2007). *FOTRES: Forensisches Operationalisiertes Therapie-Risiko-Evaluations-System*. Bern: Zytglogge
- Urbaniok, F., Endrass, J., Rossegger, A., Noll, T., Gallo, W. T., & Angst, J. (2007). The prediction of criminal recidivism: The implication of sampling in prognostic models. *European archives of psychiatry and clinical neuroscience, 257*(3), 129-134. doi: 10.1007/s00406-006-0678-y
- Urbaniok, F., Rossegger, A., Steinfeld, O., & Endrass, J. (2006). Drohungen als Vorboten schwerer Gewalt. *Fortschritte der Neurologie Psychiatrie, 74*(6), 337-345. doi: 10.1055/s-2005-915574
- van den Heuvel, C., Alison, L., & Power, N. (2014). Coping with uncertainty: Police strategies for resilient decision-making and action implementation. *Cognition, Technology & Work, 16*(1), 25-45. doi: 10.1007/s10111-012-0241-8
- Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C., & Habbema, J. D. F. (2005). Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology, 58*, 475-483. doi: 10.1016/j.jclinepi.2004.06.017
- Viljoen, J. L., McLachlan, K., & Vincent, G. M. (2010). Assessing violence risk and psychopathy in juvenile and adult offenders: A survey of clinical practices. *Assessment, 17*, 377-395. doi: 10.1177/1073191109359587

- von Franqué, F. (2013). Strukturierte, professionelle Risikobeurteilungen. In M. Rettenberger & F. von Franqué (Eds.), *Handbuch psychologischer Instrumente zur Kriminalprognose* (pp. 357-380). Göttingen: Hogrefe.
- Vossekuil, B., Reddy, M., & Fein, R. (2000). U.S.S.S. Safe school initiative: An interim report on the prevention of targeted violence in schools. Washington: United States Secret Service National Threat Assessment Center.
- Walker, K., Bowen, E., & Brown, S. (2013). Desistance from intimate partner violence: A critical review. *Aggression and Violent Behavior, 18*(2), 271-280. doi: 10.1016/j.avb.2012.11.019
- Warren, L. J., Mullen, P. E., & Ogloff, J. R. P. (2011). A clinical study of those who utter threats to kill. *Behavioral Sciences & the Law, 29*(2), 141-154. doi: 10.1002/bsl.974
- Warren, L. J., Mullen, P. E., Thomas, S. D. M., Ogloff, J. R. P., & Burgess, P. M. (2008). Threats to kill: A follow-up study. *Psychological Medicine, 38*(4), 599-605. doi: 10.1017/S003329170700181X
- Waypoint Centre for Mental Health Care. (2012). Retrieved July 2012, from [http://www.mhpc.on.ca/Site\\_Published/internet/SiteContent.aspx?Body.QueryId.Id=1673&LeftNavigation.QueryId.Categories=62](http://www.mhpc.on.ca/Site_Published/internet/SiteContent.aspx?Body.QueryId.Id=1673&LeftNavigation.QueryId.Categories=62)
- Waypoint Centre for Mental Health Care. (2014). ODARA 101 Bibliography. Retrieved March 2014, from [http://www.waypointcentre.ca/UserFiles/Servers/Server\\_9960/File/Research/Odara/Bibliography v3.pdf](http://www.waypointcentre.ca/UserFiles/Servers/Server_9960/File/Research/Odara/Bibliography v3.pdf)
- Webster, C. D., Eaves, D., Douglas, K., & Wintrup, A. (1995). *The HCR-20 scheme: The assessment of dangerousness and risk* (1 ed.): Simon Fraser University and Forensic Psychiatric Services Commission of British Columbia.

- Weller, M., Hope, L., & Sheridan, L. (2012). Police and public perceptions of stalking: the role of prior victim–offender relationship. *Journal of Interpersonal Violence, 28*(2), 320-339. doi: 10.1177/0886260512454718
- Wenk, E. A., Robison, J. O., & Smith, G. W. (1972). Can violence be predicted? *Crime & Delinquency, 18*(4), 393-402. doi: 10.1177/001112877201800408
- Whitaker, D. J., Le, B., & Niolon, P. H. (2010). Persistence and desistance of the perpetration of physical aggression across relationships: Findings from a national study of adolescents. *Journal of Interpersonal Violence, 25*(4), 591-609. doi: 10.1177/0886260509334402
- Williams, K. R., & Houghton, A. B. (2004). Assessing the risk of domestic violence re-offending: A validation study. *Law and Human Behavior, 28*, 437-455.
- World Health Organisation. (2013). Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence. Geneva: Department of reproductive health and research.
- Zoder, I. (2008). Tötungsdelikte in der Partnerschaft [Homicides in intimate relationships] *Statistik der Schweiz*. Neuchâtel.