The Predictive Validity of the Historical Clinical Risk Management – 20 Version 3 (HCR-20\textsuperscript{V3}) and the Violence Risk Appraisal Guide - Revised (VRAG-R)

Ms. Delene M. Brookstein
BAppSC (Psych) (Hons) Deakin University, 2011

under the supervision of

Professor Michael Daffern and Professor James Ogloff

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Psychology (Clinical and Forensic Psychology)
Centre for Forensic Behavioural Science
Swinburne University of Technology

© Delene M. Brookstein 2016
All rights reserved. This work may not be reproduced in whole or in part, by photocopy or other means, without written permission of the author.
DECLARATION

I, the undersigned, thereby declare that this thesis,

1. comprises of original work and contains no material which has been accepted for the award to the candidate of any other degree or diploma at any university or equivalent institution;

2. to the best of the candidate’s knowledge contains no material previously published or written by another person except where due reference is made in the text of the examinable outcome; and

3. where the work is based on joint research or publications, discloses the relative contributions of the respective workers or authors.

Signature

Ms. Delene M. Brookstein
# TABLE OF CONTENTS

List of Tables ix
List of Figures xiii
List of Appendices xv
Abbreviations xvi
Acknowledgements xix
Abstract xxi

## Chapter One

*Introduction to Violence Risk Assessment* 1

1.1 Introduction 1
1.2 The Historical Context: Dangerousness and Prediction 1
1.3 Defining Violence Risk Assessment 3
1.4 Three Generations of Violence Risk Assessment: A Short Course in the History and Development of Approaches to Assessing Violence Risk 7
   1.4.1 First Generation: Unstructured Clinical Judgment 7
      1.4.1.1 Base Rates 8
      1.4.1.2 Transparency of Decision-Making 11
      1.4.1.3 Illusory Correlations 11
      1.4.1.4 The Link between Violence and Mental Disorder 12
      1.4.1.5 Naturalistic Experiments 16
      1.4.1.6 Empirical Research 18
      1.4.1.7 Conclusion 22
   1.4.2 Second Generation: Actuarial Assessment 22
      1.4.2.1 Defining Actuarial Risk Assessment 23
      1.4.2.2 Empirical Research 24
      1.4.2.3 The Actuarial Advantage 26
      1.4.2.4 Criticisms of the Actuarial Approach 27
      1.4.2.5 Broken Legs: Rare Events impinging upon Risk Assessments 30
1.4.2.6 Understanding the Superiority of the Actuarial Approach
1.4.2.7 Resistance to the Actuarial Movement
1.4.2.8 Conclusion
1.4.3 Third Generation: Structured Professional Judgment
1.4.3.1 Advantages of the Structured Professional Judgment Approach
1.4.3.2 Clinician Preferences in Risk Communication
1.4.3.3 Disadvantages of the Structured Professional Judgment Approach
1.4.4 Statistical Advances in Violence Risk Assessment
1.5 The Victorian Context
1.6 Conclusion

Chapter Two
Development and Revision of the Historical Clinical Risk Management – 20 (HCR-20) and the Violence Risk Appraisal Guide (VRAG)

2.1 Introduction
2.2 The Historical Clinical Risk Management-20 (HCR-20)
  2.2.1 The Original HCR-20
  2.2.2 The Historical Clinical Risk Management-20 Version 2 (HCR-20^v2)
  2.2.3 The Historical Clinical Risk Management-20 Version 3 (HCR-20^v3)
2.3 HCR-20^v3 Key Changes
  2.3.1 Changes to Risk Factor Labels
  2.3.2 Changes to Content
  2.3.3 Sub-Items for Complex Risk Factors
  2.3.4 Inclusion of Relevance Ratings
  2.3.5 Inclusion of Item Indicators
  2.3.6 Changes to the Rating System
  2.3.7 Exclusion of PCL-R Psychopathy Checklist - Revised (PCL-R)
2.4 HCR-20 Summary

2.5 The Violence Risk Assessment Guide (VRAG)
   2.5.1 The Violence Risk Appraisal Guide – Revised (VRAG-R)

2.6 VRAG-R Key Changes

2.7 VRAG Conclusion

2.8 Contrasts between the HCR-20 and VRAG
   2.8.1 Applicability
   2.8.2 Subjectivity
   2.8.3 Changes in Risk
   2.8.4 Idiosyncratic Assessment

2.9 Chapter Conclusion

Chapter Three

Psychometric Properties of the HCR-20 and VRAG

3.1 Introduction

3.2 Methodological Issues in Evaluating Violence Risk Assessment Tools

3.3 Literature Search

3.4 Structured Professional Judgment Assessments: The Historical Clinical Risk Management – 20 Literature Review
   3.4.1 Predictive Validity
      3.4.1.1 Research in Forensic Psychiatric Populations
      3.4.1.2 Civil Psychiatric Populations and Mixed Samples
      3.4.1.3 Australian Based Research
      3.4.1.4 The Historical Clinical Risk Management - 20 Version 3
      3.4.1.5 Conclusion

3.4.2 Strength of Scales
      3.4.2.1 Conclusion

3.4.3 Inter-rater Reliability
      3.4.3.1 Conclusion

3.4.4 Concurrent Validity
      3.4.4.1 Conclusion

3.4.5 Incremental Validity
      3.4.5.1 Conclusion
3.5 Actuarial Assessments: The Violence Risk Appraisal Guide

<table>
<thead>
<tr>
<th>Literature Review</th>
<th>144</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5.1 Predictive Validity</td>
<td>146</td>
</tr>
<tr>
<td>3.5.1.1 Forensic Psychiatric Populations</td>
<td>148</td>
</tr>
<tr>
<td>3.5.2 Inter-rater reliability</td>
<td>152</td>
</tr>
<tr>
<td>3.5.3 Use of the VRAG in Australia</td>
<td>153</td>
</tr>
<tr>
<td>3.5.4 Violence Risk Appraisal Guide – Revised (VRAG-R)</td>
<td>154</td>
</tr>
</tbody>
</table>

3.6 Conclusion | 156

3.7 Comparisons between the HCR–20 and VRAG | 157

3.8 Conclusion | 170

**Chapter Four**

*Research Methodology*

4.1 Introduction | 173

4.2 Research Setting | 173

4.2.1 The Victorian Institute of Forensic Mental Health | 173

4.2.2 Legislative Framework | 175

4.3 Research Overview | 176

4.3.1 Research Design | 176

4.3.2 Research Aims | 178

4.3.3 Hypotheses | 179

4.4 Sources of Data | 181

4.5 Ethics | 181

4.5.1 Ethical considerations | 182

4.6 Accessing the data | 183

4.6.1 Archival Data | 183

4.6.2 Victoria Police Data | 183

4.6.3 Department of Justice Data and Regulation | 184

4.6.4 Comparative Data (Sourcing the Sample) | 184

4.6.5 Refining the Sample | 185

4.6.6 Completing the Datasets | 187

4.7 Procedure | 189

4.7.1 Data Linkage | 189

4.7.2 Avoiding Bias | 190
### 4.8 Definitions

- **4.8.1 Violence**
- **4.8.2 Violent Recidivism**
- **4.8.3 General Recidivism**
- **4.8.4 Recidivism Related Data**

### 4.9 Offence Hierarchy: Coding Violence Severity

### 4.10 Coding Offending Versatility

### 4.11 Administration of Violence Risk Assessment Tools

### 4.12 User Qualifications

### 4.13 HCR-20\(^{V3}\) Administration

- **4.13.1 Presence Ratings**
- **4.13.2 Relevance Ratings**
- **4.13.3 Summary Risk Ratings**
- **4.13.4 Risk Formulation**
- **4.13.5 Missing HCR-20\(^{V3}\) Scores and Omitting Participants**
  - **4.13.5.1 Prorating HCR-20\(^{V3}\) Scores**
- **4.13.6 HCR-20\(^{V3}\) Coding Timeframe**

### 4.14 VRAG-R Administration

- **4.14.1 The Cormier Lang System for Quantifying Criminal History**
- **4.14.3 Missing VRAG-R scores and Omitting Participants**
- **4.14.4 Prorating VRAG-R Scores**

### 4.15 Reliability Scoring

### 4.16 Calculating Time at Risk

### 4.17 Sample Demographics

- **4.17.1 Age, Gender, Ethnicity and Martial Status**
- **4.17.2 Legal Status, Unit Location and Length of Stay**
- **4.17.3 Education and Work History**
- **4.17.4 Previous Hospitalisation**
- **4.17.5 Suicide and Self-harm**
- **4.17.6 Substance Use**
- **4.17.7 Psychiatric Diagnoses (Admission)**
- **4.17.8 Psychiatric Diagnoses (Discharge)**
- **4.17.9 Institutional Behaviour and Use of Seclusion**
- **4.17.10 Forensic History**
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.17.11 Index Offence(s)</td>
<td>221</td>
</tr>
<tr>
<td>4.17.12 Offending in the Follow-up Period</td>
<td>222</td>
</tr>
<tr>
<td>4.17.13 Discharge Pathways and Legal Orders</td>
<td>223</td>
</tr>
<tr>
<td>4.18 Representativeness of the Sample</td>
<td>224</td>
</tr>
<tr>
<td>4.19 Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE)</td>
<td>226</td>
</tr>
</tbody>
</table>

Chapter Five

**Statistical Methods**

5.1 Introduction                                                        227
5.2 Hypotheses and Statistical Analyses                                 227
5.3 Data Cleaning                                                      230
5.4 Missing Data                                                       230
5.5 Statistical Assumptions                                            231
5.6 Statistical Analyses                                               232
  5.6.1 Predictive Validity                                            232
    5.6.1.1 The Receiver Operating Characteristic (ROC)                 233
    5.6.1.2 The ROC curve                                               234
    5.6.1.3 The Area Under Curve (AUC)                                  236
    5.6.1.4 Other Relevant Discriminatory Indices                       238
    5.6.1.5 Comparing AUC’s                                             240
  5.6.2 Correlational Analyses - Kendall's Tau                          244
  5.6.3 Chi-Square Test of Independence                                 245
  5.6.4 Binomial Logistic Regression                                    246
  5.6.5 Odds Ratios                                                    249
  5.6.6 Survival Analysis                                              250
  5.6.7 Intraclass Correlation Coefficient                             252

Chapter Six

**Results**

6.0 Introduction                                                       255
6.1 Hypothesis 1: Association                                          255
6.2 Hypothesis 2: Concurrent Validity 260
6.3 Hypothesis 3: Prediction 263
   6.3.1 HCR-20$^V_3$ 263
   6.3.2 VRAG-R 267
   6.3.3 HCR-20$^V_2$ and HCR-20$^V_3$ 270
   6.3.4 VRAG and VRAG-R 272
6.4 Hypothesis 4: HCR-20$^V_3$ Predictive Strength of Scales 278
   6.4.1 AUC’s for HCR-20$^V_3$ Scales 279
   6.4.2 Violent Recidivism 280
   6.4.3 Controlling for Time at Risk 282
   6.4.4 Addition of HCR-20$^V_3$ Total Score Variable 284
   6.4.5 Interactions Between HCR-20$^V_3$ Scales 286
6.5 Hypothesis 5: Discrimination 288
   6.5.1 HCR-20$^V_3$ 289
   6.5.2 VRAG-R 290
6.6 Hypothesis 6: Survival Time 291
   6.6.1 HCR-20$^V_3$ 291
   6.6.2 VRAG-R 297
6.7 Hypothesis 7: Incremental Validity 303
   6.7.1 Incremental Validity Within the HCR-20$^V_3$ 303
   6.7.2 Incremental Validity Between the HCR-20$^V_3$, VRAG-R and PCL-R 308
      6.7.2.1 HCR-20$^V_3$ Total Scores and PCL-R Comparison 309
      6.7.2.2 HCR-20$^V_3$ SRR’s and PCL-R Comparison 310
      6.7.2.3 VRAG-R Total Scores and PCL-R Comparison 311
      6.7.2.4 HCR-20$^V_3$ and VRAG-R Comparison 312
6.8 Hypothesis 8: Inter-Rater Reliability 315

Chapter Seven
Discussion 317

7.1 Overview 317
7.2 Scoring Description 318
7.3 Main Findings 318
   7.3.1 Associations with Violence 318
7.3.2 Concurrent Validity 320
7.3.3 Predictive Validity 321
7.3.4 HCR-20\textsuperscript{V3} Predictive Strength of Scales 324
7.3.5 Discrimination 326
7.3.6 Survival Time 327
7.3.7 Incremental Validity 329
  7.3.7.1 Within Tool Incremental Validity 329
  7.3.7.2 Between Tool Incremental Validity 331
7.3.8 Inter-rater Reliability 332
7.4 Findings Overview and Clinical Implications 333
7.5 Limitations 338
7.6 Future research 342
7.7 Conclusion 344

References 346
Appendices 391
LIST OF TABLES

Chapter Two

Table 1: Violence Risk Factors on the Original HCR-20 and the HCR-20\(^{V2}\) across Historical, Clinical and Risk-Management Domains

Table 2: Comparison of Items on the HCR-20\(^{V2}\) to HCR-20\(^{V3}\) Across the Three Scales

Table 3: Risk Factors and Sub-Items on the HCR-20\(^{V3}\)

Table 4: Items on the Violence Risk Appraisal Guide (VRAG), including Pearson Correlations between Variables and Violent Recidivism

Table 5: VRAG Bin Distribution and Probability of Violent Recidivism at 7-Year and 10-Year Follow-Up Periods

Table 6: Items on the Violence Risk Appraisal Guide - Revised (VRAG-R) including Pearson Correlations between Variables and Violent Recidivism

Table 7: Rates of Violent Recidivism in each VRAG-R Category after 5-Years and 12-Years Fixed Periods of Opportunity

Table 8: Comparison of Items on the VRAG and VRAG-R

Chapter Three

Table 9: Odds of Violence based on four Categories of Violent Outcomes and HCR-20 Scores derived from Douglas, Ogloff, Nicholls and Grant (1999)

Table 10: AUC’s for HCR-20\(^{V3}\) Presence Ratings, Relevance ratings and SRR’s in short-term and long-term follow-ups derived from Strub, Douglas and Nicholls (2014)
Table 11: Summary of HCR-20, HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} Predictive Validity Studies

Table 12: AUC’s for General Recidivism and Violent Recidivism based on HCR-20\textsuperscript{V2} Scale Scores derived from Gray, Taylor and Snowden (2008)

Table 13: AUC’s for Violent Outcomes based on HCR-20 Scale Scores derived from Douglas, Ogloff, Nicholls and Grant (1999)

Table 14: Summary of Studies on the HCR-20’s Scale-level Predictive Validity

Table 15: Intraclass Correlation Coefficients for the HCR-20\textsuperscript{V3} derived from Douglas and Belfrage (2014)

Table 16: Summary of Inter-rater Reliability Outcomes for the HCR-20

Table 17: Concurrent Validity between the HCR-20 and other Comparative Violence Risk Assessment Instruments

Table 18: AUC’s for the VRAG at fixed follow-up points derived from Snowden, Gray, Taylor and MacCulloch (2007)

Table 19: Predictive Validity AUC’s for Any In-patient Aggression across Short-term and Medium-term derived from Chu, Thomas, Ogloff and Daffern (2011)

Table 20: Studies comparing the Predictive Validity of the HCR-20\textsuperscript{V2}, VRAG and PCL-R/SV

Chapter Four

Table 21: Reasons for Exclusion from the Sample

Table 22: Violent Offence Categories with Examples

Table 23: Non-violent Offence Categories with Examples
Chapter Six

**Table 24:** Kendall’s Tau Correlations between Violence Risk Assessment Tool Scores and Violent Recidivism

**Table 25:** Cross-tabulation of HCR-20\textsuperscript{V3} SRR scores and Post-Release Violence

**Table 26:** Cross-tabulation of VRAG-R Bins and Number of Participants Engaging in Post-Release Violence

**Table 27:** Kendall’s Tau Correlations between Total Scores on Violence Risk Assessment Tools (HCR-20\textsuperscript{V2}, HCR-20\textsuperscript{V3}, VRAG and VRAG-R).

**Table 28:** Pearson’s Correlations between Total Scores on Violence Risk Assessment Tools (HCR-20\textsuperscript{V2}, HCR-20\textsuperscript{V3}, VRAG and VRAG-R)

**Table 29:** Pearson’s Correlations between HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} Scales.

**Table 30:** Indices of Discrimination for HCR-20\textsuperscript{V3} SRR as Cut-Off

**Table 31:** Indices of Discrimination for VRAG-R Bin Numbers as Cut-off

**Table 32:** Summary of the Predictive Performance of the HCR-20\textsuperscript{V3} and VRAG-R with Violent Recidivism as Outcome

**Table 33:** Indices of Discrimination for the HCR-20\textsuperscript{V2} and the HCR-20\textsuperscript{V3} with Violent Recidivism as Outcome

**Table 34:** Summary of the Predictive Performance of the VRAG and VRAG-R with Violent Recidivism as Outcome

**Table 35:** AUC’s for the H, C and R Scales of the HCR-20\textsuperscript{V3} and Violent Recidivism

**Table 36:** Sequential Logistic Regression Predicting Likelihood of a Violent Offence Charge Post-Release based on H-Scale, C-Scale and R-Scale Prorated Scores
Table 37: Sequential Logistic Regression Predicting Likelihood of a Violent Offence Charge Post-Release based on Time at Risk, H-Scale, C-Scale and R-Scale Prorated Scores

Table 38: Sequential Logistic Regression Predicting Likelihood of a Violent Offence Charge Post-Release based on Time at Risk, HCR-20V3 Scales and HCR-20V3 Total Scores

Table 39: Sequential Logistic Regression Predicting Likelihood of a Violent Offence Charge Post-Release based on Time at Risk, HCR-20V3 Prorated Scale Scores, and Interactions amongst HCR-20V3 Scales

Table 40: Sequential Logistic Regression Predicting Likelihood of Post-Release General Recidivism based on the HCR-20V3 Presence and Relevance Ratings

Table 41: Inter-rater Reliability of the HCR-20V3 and VRAG-R
LIST OF FIGURES

Figure 1: Dataset Composition

Figure 2: Graphical Elements of the ROC Curve and AUC Index

Figure 3: Clustered Bar Chart displaying Violent Recidivism across VRAG-R Bins

Figure 4: ROC Curves for the HCR-20$^{V3}$ Total Scores and SRR’s with Violent Recidivism as Outcome.

Figure 5: ROC Curve for the HCR-20$^{V3}$ Relevance Ratings with Violent Recidivism as Outcome

Figure 6: ROC Curves for the VRAG-R Prorated Total Scores and Bins with Violent Recidivism as Outcome

Figure 7: ROC Curves for the HCR-20$^{V2}$ and the HCR-20$^{V3}$ with Violent Recidivism as Outcome

Figure 8: ROC Curves for the VRAG and VRAG-R (Total Scores and Bins) with Violent Recidivism as Outcome

Figure 9: VRAG and VRAG-R Total Score Comparisons

Figure 10: VRAG and VRAG-R Bin Number Comparisons

Figure 11: AUC’s for HCR-20$^{V3}$ Scales

Figure 12: Kaplan-Meier Survival Curves for HCR-20$^{V3}$ Total Prorated Score Median Splits and Time to First Offence

Figure 13: Kaplan-Meier Survival Curves for HCR-20$^{V3}$ SRR’s and Time to First Offence

Figure 14: Kaplan-Meier Survival Curves for HCR-20$^{V3}$ Total Prorated Score Median Splits and Time to First Violent Offence

Figure 15: Kaplan-Meier Survival Curves for HCR-20$^{V3}$ SRR’s and Time to First Violent Offence
Figure 16: Kaplan-Meier Survival Curves for VRAG-R Total Prorated Score
Median Splits and Time to First Offence

Figure 17: Kaplan-Meier Survival Curves for VRAG-R Grouped Bins and
Time to First Offence

Figure 18: Kaplan-Meier Survival Curves for VRAG-R Total Prorated Score
Median Splits and Time to First Violent Offence

Figure 19: Kaplan-Meier Survival Curves for VRAG-R Grouped Bins and
Time to First Violent Offence
LIST OF APPENDICES

Chapter One

Appendix A: Summary of Three Generations of Violence Risk Assessment

Chapter Two

Appendix B: Violence Risk Factors and Coding Scheme for the HCR-20 (Version 2)

Appendix C: Violence Risk Factors and Sub-Items on the HCR-20 (Version 3)

Appendix D: Violence Risk Appraisal Guide (VRAG) scoring rubric

Appendix E: Violence Risk Appraisal Guide – Revised (VRAG-R) scoring rubric

Chapter Four

Appendix F: Ethics Approval Notices and Letters of Support

Appendix G: Offence Categories for Criminal Versatility

Appendix H: The Cormier-Lang System for Quantifying Criminal History

Appendix I: The PCL-R Facet 4 (Antisociality) scoring rubric

Appendix J: Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE) Checklist

Chapter Five

### ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% CI</td>
<td>95% Confidence Interval</td>
</tr>
<tr>
<td>APB</td>
<td>Victorian Adult Parole Board</td>
</tr>
<tr>
<td>ATSI</td>
<td>Aboriginal and Torres Strait Islanders</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under Curve</td>
</tr>
<tr>
<td>C-Scale</td>
<td>Clinical Scale</td>
</tr>
<tr>
<td>CALD</td>
<td>Culturally and Linguistically Diverse</td>
</tr>
<tr>
<td>CBO</td>
<td>Community Based Order</td>
</tr>
<tr>
<td>CFBS</td>
<td>Centre for Forensic Behavioural Science</td>
</tr>
<tr>
<td>CMI</td>
<td>Client Management Interface</td>
</tr>
<tr>
<td>CMIA</td>
<td>Crimes (Mental Impairment and Unfitness to be Tried) Act 1997 (Vic)</td>
</tr>
<tr>
<td>CSO</td>
<td>Custodial Supervision Order</td>
</tr>
<tr>
<td>DSM-III</td>
<td>Diagnostic and Statistical Manual of Mental Disorders – Third Edition</td>
</tr>
<tr>
<td>DSM-V</td>
<td>Diagnostic and Statistical Manual of Mental Disorders – Fifth Edition</td>
</tr>
<tr>
<td>DV</td>
<td>Dependent Variable</td>
</tr>
<tr>
<td>ESB</td>
<td>English Speaking Background</td>
</tr>
<tr>
<td>FET</td>
<td>Fisher’s Exact Test</td>
</tr>
<tr>
<td>H-Scale</td>
<td>Historical Scale</td>
</tr>
<tr>
<td>HCR-20</td>
<td>Historical Clinical Risk Management – 20</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V2}</td>
<td>Historical Clinical Risk Management – 20 (Version 2)</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V3}</td>
<td>Historical Clinical Risk Management – 20 (Version 3)</td>
</tr>
<tr>
<td>HoNOS</td>
<td>Health of the Nation Outcome Scales</td>
</tr>
<tr>
<td>IJFMHS</td>
<td>International Journal of Forensic Mental Health Services</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>ICC</td>
<td>Intraclass Correlation Coefficient</td>
</tr>
<tr>
<td>ICO</td>
<td>Intensive Correction Order</td>
</tr>
<tr>
<td>IV</td>
<td>Independent Variable</td>
</tr>
<tr>
<td>K-S Test</td>
<td>Kolmogorov-Smirnov Test</td>
</tr>
<tr>
<td>LEAP</td>
<td>Law Enforcement Assistance Program</td>
</tr>
<tr>
<td>LSD</td>
<td>Lysergic acid diethylamide</td>
</tr>
<tr>
<td>MDO</td>
<td>Mentally Disordered Offender</td>
</tr>
<tr>
<td>NESB</td>
<td>Non-English speaking background</td>
</tr>
<tr>
<td>NOS</td>
<td>Not Otherwise Specified</td>
</tr>
<tr>
<td>NPV</td>
<td>Negative Predictive Value</td>
</tr>
<tr>
<td>NGRMI</td>
<td>Not Guilty by Reason of Mental Impairment</td>
</tr>
<tr>
<td>OCD</td>
<td>Obsessive-Compulsive Disorder</td>
</tr>
<tr>
<td>OR</td>
<td>Odds Ratio</td>
</tr>
<tr>
<td>PCL-R</td>
<td>Psychopathy Checklist – Revised</td>
</tr>
<tr>
<td>PCL:SV</td>
<td>Psychopathy Checklist: Screening Version</td>
</tr>
<tr>
<td>PPV</td>
<td>Positive Predictive Value</td>
</tr>
<tr>
<td>R-Scale</td>
<td>Risk Management Scale</td>
</tr>
<tr>
<td>RAGEE</td>
<td>Risk Assessment Guidelines for the Evaluation of Efficacy</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>SORAG</td>
<td>Sex Offender Risk Appraisal Guide</td>
</tr>
<tr>
<td>SPJ</td>
<td>Structured Professional Judgment</td>
</tr>
<tr>
<td>SRR</td>
<td>Summary Risk Rating</td>
</tr>
<tr>
<td>TEH</td>
<td>Thomas Embling Hospital</td>
</tr>
<tr>
<td>VIFMH</td>
<td>Victorian Institute of Forensic Mental Health</td>
</tr>
<tr>
<td>VPS</td>
<td>Violence Prediction Scheme</td>
</tr>
<tr>
<td>VRAG/-R</td>
<td>Violence Risk Appraisal Guide/ -Revised</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to thank Professor Michael Daffern, who has encouraged me along every step of this journey. I cannot thank you enough for your leadership, input and professionalism. I could not have asked for a better supervision experience, nor a greater mentor to guide me through my training. Thank you for your steadfast support. Thank you also to Professor James Ogloff who provided additional encouragement and guidance when most needed. I would also like to thank Dr. Chi Meng Chu, Dr. Rachel Campbell, Cierran Harries and Dr. Bronwyn McKeon for their collaboration, and Professor Denny Meyer and Dr. Dan Shea for their statistical assistance.

The DPsych was a daunting experience that produced many trials, and in doing so, further developed my perseverance, patience and courage. I could not have endured this without the support of my fellow classmates: Marie Henshaw, Innes Seric, Mingwai Tam, Rana Abou-Sinna, Svenja Senkans, Benjamin Spivak, Nina Stephenson, Carmel Kodituwakku, Michael Ganci and Shireen Mahtani. Thanks also to Cindy Dickenson, Lillian De Bortolli, Stephane Shepherd and Margaret Nixon. Thank you to my overseas family who made their love and support known across the oceans.

Finally, I would like to thank my partner Jude Perera, one of my greatest motivations was thinking about post Dpsych life with you. Thank you to my siblings, Denaé and Dwayne, for their unwavering support. Mum, thank you for all your sacrifices. Your strength and tenacity is reflected in the achievements of your children. Dad, your trials and tribulations bore good fruit. I would like to dedicate this work to my grandparents, Maud and Jack Laing, who always told me that my cup would overflow until I drank from my saucer. Thank God, you were right.
Abstract

The field of violence risk assessment has progressed across three generations of thought, from unstructured clinical judgment to actuarial and structured professional judgment approaches. Two violence risk assessment tools, the Violence Risk Appraisal Guide (VRAG) and Historical Clinical Risk Management – 20 (HCR-20) are among the most commonly used tools for evaluating violence risk within the actuarial and structured professional judgment approaches respectively. These tools boast a large international evidence base in their ability to predict future violence. However, limited research has stemmed from Australian cohorts. Recently, both tools have been revised to reflect contemporary developments and best practice. Limited preliminary evidence exists for these revised tools (the VRAG - Revised and HCR-20, Version 3), despite their implementation in practice. The aim of the current study was to evaluate the predictive, concurrent and incremental validity and inter-rater reliability of these tools within a sample of 100 Australian forensic psychiatric patients. The study was pseudo-prospective in nature. Participants were followed-up for approximately 13 years post-discharge from the Thomas Embling Hospital, located in Melbourne, Victoria. Recidivism data were sourced from official Victoria Police records. Results indicated that the tools significantly predicted general and violent recidivism, to similar levels of accuracy observed in earlier versions. They demonstrate concurrent validity with previous versions, and good to excellent inter-rater reliability. To our knowledge, this represents the first evaluation of the HCR-20v3 and VRAG-R in Australia, adding to a limited preliminary body of research. Future research, and implications for clinicians and the field of violence risk assessment more broadly are discussed.
Chapter One
Introduction to Violence Risk Assessment

1.1 Introduction

The purpose of this chapter is to introduce the reader to the area of violence risk assessment by reviewing historical underpinnings of dangerousness and prediction, and defining violence risk assessment. The three generations of violence risk assessment and associated empirical research are briefly overviewed. The advantages, drawbacks and notable criticisms of each approach are highlighted. This provides the theoretical foundation for the research, and is an important precursor to contextualising the violence risk assessment tools evaluated in this thesis. The chapter concludes with a discussion on violence risk assessment within the Victorian context.

1.2 The Historical Context: Dangerousness and Prediction

The prediction of violence is considered to be “one of the most complex and controversial issues in behavioral science and law” (Borum, 1996, p. 945). Violence risk assessment is a prominent clinical activity, applicable to many criminal proceedings including granting of bail, sentencing, parole and conditions of release from custody (Mathiesen, 1998; Ogloff & Davis, 2005). In contemporary times, it remains a contentious issue at the intersection of mental health and criminal justice (Mullen & Ogloff, 2009). Before violence risk, the concept of harm to others was captured under the notion of ‘dangerousness’. Dangerousness began to gain prominence in 1960’s, where it was considered in civil law proceedings as a
criterion for the involuntary hospitalisation of mentally ill persons (Appelbaum, 1997; Mestrovic & Cook, 1986; Monahan, 1988). For example, the American Psychiatric Association’s (1983) guidelines for psychiatric hospitalisation included the involuntary detainment of mentally disordered persons who were considered “likely to cause harm to others” (p. 672). Brooks (1978) noted the vagueness of definitions of dangerousness in civil commitment statutes, but noted that the key focus was on persons considered “likely to injure or harm others, or himself” (p. 39). Similar definitions have been echoed by others who highlight the key component of exhibiting behaviour that is considered harmful to others (Monahan, 1988). Green, Schneider, Griswold, Belfi, Herrera and DeBlasi (2016) have drawn attention to the conceptualisation of dangerousness as “risk of physical harm to self or others as a result of mental illness” (p. 51, emphasis added).

Since the 1960’s, the concept of dangerousness has remained a central component of mental health practice and law (McNeil, 1998; Monahan 1988). Historical understandings of predicting dangerousness began to change conceptually in the 1980’s - 1990’s, resulting in what is now known as violence risk assessment (Mossman, 2013). Danger-based commitment laws changed significantly from the 1980’s onwards in America, Australia, Canada and other parts of the world (Appelbaum, 1997). This was a reflection of changes in theoretical and conceptual changes within the field, from ‘predicting dangerousness’ to ‘assessing risk for violence’ (Mossman, 2013; Steadman, 2000). Violence risk assessment differs from risk prediction in that, whilst prediction refers to a definitive binary statement about a future behaviour (yes or no), risk assessment refers to a probabilistic judgment, a state of potential (Jackson & Guyton, 2007). This important conceptual
nuance acknowledges the fluidity of violence and the need for ongoing assessment of the likelihood of violence (Hart, 1998). As an all encapsulating term, dangerousness has effectively been divided into separate considerations in violence risk assessment: risk factors, risk level, and the amount and type of harm (Ogloff et al., 2005).

1.3 Defining Violence Risk Assessment

In their article ‘Coming to terms with the terms of risk’, Kraemer, Kazdin, Offord, Kessler, Jensen and Kupfer (1997) defined risk as the “probability of an outcome within a population of subjects” (p. 337). Violence is defined as “actual, attempted, or threatened infliction of bodily harm on another person” (Douglas, Hart, Webster, & Belfrage, 2013; p. 36). In contemporary times, harm has also been extended to include serious psychological harm (Douglas et al., 2013). Hart (1998) operationally defined violence risk assessment as “the process of evaluating individuals to (1) characterise the likelihood they will commit acts of violence and (2) develop interventions to manage or reduce that likelihood” (p. 122). Hart’s definition touches on three important elements of violence risk assessment: 1) it is an ongoing process that involves continuing evaluation; 2) that specifies a likelihood; and 3) informs risk management (Douglas, Cox, & Webster, 1999).

The specification of likelihood is a basic requirement, and contemporary violence risk assessment has seen the emergence of more detailed commentary on: 1) the nature of the violence; 2) the types of violence that may occur; 3) the severity of the violence; 4) the frequency of the violence; 5) the imminence of violence; and finally 6) the likelihood or probability that violence will occur (Roesch, Zapf, & Hart, 2010). Therefore, a
richer level of information on the features of violence perpetration is sometimes provided regarding a particular type of violence, within a particular timeframe under stated conditions (Roesch et al., 2010).

Risk can be reported in terms of status or state (Douglas & Skeem, 2005). Risk status refers to a baseline level of risk that is based on static risk factors and relevance to others (i.e., interindividual), whereas risk state refers to variable risk that is based on dynamic risk factors and relevance to the person (i.e., intrapersonal). Therefore, risk state is amenable to change, whereas risk status is not, at least in the short- to medium-term (Dvoskin & Heilbrun, 2001). While risk status does provide insight for clinicians about the long-term level of risk posed and level of intervention required, as a fixed indicator, it has limited utility for monitoring, treating or guiding risk management (Douglas & Kropp, 2002). In contrast, risk state is fluid and fluctuates with changes to risk factors, resulting in an indicator that is sensitive to change, and has utility for daily monitoring, highlighting targets for interventions.

Risk management is a reflection of conceptual understandings of violence risk as a fluctuating entity. Risk management has been defined as “the process of systematically focusing on methods of reducing both the severity and frequency of recognized adverse clinical risks for each individual patient” (Snowden, 1997, p. 33). Mullen and Ogloff (2009) have suggested a broad four stage process in assessing and managing risk: 1) evaluation of risk level and priority for intervention; 2) identification of targets for change and factors which may increase the person’s chances of engaging in violent behavior; 3) development of management strategies; and 4) evaluation of the effectiveness of these management strategies. Risk management then is
essentially a strategy for reducing the probability that a person will perpetrate violent behavior, and requires a knowledge of factors which are associated with an increase or decrease in the chances of engaging in violent behavior (Roesch et al., 2010). Risk assessment and risk management are therefore linked as the risk assessment process aids identification and knowledge of risk factors, which then present as targets for change in risk management. However, aligning risk assessment with risk management must go beyond identifying the presence of risk factors, to their relevance for the perpetration of violence and the formulation of risk. As Doyle and Logan (2012) have explained, “risk assessments that produce risk judgments based on summed scores or on an appraisal based on the risk factors that are present may only have a loose link to risk management. Assessments generating conclusions about level of risk – high, moderate, low – imply a volume of risk management, although not necessarily its focus” (p. 408). A thorough, detailed risk assessment grounded in theory-driven formulation can provide a solid foundation for moving seamlessly from the paradigm of risk assessment to risk management (Conroy & Murrie, 2012).

Hart (2009) defines risk factors as “a thing (condition, characteristic, event etc.) that demonstrably precedes the occurrence of a hazard and…may play a role in causing it” (p. 146). Another definition by Kraemer and colleagues (1997) states that violence risk factors are a “measurable characterisation of each subject in a specified population that precedes the outcome of interest and which can be used to divide the population into two groups (the high-risk group and the low-risk group)” (Kraemer at al., 1997, p. 338). The dichotomous ‘high-risk’ and ‘low-risk’ categories distinguish groups at the most basic level, but risk factors can be used to inform more discrete
outcomes (e.g., trichotomous Low, Moderate or High risk categories). These categories avoid inappropriate labeling of the individual as permanently violent or non-violent by acknowledging the fluctuating nature of dynamic risk variables, contextual determinants, and ultimately the probability of violence risk.

Because violence is considered a complex phenomenon, violence risk assessments are typically based on a combination of violence risk factors (i.e., biological, psychological or sociological characteristics that increase the likelihood of violent behaviour) (Brown & Singh, 2014; Monahan, 1988). Risk factors are generally considered static (historical, unchangeable variables) or dynamic (changeable) (Andrews & Bonta, 2010). Static and dynamic risk factors are also known as fixed markers and variable risk factors respectively within the risk assessment literature more generally (Kraemer et al., 1997). Modifiable dynamic risk factors can be further understood as being acutely dynamic (likely to change) or stably dynamic (unlikely to change) (Andrews et al., 2010). Violence risk factors are essentially predictive variables that are statistically or empirically associated with violence (Mullen et al., 2009). Causality is not assumed in linking predictor variables to violence, but rather an association with an increased level of risk (Mullen et al., 2009).

Violence risk assessments may also include protective risk factors (i.e., biological, psychological or sociological characteristics that decrease the likelihood of violent behaviour) as part of a balanced risk assessment (Brown et al., 2014; Rogers, 2000). This also has implications for risk management by encouraging treatment that diminishes risk factors but also enables protective factors (Blum & Ireland, 2004; de Vogel et al., 2014).
1.4 Three Generations of Violence Risk Assessment: A Short Course in the History and Development of Approaches to Assessing Violence Risk

Commentators have discussed five decision-making approaches to risk assessment: Unstructured clinical judgment, anamnestic assessment, actuarial prediction, adjusted-actuarial prediction, and structured professional judgment (Melton, Petrila, Poythress, Slobogin, Lyons, & Otto, 2007). These approaches, which have developed in succession and together, represent three generations of violence risk assessment (see Appendix A, Summary of Three Generations of Violence Risk Assessment).

1.4.1 First Generation: Unstructured Clinical Judgment

Prior to and into the 1970’s, violence risk assessment was typically performed using unstructured clinical judgment. Under this approach, risk was assessed qualitatively by the clinician, who would use a personally derived method for gathering and synthesizing self-identified risk-relevant information into a meaningful whole. This process occurs entirely in an unstructured manner, which generally means that the risk assessment is completed without the aid of a risk assessment tool (Brown et al., 2014).

Anamnestic assessment is a subcategory of the unstructured clinical judgment approach because, although it may occur in a slightly more structured way, it is still based on information and data sources that the clinician deems necessary. Anamnestic assessment involves examination of past violent events to identify the cognitive, behavioural, affective and contextual precipitants, and using this information to identify patterns or themes and hypotheses about the role of risk factors in future violent events.
(Otto, 2000). More recently, anamnestic assessment has been considered an adjunct to structured professional judgment (Melton et al., 2007).

Benefits of the unstructured clinical judgment approach included superior flexibility, tailoring of the assessment to the individual and inexpensiveness (i.e., no materials are required) (Brown et al., 2014). However, the flexibility of the approach also created an issue because personal approaches of assessors could vary significantly, as could rationales behind risk-related decisions. Judgments might have incorporated a range of aspects including clinician experience, knowledge base, gut feelings or intuitions (Mossman, 2013), resulting in low inter-rater reliability (Melton et al., 2007; Monahan & Steadman, 1994).

Monahan (1981) identified several issues in clinical predictions of violence: Lack of specificity in defining the predicted outcome; ignoring statistical base rates of violence; relying upon illusory correlations (variables that had little to no relationship with violence); and failing to incorporate environmental or contextual information into risk assessments. There is also criticism about clinician knowledge of valid risk factors and integration of information in a systematic way (Mullen et al., 2009).

1.4.1.1 Base Rates

Base rate refers to the prevalence of a specific event, behavior or condition (such as the occurrence of violence) within a defined population over a specified time period. Therefore, in the context of violence, base rate represents the “statistical prevalence of violent behaviour in a given group…the frequency with which violence is committed in a given time period” (Monahan, 1981, p. 33). Events with low base rates are thought to be
difficult to predict, because of the rarity of their occurrence. As a result, a person being asked to predict violence will probably incorrectly identify more people as violent, than those who actually are (Ogloff et al., 2005). Early research suggested that predictions of violence were flawed because such events were too uncommon to be predicted (Monahan, 1981). Early research by Steadman, Cocozza and Melnick (1978) looked at arrest rates of patients released from psychiatric facilities in New York in 1968 and 1975, and compared these to arrest rates from 25 years prior. They found that base rates of violence were fairly low, ranging from 2-5% over 1-year follow-up period.

Borum (1996) later challenged the assumption that base rates of violence were too low for violence to be predicted and highlighted that subsequent research (see below, Klaases & O'Connor, 1988a; 1988b; 1988c; 1989; Lidz, Mulvey, & Gardner, 1993 Steadman, Monahan, Appelbaum, Grisso, Mulvey, Roth, Robbins, & Klassen, 1994) demonstrated that base rates of violence were much higher than previously suggested. In a series of studies, Klaasen and O'Connor (1988a; 1988b, 1988c, 1989) looked at the probability of violence in male mental health patients from urban communities. Only men with the potential for violence (i.e., having a history of violent acts; thoughts/fears of harming others; command hallucinations to harm others; making threats to harm others; or attempted/ actual assault, rape, robbery, arson or homicide) were included. Researchers included arrest and rehospitalisation for an act that would have resulted in arrest for a violent crime, and found that 25-30% met that criterion after 1-yr follow up post-discharge (1989). Base rates of violence are also dependent on the length of follow-up and number of sources available to inform the outcome (e.g., self-
report, collateral and official records, being criminal, psychiatric or civil records). Douglas and Ogloff (2003) measured the post-release community violence of 193 involuntarily held civil psychiatric patients using three forms of archival data (criminal records, psychiatric hospital re-admission records and civil hospital re-admission records). They found that base rates varied as a function of sources, whereby the base rate identified through several combined sources surpassed the base rate identified through individual sources. Several other studies have supported these findings (e.g., Doyle, Carter, Shaw, & Dolan, 2012; Dolan & Doyle, 2000; Mulvey, Shaw, & Lidz, 1994).

Associated with the base rate issues were concerns around ceiling effects and the upper levels of predictive accuracy that could be reached. Monahan (1984; see also Monahan, 1988) suggested that there may be “a ceiling on the level of accuracy that can ever be expected of clinical prediction of violent behaviour” (p. 11). Monahan (1988) initially estimated that “the upper bound accuracy that even the best risk assessment technology could achieve was on the order of .33” (p. 250). Shah (1981) was more optimistic and said “to say that something is difficult to do (namely, to achieve high levels of accuracy in predicting events with low base rates) is not the same as asserting that the task is impossible and simply cannot be done” (p.161, emphasis in original). Subsequent research challenged these opinions and indicated that levels of accuracy could far surpass chance. However, most researchers and clinicians accept that “violence prediction will never be entirely accurate, given that violence itself is a complex concept” (Dolan et al., 2000, p. 309).
1.4.1.2 Transparency of Decision-making

Another consequence of the subjectivity of unstructured clinical judgment is that the decision-making process is unclear, fluctuating on a case-by-case basis. Although clinicians are not often required to provide overly detailed descriptions of how they arrived at their estimation of risk, the truth is this procedure is likely to be very muddled in the case of unstructured clinical judgment (Buchanan, 1999). This lack of transparency also leaves the approach open to criticism of being unsystematic (Pfohl, 1978).

1.4.1.3 Illusory Correlations

Another criticism of the unstructured professional judgment approach was that it left the assessment process vulnerable to biases in human thinking. Illusory correlations represent one such biases in decision-making. An illusory correlation exists when a clinician forms the perception that two variables are correlated, when in fact no relationship exists (i.e., the variable is not a risk factor for violence) (Chapman & Chapman, 1967). This fallacy ultimately distorts the violence risk assessment. Under unstructured assessment, clinicians are exposed to the possibility of making illusory correlations, especially because the lack of correlation between some seemingly risky cues and violence is unknown (Hart, 1998). Illusory correlations may form from personal or societal perceptions and stereotypes of ‘dangerousness’, and individual experiences which have either formed or confirmed a held bias (Murray & Thompson, 2010).

As Elbogen (2002) suspected, research has shown that clinicians do form illusory correlations in risk assessment and may be misguided by personal perceptions and stereotypes of who an offender is (e.g., Quinsey &
Maguire, 1986; Chapman & Chapman, 1969). Dawes, Faust and Meehl (1989) noted that “individuals have considerable difficulty in distinguishing valid and invalid variables and commonly develop false beliefs in associations between variables” (p. 1671).

### 1.4.1.4 The Link between Violence and Mental Disorder

The terms mental illness and mental disorder have been used interchangeably within the literature and considered synonymous (Kendell, 2002). The Diagnostic and Statistical Manual of Mental Disorders – Fifth Edition (DSM-V; American Psychiatric Association, 2013) notes that “no definition can capture all aspects of all disorders” (p. 5), and defines mental disorder as “a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning” (p. 20). Similarly, the International Classification of Mental and Behavioural Disorders (ICD-10; World Health Organization, 1992) has highlighted definitional challenges, noting that “the term disorder is used throughout the classification, so as to avoid even greater problems inherent in the use of terms such as disease and illness. Disorder is not an exact term, but it is used here to imply the existence of a clinically recognisable set of symptoms or behavior associated in most cases with distress and with interference with personal functions” (p. 5). Due to the lack of a consistent operational definition that can apply to all situations herein, the term mental disorder has been adopted, as defined in the DSM-V and ICD-10. The term mental illness has been included where specified by the relevant source.
It is beyond the scope of this thesis to review individual risk factors for violence. However, given that the sample studied is forensic psychiatric in nature, it is important to acknowledge the link between mental disorder and violence. There has been a longstanding debate in the fields of mental health and violence about the presence and nature of the relationship between mental disorder and violence (Strub, Douglas, & Nicholls, 2014). Beliefs regarding the nature of the association between mental disorder and violence have changed over time.

Prior to the 1960’s, the general consensus was that persons with mental illness were more prone to violence (Ogloff et al., 2005). This was reflected in clinical decision-making around the continued detainment of mentally disordered persons for the safety of the community. Steadman and Cocozza, (1978) noted that one of the erroneous assumptions underlying the unstructured clinical judgment movement was beliefs that there was a “strong, understood relationship between mental illness and violent behaviour” (p. 226). Steadman and Cocozza (1978) described these assumptions as being invalid, and drew attention to the popular belief amongst clinicians during these times that “one must be crazy in order to be an habitually violent offender” (p. 230).

By the late 1960’s onwards, opinions on the relationship between mental illness and violence shifted (Ogloff et al., 2005). Early studies conducted during the pre-deinstitutionalisation period (e.g., Cohen & Freeman, 1945; Pollock, 1938) indicated that psychiatric patients actually had a lower arrest rate than that observed in the general population. Over the next two decades, from the late 1970’s to 1980’s, emerging research focused on critical evaluation of whether persons with mental disorder exhibited similar
levels of violence to non-disordered persons (see reviews by Harris & Lurigio, 2007; Ogloff et al., 2005). This research brought into question “the conventional wisdom among social scientists that no significant relationship existed between violence and mental illness” (Borum, 1996; p. 946). A body of research began to form suggesting that there was indeed a relationship between mental disorder and violence. Some studies demonstrated that rates of violence amongst samples of persons with mental disorder were three to six times greater than rates of violence amongst non-disordered samples (e.g., Link, Andrews, & Cullen, 1992; Swanson, Holzer, Ganju, & Jono, 1990).

Link, Andrews and Cullen (1992) compared mental patients (N = 375) to community residents (N = 521) in New York City using official and self-report measures of violent behaviour. Outcome data was obtained prior to self-reports regarding arrest history, fighting/physically assaulting others and weapon use. Results indicated that mental patients had higher rates on all measures, a difference which was not accounted for by sociodemographic variables. Swanson and colleagues (1990) used data from the Epidemiologic Catchment Area Survey to examine the relationship between violence and mental illness amongst community-based adult cohorts. Psychiatric assessments were based on the Diagnostic Interview Schedule. Results indicated that reported violent behaviour increased with the number of psychiatric diagnoses. Such outcomes caused Monahan (1992) to suggest, “mental disorder may be a robust and significant risk factor for the occurrence of violence” (p. 519). However, the conclusion reached was that mental disorder was not causative of violence, most people with mental disorders did not commit violent offences, and the strongest risk factors for violence in psychiatric populations were also the strongest risk factors in non-psychiatric
populations (e.g., Psychopathy, Anti-social Personality Disorder and substance use) (Bonta, Law, & Hanson, 1998; Harris et al., 2007; Ogloff et al., 2005). Link and colleagues (1992) also noted that elevated risk was not simply about having a mental illness, but more so the presence of active symptoms, and the importance of distinctions between the type of mental illness and associated symptom level nuances (Binder & McNiel, 1988; Link & Stueve, 1994).

By the late 1990’s, enough research had emerged to support the view that there was indeed a link between mental disorder and violence (Harris et al., 2007). The definitive MacArthur Violence Risk Assessment Study (Monahan, Steadman, Silver, Appelbaum, Robbins, Mulvey, Roth, Grisso, & Banks, 2001) is the largest study of community violence amongst male and female mentally disordered persons ever undertaken. Conducted in the USA, the study involved more than 1,000 acute civil psychiatric patients who were assessed for a large range of risk factors (130+), across dispositional, historical, contextual and clinical domains. Participants were selected from acute inpatient facilities across three sites in America. Certain selection criteria were maintained: 1) civil admission; 2) aged 18 – 40; 3) English-speaking; 4) White, African-American or Hispanic; and 4) presented with certain psychiatric diagnoses. Patients who met these criteria were then sampled according to age, gender and race in order to maintain comparable distribution. Outcomes from this cohort were compared to a control group of approximately 500 individuals who were geographically similar (i.e., lived in similar areas) but had not been admitted to psychiatric hospitals. Participants were followed-up in the community for 1-year post-discharge from hospital. The outcome variable, community violence, was assessed through official
records (i.e., arrest and hospital files), and interviews with patients and collaterals (e.g., family, friends, professionals, co-workers etc.). Results indicated that the relationship between mental illness and violence was far more complex than previously thought. One of the key findings was that confounding factors such as substance use were particularly important. Results indicated that violence perpetrated by psychiatric patients without substance abuse problems matched community outcomes (see also Steadman, Mulvey, Monahan, Robbins, Appelbaum, Grisso, Roth, & Silver, 1998). Ogloff and Davis (2005) commented on such findings, “simply stated, mental illness is a risk factor for violence that is made worse when those with mental illnesses also have substance abuse problems” (p. 316). The significant influence of personality pathology as a risk factor for violence was also highlighted.

These results challenged previously held beliefs about the relationship between mental disorder and violence. Authors noted that this complexity was indeed one of the reasons clinicians’ unstructured means of assessing violence was so inaccurate, as they were based on fallacious understandings of the relationship between mental disorder and violence, limited knowledge of valid risk factors and integration of same into risk assessments.

1.4.1.5 Naturalistic Experiments

A pivotal occurrence that shaped the progression of violence risk assessment approaches was a number of ‘naturalistic experiments’ in which patients were released for legal reasons, providing the opportunity for the unanticipated investigation of the accuracy of clinicians’ predictions of violence (Monahan, 1981). The Baxstrom v Herald (1966) ruling resulted in
the release or transfer of 966 patients from maximum-security hospitals in the State of New York to the community or less secure hospitals, because their detention was deemed unconstitutional. Steadman and Cocozza (1974) reported on the 4-year outcomes of this cohort in their landmark study, ‘Careers of the Criminally Insane’. They found that only 20% of the cohort had been reconvicted, mostly for non-violent offences. This fueled notions that unstructured clinical judgement was flawed (Dolan et al., 2000).

The Baxstrom decision set a precedent in the confinement and release of forensic psychiatric patients, and preceded similar decisions throughout the United States (Wood, 1980). Another opportunity for a naturalistic experiment was seized by Thornberry and Jacoby (1979). Their study paralleled Steadman and Cocozza’s (1974) research in that, a court had ordered the release of patients from Fairview State Hospital for the Criminally Insane in Pennsylvania (another maximum-security mental facility), based on the fact that their sentence had expired, and that ongoing commitment was unconstitutional. In 1971, these 586 patients (also known as the Dixon patients, as the decision was based on Dixon v. Attorney General of the Commonwealth of Pennsylvania) were transferred to civil hospitals, and many were later discharged into the community (65%). The court decision contrasted and over-ruled the judgments of staff, who maintained that the patients were ‘too dangerous’ for release. The results were similar to those involving Baxstrom. During the 4-year follow-up period, more than three-quarters of the group remaining arrest free, and only 14% of patients were rehospitalised for a violent crime. These results indicated that post-release violence was similar to those of people released from civil psychiatric facilities. In explaining these outcomes, Thornberry and Jacoby suggested
that assessments of violence were 1) formed on the assumed characteristics of a group of individuals (i.e., the criminally insane), and the fallacious presumption of violence within this group; and 2) prediction was influenced by the fear and embarrassment of inaccuracy, hence the high over-prediction rates. This particular study provided strong evidence that the presumption of violence within forensic psychiatric samples was fallacious, as was the commonly held perception that such persons remained dangerous long after committing violent acts. These beliefs resulted in unnecessarily long periods of confinement for persons who were assumed to be dangerous. The study was influential in that it added to the limited body of research on the confinement and release of so called ‘dangerous’ patients (Levinson, 1980).

Another landmark case which shaped the importance of violence risk assessment in duty to protect was the Tarasoff decision (Tarasoff v. Regents of the University of California, 1972) which set a legal precedent that mental health professionals should identify and know the likelihood of injury to third parties, and may be held partially responsible for violence perpetrated by their patients (Mossman, 2013; Brown et al., 2014). The underlying assumption of this ruling is that the risk of violence can be accurately assessed. The outcomes of such naturalistic experiments and assumptions underlying the unstructured clinical judgment approach cast doubt and began to breed speculation about whether prediction of violence by clinicians was indeed possible. This led to further empirical research.

1.4.1.6 Empirical Research

In what Monahan (2000) considered one of the most ‘sophisticated’ studies in unstructured clinical judgment, Lidz, Mulvey and Gardner (1993)
asked mental health professionals to assess patients in an acute psychiatric emergency department for violence towards others within the following 6-months. Results indicated that the accuracy of clinical judgment was significantly better than chance, as patients whom professionals were concerned about regarding future violence were significantly more likely to be violent post-release (53%) than those for whom they had no concern (36%). Lidz and colleagues concluded “clinical judgment has been undervalued in previous research” (p. 1010). Other research by Gardner, Lidz, Mulvey and Shaw (1996) also demonstrated that while actuarial measures may trump unstructured clinical judgment approaches, clinical judgment could perform better than chance.

However, more studies began to emerge that suggested the opposite, forming a body of research that indicated the inaccuracy of the unstructured clinical judgment approach. As these shortcomings began to surface through empirical research, Monahan (1984) noted, a “dark cloud began to form in our crystal balls” (p. 10). Monahan’s 1981 seminal monograph represents one of the strongest criticisms of the unstructured clinical judgment approach. Monahan (1981; 2007) reviewed five key studies conducted in the 1970’s that form the core of knowledge on first generation approaches (namely, Cocozza & Steadman, 1976; Kozol, Boucher, & Garofalo, 1972; Steadman, 1977; Steadman et al., 1974; and Thornberry et al., 1979). He concluded that “psychiatrists and psychologists are accurate in no more than one out of three predictions of violent behaviour over a several-year period among institutionalised populations that had both committed violence in the past…and who were diagnosed as mentally ill” (p. 47, 49). These findings contrast with psychologists’ appraisals of their own abilities to predict
dangerousness. Givelber, Bowers and Blitch (1984) surveyed 1,577 mental health professionals (psychologists, psychiatrists and social workers) about their ability to predict physical harm to another by one of their outpatients. Results showed that 5% of respondents felt there was “no way to predict” and over three-quarters of respondents felt that they could make a “probable” to “certain” prediction (p. 463).

An important finding relating to early unstructured clinical judgment risk assessment studies was that clinicians tended to severely overestimate violence. As Ennis and Litwack (1974, p. 734 - 735) stated, “psychiatrists have absolutely no expertise in predicting dangerous behaviour – indeed they may be less accurate predictors than laymen – and…they usually err by overpredicting violence”. Overestimations were not slight. For example, Belfrage (1998) found that clinicians predicted 90% of a sample of 640 forensic psychiatric patients to be “at risk of severe criminality”, when only 50% actually recidivated. Over-prediction is likely to stem from the consequences of two forms of prediction errors, false positives and false negatives, being completely disproportionate and ‘diametrically opposed’ (Davis, 2010; Ogloff et al., 2005). A false positive describes a person assessed as being high-risk, who does not offend; whereas a false negative describes a person assessed as being low-risk, who does in fact offend. The goal of risk assessment then is to minimise false positives and false negatives, whilst maximising correct predictions, true positives and true negatives (Ogloff et al., 2005). The potential consequences of false negatives are violent actions and harms to persons, which can lead to community outrage and concerns over community safety. On the other hand, false positives do not invoke consequences for the larger community, but only the
person detained based on an inaccurate assessment. Furthermore, false positives are not disproved because there is no opportunity to do so given that the supposedly ‘high-risk’ person is not released (Davis, 2010). This dilemma is known as ‘sample censoring’, where predictions cannot be validated nor disproved because the opportunity to do so never eventuates (Monahan 1988; Quinsey et al., 1986). As would be the case with assessments of dangerousness in the Baxstrom and Dixon cohorts, yet unforeseen events ultimately proved these assessments to be inaccurate.

Scholarly conclusions about first generation approaches were pessimistic to say the least. Several scholars noted that psychologists were more likely to be wrong in predicting legally relevant behaviour like violence (e.g., Morse, 1978; Monahan, 1981; Steadman et al., 1974). Ennis and Litwack (1974) described the clinical predictions of violence as “flipping coins in the courtroom” (p. 693). Such views based on research outcomes led the task force of the American Psychiatric Association (1974) to conclude that “the ability of psychiatrists or and other professionals to predict future violence is unproved” (p. 30).

With these limits on prediction in mind, it has been questioned whether unstructured clinical judgments of violence are in fact unethical (Grisso & Appelbaum, 1992). Ewing (1983) weighed in on the debate by adding that, in the context of capital sentencing proceedings and liberty, “the psychiatrist or psychologist who makes a prediction of dangerousness violates his or her ethical obligation to render judgments that rest on a scientific basis” (p. 418). However, these appraisals must be considered within the historical context, where there was limited research available to actually inform what the empirically supported risk factors were (Ogloff et al., 2005). Furthermore,
there was some evidence to suggest that clinical predictions were better than chance guesses (Lidz et al., 1996).

1.4.1.7 Conclusion

The unstructured clinical judgment approach was used in forensic evaluations for over 100 years and, despite the criticisms, continues to be used by some risk assessors (Melton et al., 2007). Following his review, Monahan (1981) made two recommendations that were employed in the development of second-generation violence risk assessment approaches: 1) increased emphasis on statistical concepts in clinical prediction; and 2) heightened sensitivity to environmental and contextual factors. There was a specific call for studies to look at other approaches to violence risk assessment, such as actuarial methods to better identify valid risk factors so as to avoid fallacious judgments. There was also a call for greater emphasis on environmental and contextual factors impacting on risk; short-term as opposed to long-term predictions; and consideration of risk in the community as opposed to institutional settings (Monahan, 1984). Regarding the state of unstructured clinical risk assessment, Monahan declared “there are so many nails now in that coffin that I propose we declare the issue officially dead” (p.13).

1.4.2 Second Generation: Actuarial Assessment

A shift from first generation unstructured approaches to second generation actuarial approaches began to occur between 1970 and 1980 (Public Safety Canada, 2009). The second generation of violence risk assessment saw the emergence of actuarial prediction approaches that
shunned the unstructured clinical judgment approach. The earlier writings of two key proponents of the actuarial approach, John Monahan and Paul Meehl, had contributed to a solid foundation of research and commentary that criticised the unstructured clinical judgment approach, describing it as a “prestigious synonym for anecdotal evidence” (Grove & Meehl, 1996, p.302). Hence, the second generation represented a push towards evidence-based assessment characterised by three central themes: 1) emphasizing the limitations of existing knowledge; 2) enthusiasm about possible improvements in predictive accuracy; and 3) evaluation of public policy involving violence prediction (Monahan, 1984). Monahan summarised these three goals as: “we know less than we thought about the accuracy of predictions; what little we do know may be improved upon; and how useful this knowledge is depends on what we do with it, compared to what we would do without it” (p.13). Whilst improvements in accuracy was a focus of the second generation, there was some skepticism about what could be achieved. As Monahan (1984) said, “no one thinks that the prediction of violence is on the verge of attaining a validity comparable to that of the prediction of the weather” (p.11).

1.4.2.1 Defining Actuarial Risk Assessment

In practical terms, actuarial risk assessment (also known within the literature as statistical, mechanical or formal risk assessment) involves assessing risk using statistical analyses. This is based on a predetermined set of predictor variables which are algorithmically combined (Davis, 2010; Grove et al., 1996; Grove, Zald, Lebow, Snitz, & Nelson, 2000). The actuarial method relies completely on empirically established relationships between data and outcome/s (i.e., a condition or event) (Dawes et al., 1989). This is
based on nomothetic and quantitative procedures. The actuarial method is generally void of human decision-making and relieves the clinician of the responsibility of knowing what variables to look for, and how to integrate them (Dawes & Corrigan, 1974; Roesch et al., 2010). Although judgments by assessors are often required to answer questions relating to risk factors in the algorithm. Actuarial instruments are based on probability estimates using the statistical relationship between combinations of risk factors and the likelihood of violence, providing automatic outcome interpretations (Davis, 2010; Dawes et al., 1989).

Risk factors are typically derived from studies where these factors have been shown to statistically outperform others in a construction sample, and are generally weighted according to the amount of variance they capture in predicting the outcome (Brown et al., 2014). Whilst clinical judgment may be needed to elicit the information for each risk factor, once this information is coded into the actuarial instrument, the final assessment of risk is essentially mechanical (Ogloff et al., 2005). Commonly known actuarial assessment tools in the forensic context include the Violence Risk Appraisal Guide – Revised (VRAG-R; Harris, Rice, Quinsey, & Cormier, 2015) and the Static-99-Revised (Static-99R; Helmus, Thornton, Hanson, & Babchishin, 2012).

1.4.2.2 Empirical Research

The first published research on violence risk assessment using actuarial methods emerged in the mid 1990’s (e.g., Gardner et al., 1996; Rice & Harris; 1995). The overall superiority of actuarial approaches over unstructured clinical judgment approaches has now been established, with research suggesting that actuarial approaches perform significantly better
This holds true for a range of outcomes. Within the social sciences, approximately 100 comparative studies have demonstrated the superiority of actuarial methods over clinical judgment (Dawes et al., 1989; see also Sawyer, 1966). Within violence specifically, the superiority of actuarial methods has been demonstrated for several forms of violent prediction including sexual, violent and general recidivism (Hanson & Bussiere, 1998); and with mentally ill offenders (Bonta et al., 1998).

Meehl’s classic 1954 publication ‘Clinical versus statistical prediction: A theoretical analysis and a review of the evidence’ provided a comprehensive comparison of the two approaches, based on the limited research available at the time. Meehl drew on 20 studies and found that the two methods performed equally in eleven studies. Actuarial approaches were superior in eight studies, and clinical judgment was superior in one study (however, measures of clinical judgment were crude). Meehl concluded that actuarial methods for predicting almost any outcome were superior to clinical judgment approaches. In considering these outcomes, Meehl stated “the less efficient of two prediction procedures in dealing with such matters is not only unscientific and irrational, it is unethical” (1954, p. 320). Regarding the debate on clinical versus actuarial methods, Meehl later stated, “there is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly” (1986; p. 373).

Another more recent meta-analysis by Grove and colleagues (2000) involved 617 distinct empirical comparisons between clinical and actuarial predictions, based on 136 studies. Eligibility for the meta-analysis was broad, including: Any study published in English since the 1920’s, where prediction
concerned human behaviour or health-related phenomena and wherein outcomes of clinical prediction and actuarial prediction were described. Of 136 studies, 64 studies favoured actuarial approaches, 64 studies showed equal results, and eight studies favored clinical judgment approaches. Based on this study, Grove and colleagues (1996) concluded, “the mechanical method is almost invariably equal to or superior to the clinical method” (p.293). This research further consolidated early research findings by Meehl (1954), with a much broader sample. As Grove and Meehl noted, “forty years of additional research published since (Meehl’s) review has not altered the conclusion he reached. It has only strengthened that conclusion” (p.299).

Interestingly, the superiority of actuarial measures emerged despite allowances for clinical judgment determinations to be based on more information than available to actuarial judgments. This imbalance occurs across the majority of studies, so the clinical approach was advantaged and clinical assessors always had more information available to them than actuaries.

1.4.2.3 The Actuarial Advantage

As outlined above, research has demonstrated the superiority of actuarial approaches. In addition to improvements in prediction, actuarial assessments are advantaged by objectivity, transparency and ease of administration (Brown et al., 2014). Actuarial instruments promote predictive validity and reliability by identifying and combining predictors in a specific and transparent manner, with a clear decision-making process (Davis, 2010). The clarity of such a process is critical to assuring others, such as the courts, that the determination was reasonably made (Borum, 1996). This contrasts with
unstructured clinical judgment, where it is difficult to determine the factors and process clinicians used to form their final impressions. Within the scientific method, this transparency is also crucial for replication (Buchanan, 1999). As Dawes and colleagues (1989) noted, “actuarial methods are explicit, in contrast to clinical judgment, which rests on mental processes that are often difficult to specify. Explicit procedures facilitate informed criticism and are freely available to other members of the scientific community who might wish to replicate or extend research” (p. 1673). Indeed, actuarial methods are also reasonably easy and efficient to administer. Most are based on static factors, providing baseline assessments of risk that require no assessment of current mental state or future contextual considerations.

Actuarial approaches also have the advantage of using algorithmic procedures, which ensure that prominent risk factors are considered and illusory correlations are excluded. By not having a clinical judgment component, actuarial assessments avoid the biases associated with clinician discretion and human error in decision-making. Thus, the clinician is protected from the possibility of false negatives (Mullen et al., 2009).

1.4.2.4 Criticisms of the Actuarial Approach

Despite these advantages, actuarial approaches are not without limitation. Actuarial approaches have been criticised for being atheoretical and based solely on available variables that show associations with recidivism within construction samples (Public Safety Canada, 2009). Furthermore, actuarial tools have been criticised for a lack of generalisability to populations other than those they were developed from, particularly given that the actuarial approach focused on specified populations (Davis & Ogloff, 2008;
Advocates of the clinical approach have argued that conclusions derived from groups may have limited value in predicting the behaviour of individuals (Allport, 1940). Furthermore, although actuarial measures move from group construction to individual application, the assessment cannot be tailored to the individual, nor tempered to include other risk relevant information. As Howe and colleagues (2016) commented:

One of the primary drawbacks is that actuarial instruments combine risk factors for every individual in the same way. All risk factors receive the same empirical weight, and all risk factors are summed according to the same algorithm, regardless of potentially relevant differences in the individual being evaluated or the situation in which risk of violence is being assessed. There are likely instances in which the algorithm may produce a risk estimate that is not in agreement with the clinician’s professional judgment, at which point the clinician must struggle to reconcile two conflicting sources of information (pp. 399).

Because risk factors are selected purely on a statistical basis, this has oftentimes resulted in illogical risk factors emerging. For example, having a diagnosis of schizophrenia was negatively weighted in the Violence Risk Appraisal Guide (VRAG; Quinsey et al., 2006), although researchers and clinicians alike would probably disagree with schizophrenia being considered a “protective factor” (Kroner, Stadtland, Eidt, & Nedopil, 2007, p.97).

The static nature of actuarial tools means they are ill suited for measuring change in risk over time. This is problematic given that conceptual understanding of risk as a fluctuating entity. Actuarial approaches have also
been criticised for failing to prioritise clinically relevant information, as they generally consist of static risk factors (Hart, 1998). These static risk factors often address the individual's forensic history, making them non-applicable to predictions of first violent offences. The lack of dynamic risk factors focusing on current mental functioning also creates issues for addressing violence that stems from mental illness. Furthermore, it has flow-on effects for informing risk management (Hart, 1998). This limits the utility of the assessment as a foundation for risk management plans. The method has also been criticised for not being able to foster any ideas around the cause of violence (Buchanan, 1999; Floud & Young, 1981). Buchanan (1999) argues that such explanations of human behaviour should be essential to violence risk assessment. Grove and Meehl (1996) counteract this argument by noting that “putting scholarly thirst for alleged understanding ahead of an institutionally defined pragmatic task…It does not appear that some surplus understanding over and above those components of diagnosis that have actuarial predictive value accomplishes much of anything” (p. 314).

Another criticism of actuarial assessment approaches is the issue of outcomes and time span. Based on construction procedures, actuarial tools such as the VRAG-R provide violent recidivism probability estimates at 5-year and 12-year time points. This could be problematic, as Buchanan (1999) put it, “a table providing estimates of the risk of interpersonal violence over a 5-year period…is of little use to a clinician who wishes to predict the risk of his patient committing arson before his next out-patient appointment” (p. 469). However, it is also the advantage of actuarial instruments that longitudinal estimates of risk can be provided.
1.4.2.5 Broken Legs: Rare Events impinging upon Risk Assessments

The rigid approach adopted by actuarial tools means that, whilst structure and reliability is increased, there is no room for the consideration of rare yet important factors in the individual case. Furthermore, because actuarial tools are generally developed from construction samples, where the selected risk factors are those with the strongest associations, rare factors are highly unlikely to be in the sample, let alone emerge as significant predictors. Within the violence risk assessment literature, these rare factors have come to be known as the issue of ‘broken legs’ (Grove et al., 1996; Meehl, 1954; Davis, 2010; Dawes et al., 1989). This term is based on a hypothetical situation in which an actuarial tool is developed demonstrating great accuracy in predicting weekly movie attendance. However, if the subject of the prediction breaks his leg, a rare event/factor, which the tool is unable to account for, then the prediction may be flawed.

A frequently mentioned broken leg in the violence risk assessment context is threats of serious harm. Hart (1998) suggested that risk factors on actuarial instruments mean nothing if the person genuinely expresses intent, and as such it would be foolish to exclude such relevant information. Warren, Mullen, Thomas, Ogloff and Burgess (2008) examined serious violence in a sample of 613 participants convicted of making threats to kill. They found that, within a 10-year period, approximately 45% were convicted of violent offences, suggesting high rates of violence following threats to kill. Capgras delusions and diagnoses such as delusional morbid jealousy are further examples of factors relevant to violence risk, which are far too uncommon to emerge as risk factors in actuarial tools (De Pauw & Szulecka, 1988; Buchanan, 1999; Mullen et al., 2009).
1.4.2.6 Understanding the Superiority of the Actuarial Approach

Many reasons have been offered to explain the superiority of actuarial approaches. Grove and Meehl (1996) describe unstructured clinical judgment as a “subjective, impressionistic, in-head approximating job of actuarial computation” (p. 315), and offer an explanation of why actuarial procedures are superior to clinical methods:

The clinician’s brain is functioning as merely a poor substitute for an explicit regression equation or actuarial table. Humans simply cannot assign optimal weights to variables, and they are not consistent in applying their own weights…the human brain is a relatively inefficient device for noticing, selecting, categorizing, recording, retaining, retrieving, and manipulating information for inferential purposes (pp. 315-316).

1.4.2.7 Resistance to the Actuarial Movement

Gardner and colleagues (1996) suggested that clinicians may not be adopting actuarial approaches because of their mathematical basis (i.e., calculations are required). They provide a progress report on the development of actuarial methods for predicting violence and discuss strategies (e.g., regression trees) for making actuarial predictions of violence more “understandable, acceptable, and efficient in routine clinical use” (p. 47).

Grove and Meehl (1996) discuss 17 anti-actuarial arguments and resistance within the field to the actuarial approach. Their article came as a reaction to the fact that, in the 1990’s, “despite 66 years of consistent
research findings in favour of the actuarial method, most professionals continue to use a subjective, clinical judgment approach when making predictive decisions” (p.299). Some of the arguments centered on using both methods, and concerns over idiosyncrasy. Grove and Meehl (1996) reject both arguments. They suggest that it is impossible to use two methods that offer outcomes that are opposed. This view is based on the assumption of disharmony between clinical and actuarial approaches, which is not necessarily the case. Grove and Meehl then call it a matter of science and professional ethics to opt for the method that is empirically supported.

They also address the issue of anti-actuarialists who feel that aggregated statistical figures are irrelevant in dealing with unique persons. They provide a useful hypothetical situation to illustrate their rebuttal of this argument:

Suppose you are suffering from a distressing illness…and your physician says that it would be a good idea to have surgeon X perform a certain radical operation in the hope of curing you. You would naturally inquire whether this operation works for this disease and how risky it is. The physician might say “Well, it doesn’t always work, but it’s a pretty good operation. It does have some risk. There are people who die on the operating table, but not usually”. You would ask, “Well, what percentage of the time does it work?”…How would you react if your physician replied, “Why are you asking me about statistics? We are talking about you – an individual patient. You are unique. Nobody is exactly like you. Do you want to be a mere statistic? What differences do those percentages make, anyway? (pp. 305).
Dawes and colleagues (1989) also dispute this anti-actuarial position, and explain that, “although individuals and events may exhibit unique features, they typically share common features with other persons or events that permit tallied observations or generalisations to achieve predictive power” (p. 1672).

Advocates of the actuarial method have called into question the rational, ethical and scientific backbone of anti-actuarialists. As Dawes and colleagues (1989) commented, “failure to accept a large and consistent body of scientific evidence over unvalidated personal observation may be described as a normal human failing or, in the case of professionals who identify themselves as scientific, plainly irrational” (p. 1673). Grove and Meehl (1996) echoed this sentiment, “we know of no social science controversy for which the empirical studies are so numerous, varied, and consistent as this one…one must classify continued rejection (or disregard) of the preactuarial generalisation as clear instances of resistance to scientific discovery…to use the less efficient of two prediction procedures in dealing with such matters is not only unscientific and irrational, it is unethical” (pp. 318-320).

1.4.2.8 Conclusion

Comparative studies clearly indicate an increase in the accuracy of predictions in the second generation of violence risk assessment using actuarial approaches, indicating that actuarial approaches improved upon first generation unstructured clinical judgment approaches. The general consensus amongst scholars was that actuarial approaches showed superior predictive validity, and “perform(ed) better than the best guesses of most experts” (Mullen et al., 2009, p. 1997; see also Grove et al., 1996; Harris & Rice, 1997). Despite these improvements in performance, actuarial
assessment was still considered to be theoretically unsound by many, who
held the view that risk assessment should be based on clinical formulation,
not calculation (Mullen et al., 2009). Whilst the two approaches were seen as
being the antithesis of each other, it was considered that the two may exist on
a continuum, with completely structured and unstructured on either end, and
partially structured methods between them (Skeem & Monahan, 2011). A
third generation of violence risk assessment emerged, representing this
middle ground.

1.4.3 Third Generation: Structured Professional Judgment

The third generation of violence risk assessment essentially
represented “a melding of the clinical and actuarial approaches” (Roesch et
al., 2010, p. 46). The structured professional judgment (SPJ) approach (also
known as structured clinical judgment or adjusted actuarial prediction) is
essentially a hybrid approach to violence risk assessment, combining the
preceding two generations. Dolan and Doyle (2000, p. 303) described the
third generation as “bridging the gap between clinical and actuarial
measures”, where the predictive power of the actuarial approach is met with
the flexibility of clinical judgment (Davis, 2010). In this way, the SPJ approach
acknowledges both the merits and limitations of clinical judgment and
actuarial assessment by allowing the clinician to conduct risk assessments by
integrating information in a systematic and structured way (that is scientifically
based and grounded in theory), but to also render these outcomes based on
clinical knowledge and expertise (Hart, 1998). Therefore, an amalgamation of
the two approaches under SPJ seeks to minimise the weaknesses of
unstructured and actuarial approaches, whilst retaining their strengths. In
contemporary practice, the SPJ approach has been well adopted “by mental health professionals around the world who felt acknowledged in their forensic clinical expertise and at the same time strengthened by the empirical basis of the SPJ checklists” (de Vogel, van den Broek, & de Vries Robbé, 2014, p. 109).

‘Structured’ herein distinguishes SPJ from first generation unstructured approaches, and indicates the guided nature of decision-making. This structure is employed in several different ways: through specifying the risk factors to be considered; guiding systematic data collection through carefully defining and operationalising risk factors with scoring guidelines; and through guided tempering of the anchored assessment, resulting in a clinician-decided final risk judgment (Hart, 1998; Ogloff et al., 2005; Roesch et al., 2010). In this way, SPJ tools act as an ‘aide memoire’ to clinicians, directing them to key aspects of an empirically supported risk assessment (Douglas et al., 2013). Because SPJ tools are based on reviews of the literature, they are considered comprehensive in nature, placing equal emphasis on the science and research of violence risk assessment, and the actual practice of it (Douglas & Belfrage, 2014).

Historically, risk has been presented as a stable construct using dichotomous outcomes (i.e., dangerous versus not dangerous) (Monahan et al., 1994). The third generation embraced conceptual changes in violence risk assessment, where risk is conceptualised as a variable, fluctuating entity (Davis, 2010; Douglas et al., 2013). This is reflected in the re-emergence of dynamic risk factors into assessment tools, and probabilistic outcomes such as Low, Moderate or High risk of violence as opposed to fixed yes/no outcomes. It has also expanded to include more detailed and informative
assessments, which address risk, but also how risk may present itself (i.e., seriousness of harm, type of violence, potential scenarios etc.) (Davis, 2010).

Commonly utilised SPJ tools in forensic assessment includes the Historical Clinical Risk Management-20 (Version 3) (HCR-20\textsuperscript{V3}; Douglas, Hart, & Webster, 2013), the Sexual Violence Risk-20 (SVR-20; Boer, Hart, Kropp, & Webster, 1997), the Short-Term Assessment of Risk and Treatability (START; Webster, Martin, Brink, Nicholls, & Desmarais, 2009), the Spousal Assault Risk Assessment Guide (SARA; Kropp, Hart, Webster, & Eaves, 1994) and the Stalking Risk Profile (SRP; MacKenzie, McEwan, Pathé, James, Ogloff, & Mullen, 2009).

Although past research favoured actuarial over unstructured professional judgment approaches, the introduction of structured clinical judgment (as opposed to unstructured) challenged the superiority of actuarial approaches. Some research now indicates that the structured approach to clinical decision-making can be as accurate, and in some cases more accurate, than a purely actuarial approach (Campbell, French, & Gendreau, 2009; Davis, 2010; see Chapter Two, Literature Review for a more in depth review).

1.4.3.1 Advantages of the Structured Professional Judgment Approach

SPJ approaches are more focused on individuals than groups of individuals, but the risk factors which clinicians are directed to use are drawn from research which may be based on groups (Brown et al., 2014). A key feature of the third generation approach was extending on from risk assessment to risk management, through the inclusion of dynamic risk factors. In this way, SPJ tools are more comprehensive, informing both risk...
status and risk state, whereas actuarial tools are only informative of risk status. This emphasis on both risk state and status lays the foundation for risk management, whilst the clinician is still in the assessment phase. By the time the assessment is finalised, the clinician already has information on risk status informing the level of intervention required, and the specific dynamic risk factors that should be addressed in the treatment plan. The clinical judgment aspect of SPJ also allows for the consideration of unique, case-specific risk factors that the clinician deems relevant, such as the ‘broken legs’ discussed previously (Dawes et al., 1989). Clinical judgment and formulation also foster ideas around the causes of violence. Actuarial methods have been criticised on this point for restricting deeper understandings of the perpetration of violence and future risk management.

1.4.3.2 Clinician Preferences in Risk Communication

The SPJ approach also provides clinicians with another alternative for communicating risk, as opposed to statistical probabilities in the actuarial approach. SPJ outcomes are usually described using terms of Low, Moderate or High risk. Heilbrun, Phillipson, Berman and Warren (1999) found that clinicians are reluctant to use numerical risk estimates. Of 55 clinicians, only one clinician preferred numerical probabilistic communications of risk, compared to the remainder who preferred categorical communications of risk. Participants cited several reasons for this preference, such as “I don’t want to be held accountable for being that precise” (p. 401). These findings have been replicated by others (e.g., Heilbrun, O’Neill, Strohman, Bowman, & Philipson, 2000; Huss & Zeiss 2004), who found that mental health professionals prefer and find categorical forms of risk communication to be
more valuable. Although research on the importance of language in communicating risk is still in its infancy, the limited research does suggest that clinicians prefer categorical communications of risk, which has inherent implications for management (i.e., high risk indicates an urgent need for intervention) (Heilbrun et al., 2000; Douglas et al., 2013).

1.4.3.3 Disadvantages of the Structured Professional Judgment Approach

As a combination of the two approaches, SPJ approaches are not as objective as actuarial approaches (because of the re-emergence of human decision making into the risk assessment process), and not as subjective as unstructured approaches (because of structured frameworks) (Brown et al., 2014). Within the SPJ approach, actuarial or structured outcomes are open to adjustment by the clinician, either based on differences of the individual to the intended population (e.g., gender, race, population or cultural consideration), or the presence of ‘broken legs’. Some have argued however that the tempering of actuarial outcomes with clinical judgment weakens their reliability and validity (Quinsey et al., 2006). Grove and Meehl (1996) are critical of those who support the option of adjustment, by noting that advocates of clinical adjustment “focus their attention on the cases in which they could have saved an actuarial mistake, neglecting the obvious point that any such decision policy…will also involve making some mistakes in the opposite direction” (p. 308).

A question over the extent of the adjustment is also important. This issue was addressed by developers of the VRAG, who originally stated: “if adjustments are made conservatively and only when a clinician believes, on
good evidence, to be related to the likelihood of violent recidivism in an individual case, predictive accuracy may be optimised" (Webster, Harris, Rice, Cormier, & Quinsey, 1994, p. 57). Four years later, Quinsey and colleagues (1998) pronounced a different opinion on the matter:

What we are advising is not the addition of actuarial methods to existing practice, but rather the complete replacement of existing practice with actuarial methods. This is a different view than we expressed in Webster et al 1994, where we advised the practice of adjusting actuarial estimates of risk by up to 10% when there were compelling circumstances to do so...we no longer think this practice is justifiable. Actuarial methods are too good and clinical judgment too poor to risk contaminating the former with the latter (pp. 171).

Limited research exists on the tempering of purely actuarial tools with clinical judgment. Some research shows that there is greater accuracy when clinicians rely solely on actuarial tools and avoid making ‘discretionary judgments’ (see Sawyer, 1966; Goldberg, 1968; Quinsey et al., 2006). Dawes and colleagues (1989) note that when clinicians are given free reign, they identify “too many exceptions” to the actuary (p. 1671). As Dawes and colleagues suggest, “if clinicians were more conservative in overriding actuarial conclusions they might gain an advantage, but this conjecture remains to be studied adequately” (p. 1671).

Guy (2008) conducted a comprehensive evaluation of the SPJ approach, and compared the predictive validity of SPJ and non-SPJ approaches using meta-analytic techniques based on 113 disseminations
Findings provided support for the SPJ decision-making model, indicating that structured clinical judgments performed just as well or better than SPJ tools when used actuarially, and that clinical judgments did not spoil actuarial judgments. Furthermore, differences between the two approaches were negligible, suggesting comparative predictive performance.

The process of computing a SPJ assessment is likely to be more time consuming because the clinician requires time to consider the formulation component of the assessment. Webster, Douglas, Eaves and Hart (1997) are notable advocates of the SPJ movement. They suggest that SPJ assessments are most effective when: a) assessments are carried out according to well defined schema; b) assessors have a good level of agreement, based in training, knowledge and expertise; c) outcome variables (i.e., violence) are well defined with clear follow-up periods; d) instances of violence acts are detectable and able to be recorded; e) relevant information is available and able to be substantiated; f) adjustments of actuarial estimates are well justified. At a broad level, they call for the following cornerstones of assessment: accessibility, scientific integrity, testability, administrative feasibility, and efficiency (Webster et al., 1995).

1.4.4 Statistical Advances in Violence Risk Assessment

Another significant development in violence risk assessment has been the application of new statistical methods for evaluating the accuracy of violence risk assessment tools. Violence risk assessment tools are generally evaluated clinically (i.e., using beta-testing and consumer satisfaction ratings) or empirically (i.e., reliability and validity analyses). The predictive validity
index indicates the degree to which an instrument predicts a criterion measure (e.g., violence) (Cohen & Swerdlik, 2010). Judgments of predictive validity are sometimes based on statistical evidence in the form of contingency tables (e.g., false positive and false negative cross-tabulations) and correlations between test scores and outcomes (Cohen et al., 2010; Rice et al., 1995). However, such techniques are limited by prediction errors and not including base rates of violence (Rice et al., 1995).

During the 1990's, researchers began to evaluate the accuracy of predictions using statistical methods that separated the effects of base rates and decision thresholds (Mossman, 2013). This was important because, as Mossman (1994) notes, "in many studies, prediction accuracy is quantified as the fraction of correct predictions. However, if only 5% of the subjects are violent, a clinician who always predicts "no violence" would be right 95% of the time, whereas a clinician who errs on the side of caution and predicts violence 20% of the time can be correct about no more than 85% of the subjects" (p. 783). Mossman (1994) recommended the use of the Receiver Operating Characteristic (ROC) and the Area Under Curve (AUC) it produces to overcome these limitations by establishing accuracy of predictions independent of base rates of violence and prediction errors (Mullen et al., 2009; Rice et al., 1995).

The Receiver Operating Characteristic (ROC) analysis and Area Under Curve (AUC) it produces was a novel approach to evaluating risk. As Mossman (1994; 2013) explained, the “ROC analysis recognises that diagnosticians have varying levels of confidence about whether an either/or event will occur and that proper descriptions of detection accuracy must reflect these varying levels” (p. 24). The ROC produces an Area Under the
Curve (AUC), which ranges from 0 - 1 and, in a single value, provides a gross measure of discriminant ability of the tool. Risk assessment tools with near perfect accuracy will have AUC's approaching 1 (Dolan et al., 2000). The ROC allows for the evaluation of prediction approaches independent of underlying base rates and biases for certain prediction outcomes (e.g., Type I or Type II prediction errors) (Mossman, 1994). Applications of the ROC began appearing in publications relating to psychiatry in the late 1980's (e.g., Murphy, Berwick, Weinstein, Borus, Budman, & Klerman, 1987). Specific applications of the ROC to violence began to emerge in the mid 1990's (e.g., Mossman, 1994a).

Given these advantages of the ROC were not yet known in first generation the studies of violence, Mossman (1994) conducted a meta-analysis in which he reanalyzed 58 datasets from 44 previously published studies in first and second generation violence prediction (1972-1993) using the ROC. The 44 studies represented over 16,000 participants, what Mossman considered to be a “representative sample of the different types of violence prediction studies published over the past 2 decades” (p.786). Four broad approaches to prediction were included: clinical judgment, past behaviour, retrospective and prospective discriminant functions (including rating systems). First generation approaches are represented by pre-1986 datasets, whilst second generation studies are represented by post-1986 datasets. Results showed that overall, prediction was significantly better than chance for 47 of the 58 datasets (81%). The median AUC was 0.73. The AUC’s of first and second generation approaches were 0.75 and 0.84 respectively. In terms of follow-ups, 27 of the 36 first generation datasets involved long term predictions (> 1 year), in contrast, 15
of the 22 second-generation datasets dealt with short-term (< 1 year) predictions. For 10 of the first-generation studies, prediction was no better than chance, but more recent studies were better than chance. These results indicated that clinician predictions of violence were on average, better than chance; however, AUC’s produced from second-generation studies were superior. These results challenged previous beliefs around clinical judgment, but affirmed general beliefs around the superiority of actuarial approaches. Advances in statistical methods for assessing violence risk have indicated that the discrepancy between approaches is not as large as first thought. When evaluated using the ROC, clinical judgment can perform better than chance (Mossman, 1994, 2000).

By the end of the 20th century, ROC was the analysis of choice in violence risk assessment research (Mossman, 2013). See Chapter 5 Statistical Methods, Section 5.6.1 Predictive Validity for a more detailed description of the ROC and AUC.

1.5 The Victorian Context

The role of violence risk assessment in the Victorian legal system has evolved dramatically over the past two decades. With regard to forensic mental health, prior to the 1990’s, forensic psychiatric patients were held indefinitely under the ‘Governor’s pleasure’. As the name suggests, release of patients was dependent upon the will of the Governor. In reality however, these determinations were actually made by state cabinets on behalf of the Governor (Ogloff et al., 2005). Release was therefore politically determined, often resulting in forensic psychiatric patients having longer periods of detention than those without serious mental illness who were convicted and sentenced (Ogloff et al., 2005). This was largely due to fear over false
negative judgments which have the potential to contribute to political discontent. On the other hand, false positives may have flourished because forensic psychiatric patients were never released to prove predictions wrong (Ogloff et al., 2005). In this way, clinical judgments were actually self-fulfilling prophecies; predictions were biased and influenced outcomes (Einhorn & Hogarth, 1978).

The failures of violence risk assessment not only lie in the releasing of high-risk persons, but also the detainment of low-risk persons (Douglas, Ogloff, Nicholls, & Grant, 1999). This unfounded deprivation of liberty represented a human rights issue, and states in Australia began to amend laws to correct this. These laws sought to balance public safety with the rights of the individual (Ogloff et al., 2005). The Crimes (Mental Impairment and Unfitness to be Tried) Act 1997 (Vic) (CMIA) was introduced in 1997 and replaced the Governors Pleasure system (Victorian Law Reform Commission, 2016). In Victoria, persons found Not Guilty by Reason of Mental Impairment (NGRMI) are processed under the Crimes (Mental Impairment and Unfitness to be Tried) Act 1997 (Vic). Under this, both custodial and non-custodial supervision orders are possible. These are indefinite orders that include a nominal term equivalent to the maximum prison sentence for the offence. Once the nominal term expires, review of the need for supervision (with emphasis on risk assessment) is required by law. For example, a custodial order cannot be changed to a non-custodial order “unless satisfied on the evidence available that the safety of the person subject to the order or members of the public will not be seriously endangered as a result of the release of the person” (Ss (32) 2). According to s. 40(1), regardless of any change to the order, the court must take into regard whether the released
person would ‘likely endanger themselves, another person, or other people generally’, because of their mental impairment, and must protect people from such danger. In Australia, the Victorian legislation is the most prescriptive in stating guidelines and requirements for mental health professionals to acknowledge the importance of risk assessments (Ogloff et al., 2005).

Whilst the Tarasoff decision regarding duty to warn is not legally binding in Australia, psychologists are still under legal and ethical obligations (Warren, 2006). Standard A.5.2 of the Australian Psychological Society Code of Ethics (2007) describes the ethical standard of confidentiality and requests that “Psychologists disclose confidential information obtained in the course of their provision of psychological services only under any one or more of the following circumstances… if there is an immediate and specified risk of harm to an identifiable person or persons that can be averted only by disclosing information” (p. 15). McMahon (1992) discussed the applicability of the Tarasoff ruling and duty to protect in Australia, and notes that Australian courts take a “more conservative approach than American courts and only in exceptional circumstances have they been willing to hold that a person is liable for the acts of another” (p. 15). Despite this, psychologists are still faced with the dilemma of the “competing interests of the obligation of confidentiality towards the client and public interest in disclosure” (p. 12). Although not legally binding, that is not to say that cases such as Tarasoff have not kept the anxieties of risk assessment at the forefront of the clinicians mind.

1.6 Conclusion

Over the past three to four decades, there have been considerable changes in the way violence risk is conceptualised, assessed and communicated (Mossman, 2013). Historical predictions of a dichotomous risk
outcome are now replaced by more contemporary understandings, in which risk is seen as a multi-faceted dynamic entity, where risk assessment is intertwined with risk management (Douglas & Kropp, 2002). This has subsequently affected how risk is communicated (Harris et al., 2007). Prediction has improved drastically from first generation unstructured approaches where accuracy was akin to “flipping coins” (Ennis et al., 1974, p. 693) to a third generation where structured approaches reach a level of accuracy well-above chance (Mossman, 2013). These advances have resulted in substantial improvements in clinician’s ability to assess risk and differentiate offenders in terms of their level of risk (Government of Canada, 2009). While actuarial approaches were found to have consistent superiority over unstructured clinical judgment approaches, the introduction of structured professional judgment has challenged this dominance, with predictive accuracy equivalent to actuarial approaches. Other advances such as novel approaches to evaluating violence risk assessment tools through the use of the ROC have further propelled the field.

The body of literature now clearly demonstrates that violence risk can be assessed at an accuracy that is greater than chance. The field has developed and is advanced enough that we are no longer faced with the question of “can we assess violence risk?” but rather, “how can we best assess and manage violence risk” (Douglas et al., 2013, p. 2).

Violence risk assessment is now a critical part of service delivery in a number of settings, and continues to remain high on social, political and clinical agendas (Dolan et al., 2000; Douglas et al., 2013). Psychologists and psychiatrists alike are increasingly required to assess risk of violence, oftentimes as necessitated by law (de Vogel et al., 2014; Steadman &
Cocozza, 1978). Effective violence risk assessment is essential to identifying individuals who may pose a high risk of violence, but also identifying those who won't; for those deemed to be low risk, these assessments can help reduce restrictions to liberty. SPJ and actuarial forms of assessment are now generally accepted as valid approaches to assessing violence risk, but the utility of individual tools within these approaches require evaluation. Both actuarial and SPJ tools are available, and in keeping with best practice, these tools require revision to reflect current evidence and advances in the field. The HCR-20 and VRAG are among the two most commonly used actuarial and SPJ tools, respectively (Singh et al., 2014), particularly amongst forensic psychologists (Archer, Buffington-Vollum, Stredny, & Handel, 2006). Both tools were recently revised, the HCR-20V3 in 2013, and the VRAG-R in 2015. The following chapter describes these changes.
Chapter Two

Development and Revision of the Historical-Clinical-Risk Management – 20 (HCR-20) and the Violence Risk Appraisal Guide (VRAG)

2.1 Introduction

The purpose of the current chapter is to describe the development and revision of the HCR-20\textsuperscript{V2} and VRAG and development of the HCR-20\textsuperscript{V3} and VRAG-R. The construction of tools, reasons for revision and inherent changes will be discussed. The chapter concludes by contrasting the tools, to highlight how they differ as forms of actuarial and structured professional judgment (SPJ) violence risk assessment instruments.

2.2 The Historical Clinical Risk Management – 20 (HCR-20)

As the name suggests, the HCR-20 consists of 20 violence risk factors. These factors were selected based on a comprehensive review of the violence risk assessment literature and the clinical insights of experienced forensic clinicians (Webster, Douglas, Eaves, & Hart, 1997). The risk factors are spread across three scales: Historical (H-Scale), Clinical (C-Scale) and Risk Management (R-Scale). The HCR-20 is diachronic in nature, capturing past, present and anticipated future functioning. The H-scale is concerned with past functioning, whilst the C-Scale assesses current functioning, and the R-Scale relates to future functioning. The H-Scale consists of 10 static risk factors (i.e., unchangeable historical information). This creates the foundation of the risk assessment, establishing an indication of the individual's long-term risk level. The C-Scale and R-Scale consist of five risk factors each, and provide current functioning and future functioning.
considerations through dynamic variables (i.e., changeable considerations). Being subject to change, these dynamic factors have a dual function in informing the risk assessment but also representing targets for intervention\(^1\).

This results in the HCR-20 being sensitive to changes in violence risk. The violence risk assessment is underpinned by the H-Scale, and the C-Scale and R-Scale then provide the clinical and contextual information for tempering the risk assessment further (Borum, 1996; Douglas & Webster, 1999). This clinical information relates to mental health and cognitive, affective and behavioural states. Contextual features relate to treatment, support and environmental factors that have the potential to destabilise and influence the individual’s likelihood of perpetrating violence. In scoring the R-Scale, the assessment focused on whether the person is “In” (institutional care) or “Out” (community care), which provides a context for the ratings.

2.2.1 The Original HCR-20

Published in 1995, the original HCR-20 scheme (Webster, Eaves, Douglas, & Wintrup, 1995) represented a new era in violence risk assessment through the merging of “science and practice” (Douglas et al., 1999, p. 918). Webster and colleagues described the HCR-20 as an integration of “the almost separate worlds of research on the prediction of violence and the clinical practice of assessment” (Webster et al., 1997b, p.1). Since the original publication and the release of version 2 (Webster et al., 1997), the HCR-20 has become the leading SPJ instrument, and the most widely used violence risk assessment instrument in the world (Singh et al.,

---

\(^1\) The Violence Risk Management Companion Guide (Douglas, Webster, Hart, Eaves, & Ogloff, 2001) addresses the HCR-20 risk factors in developing risk management plans.
The HCR-20\textsuperscript{V2} has been translated into 20 languages, and adopted or evaluated in more than 35 countries (Douglas et al., 2013). Its popularity is a reflection of the breadth of its applicability to males and females aged 18 and above, within institutional and community contexts, and across correctional, forensic, general psychiatric and civil psychiatric settings (Douglas et al., 2013).

2.2.2 The Historical Clinical Risk Management-20 Version 2 (HCR-20\textsuperscript{V2})

Version 2 of the HCR-20 was published in 1997 (Webster et al., 1997). The risk factors were identical to those in the original HCR-20, however item labels were altered to better reflect definitions and the type of information the item intends to capture. Table 1 below displays both the original HCR-20 and the HCR-20\textsuperscript{V2} violence risk factors across the three scales, highlighting changes in labels. See Appendix B for the HCR-20\textsuperscript{V2} scoring rubric.

\begin{table}[h]
\centering
\begin{tabular}{lll}
\hline
\textbf{Item} & \textbf{HCR-20} & \textbf{HCR-20\textsuperscript{V2}} \\
\hline
\textbf{Historical} & & \\
H1 & Previous Violence & Previous Violence \\
H2 & Age at First Violent Offence & Young Age at First Violent Incident \\
H3 & Relationship Stability & Relationship Instability \\
H4 & Employment Stability & Employment Problems \\
H5 & Alcohol or Drug Abuse & Substance Use Problems \\
\hline
\end{tabular}
\caption{Violence Risk Factors on the Original HCR-20 and the HCR-20\textsuperscript{V2} across Historical, Clinical and Risk-Management Domains}
\end{table}
### 2.2.3 The Historical Clinical Risk Management-20 Version 3 (HCR-20\textsuperscript{V3})

The HCR-20\textsuperscript{V3} (Douglas et al., 2013) represents the latest version in the HCR-20 series and builds upon the foundation laid by previous versions.
According to the authors, an updated version was required for several reasons. Firstly, as the HCR-20 is based on review of the violence risk assessment literature, updates to reflect developments in research are required. Notably, thousands of research studies on violence have been published since the release of version 2 (Douglas, Hart, Webster, Belfrage, Guy, & Wilson, 2014). With two years between the original and second version, and 16 years between the second and third versions, the need for a revision was clear. Furthermore, as a form of professional guidelines, it is imperative that the HCR-20 reflects current knowledge and best practice in assessing risk of violence (Douglas et al., 2014). Secondly, conceptual developments in the field such as increased emphasis on decision-making processes, formulation and idiosyncratic assessments; strengthening the association between assessment and prevention; and practical considerations such as scenario planning, needed to be incorporated into the tool (Douglas et al., 2014). The authors also endeavoured to provide greater guidance to users of the HCR-20 given that the tool is so heavily dependent on clinical judgment.

At a broad level, the authors noted that the goals and guiding principles for the revision were: “(1) continuity of concept; (2) exemplification and embodiment of the SPJ model; (3) practical utility; (4) enhanced clarity; (5) legal and ethical acceptability; and (6) empirical defensibility” (Douglas et al., 2014, p. 95). The revision process itself involved beta-testing and empirical testing of the tool across several countries (Douglas et al., 2014).

The “continuity of concept” and “embodiment of the SPJ approach” has meant that the HCR-20 V3 has retained the core aspects of its predecessor (Douglas et al., 2013, p.95). Douglas and colleagues (2013)
stated that “there is no real change in the overall nature or domain of information collected relevant to risk...a person with many risk factors on Version 2 will typically have many risk factors on Version 3. Similarly, it is highly unlikely that persons rated as high risk on Version 2 would subsequently be rated as low risk on Version 3” (p. 29). Therefore, the goal was to produce comparable decisions about violence risk (Douglas et al., 2014). Structurally, there are still three scales, which are spread diachronically and include 20 risk factors. There is still the same number of risk factors per scale, and risk factors continued to be rated on a three-point scale. Clinically, the tool’s outcome is still a Summary Risk Rating (SRR) of Low, Moderate or High. There is no numeric risk estimate or probability, nor any utility in Total scores or cut off scores provided for interpretation (Smith, Kelley, Rulseh, Sörman, & Edens, 2014). However, there have been key changes. Overall, seven items were broadened, four were narrowed, and one new item included (Bjorkly, Eidhammer, & Selmer, 2014). The most extensive changes were made to the R-Scale (Kötter, Franqué, Bolzmacher, Eucker, Holzinger, & Müller-Isberner, 2014). The key changes from Version 2 to Version 3 as are discussed below.

2.3 HCR-20^V3 Key Changes

Several key changes occurred in transitioning from the HCR-20^V2 to the HCR-20^V3. Outlining these changes provides the justification and basis for understanding why a comparison of the tools was investigated in the current research project. The changes are listed on the next page and discussed in turn.
• Changes to Risk Factor Labels
• Changes to Content (‘Broadening’ and ‘Narrowing’ of Risk Factors)
• Sub-items for Complex Risk Factors
• Inclusion of Relevance Ratings
• Inclusion of Item Indicators
• Changes to the Rating System
• Exclusion of the Psychopathy Checklist - Revised (PCL-R)

2.3.1 Changes to Risk Factor Labels

Extensive changes were made to risk factor labels. At the outset, a prefix was added to each scale of the HCR-20\textsuperscript{V3} to assist in guiding the clinician and clarifying the scope of the item (Douglas et al., 2013). Prefixes of “history of problems with”; “recent problems with” and “future problems with” were added to the H, C and R Scales respectively. Table 2 below summarises changes in labels from the HCR-20\textsuperscript{V2} to the HCR-20\textsuperscript{V3}. See Appendix C for the HCR-20\textsuperscript{V3} scoring rubric.

Table 2

\textit{Comparison of Items on the HCR-20\textsuperscript{V2} to HCR-20\textsuperscript{V3 Across the Three Scales}}

<table>
<thead>
<tr>
<th>HCR-20\textsuperscript{V2}</th>
<th>HCR-20\textsuperscript{V3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Historical</td>
<td>History of Problems with:</td>
</tr>
<tr>
<td>H1  Previous Violence</td>
<td>H1  Violence</td>
</tr>
<tr>
<td>H2  Young Age at First Violent Incident</td>
<td>H2  Other Antisocial Behaviour</td>
</tr>
<tr>
<td>H3  Relationship Instability</td>
<td>H3  Relationships</td>
</tr>
</tbody>
</table>

55
<table>
<thead>
<tr>
<th><strong>HCR-20^V2</strong></th>
<th><strong>HCR-20^V3</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Historical</strong></td>
<td><strong>History of Problems with:</strong></td>
</tr>
<tr>
<td>H4 Employment Problems</td>
<td>H4 Employment</td>
</tr>
<tr>
<td>H5 Substance Use Problems</td>
<td>H5 Substance Use</td>
</tr>
<tr>
<td>H6 Major Mental Illness</td>
<td>H6 Major Mental Disorder</td>
</tr>
<tr>
<td>H7 Psychopathy</td>
<td>H7 Personality Disorder</td>
</tr>
<tr>
<td>H8 Early Maladjustment</td>
<td>H8 Traumatic Experiences</td>
</tr>
<tr>
<td>H9 Personality Disorder</td>
<td>H9 Violent Attitudes</td>
</tr>
<tr>
<td>H10 Prior Supervision Failure</td>
<td>H10 Treatment / Supervision Response</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Clinical</strong></th>
<th><strong>Recent Problems with:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>C1 Lack of Insight</td>
<td>C1 Insight</td>
</tr>
<tr>
<td>C2 Negative Attitudes</td>
<td>C2 Violent Ideation or Intent</td>
</tr>
<tr>
<td>C3 Active Symptoms of Mental Disorder</td>
<td>C3 Symptoms of Major Mental Disorder</td>
</tr>
<tr>
<td>C4 Impulsivity</td>
<td>C4 Instability</td>
</tr>
<tr>
<td>C5 Unresponsive to Treatment</td>
<td>C5 Treatment / Supervision Response</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Risk Management</strong></th>
<th><strong>Future Problems with:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>R1 Plans Lack Feasibility</td>
<td>R1 Professional Services</td>
</tr>
<tr>
<td>R2 Exposure to Destabilisers</td>
<td>R2 Living Situation</td>
</tr>
<tr>
<td>R3 Lack of Personal Support</td>
<td>R3 Personal Support</td>
</tr>
<tr>
<td>R4 Noncompliance with Remediation Attempts</td>
<td>R4 Treatment / Supervision Response</td>
</tr>
<tr>
<td>R5 Stress</td>
<td>R5 Stress or Coping</td>
</tr>
</tbody>
</table>
2.3.2 Changes to Content

The purpose of changes to content was to further distinguish risk factors from one another, thereby reducing redundancy and overlap (Douglas et al., 2013). This led to a “reorganisation” of risk relevant information (Douglas et al., 2013, p. 29). A key feature of the changes was ‘broadening’ and ‘narrowing’ of risk factors; and the addition of novel information. A summary of the changes made to each risk factor is outlined below, as per the official HCR-20\textsuperscript{V3} manual (Douglas et al., 2013, p. 29-31):

**H1: Violence**

Continues to encapsulate the nature and severity of the individual’s history of violence, but was broadened to include developmental trajectory. Information captured under H2 (Young age at first violent incident) in the HCR-20\textsuperscript{V2} is now captured under this item.

**H2: Other Antisocial Behaviour**

Added to capture the nature, severity and developmental trajectory of the individual’s non-violent antisocial behaviour. In the HCR-20\textsuperscript{V2}, this information was captured through items such as H8 (Early Maladjustment); H10 (Supervision failure) and C2 (Negative Attitudes).

**H3: Relationships**

Continues to address problems with intimate relationships (as per HCR-20\textsuperscript{V2}) but was broadened to include problems with non-intimate relationships (e.g., friends, family, colleagues and neighbors).
H4: Employment
No substantial changes.

H5: Substance Use
No substantial changes.

H6: Major Mental Disorder
No substantial changes.

H7: Personality Disorder
Continues to include problems with psychopathic personality disorder (as per HCR-20\textsuperscript{V2}), but broadened to include other personality disorders (information that was previously captured under H9 (Personality Disorder) in the HCR-20\textsuperscript{V2}. PCL-R (Hare, 1991) or PCL:SV (Hare, 1998) are no longer required for scoring the presence of psychopathy.

H8: Traumatic Experiences
This risk factor was narrowed from H8 in the HCR-20\textsuperscript{V2} (Early Maladjustment) to focus on traumatic experiences specifically, but also broadened to include traumatic experiences in childhood, adolescence and adulthood. Conduct related information that was captured under H8 in the HCR-20\textsuperscript{V2} is now captured under H1 and H2 in the HCR-20\textsuperscript{V3}.

H9: Violent Attitudes
Added to capture pro-violence attitudes (i.e., attitudes that encourage or condone the use of violence. This information was previously considered as
part of C2 (Negative Attitudes) in the HCR-20\textsuperscript{V2}.

**H10: Treatment/Supervision Response**
Broadened to include problems with treatment response and responses to institutional/community supervision.

**C1: Insight**
No substantial changes.

**C2: Violent Ideation or Intent**
Narrowed from C2 in the HCR-20\textsuperscript{V2} (Negative Attitudes) to focus specifically on thoughts and plans concerning the perpetration of violence.

**C3: Symptoms of Major Mental Disorder**
No substantial changes.

**C4: Instability**
No substantial changes.

**C5: Treatment/Supervision Response**
Broadened to considerations of compliance and responsiveness to institutional and community supervision.

**R1: Professional Services**
Narrowed from the HCR-20\textsuperscript{V2} R1 (Plans Lack Feasibility) to focus specifically on planning and implementing professional services.
R2: Living Situation
Narrowed from the HCR-20\textsuperscript{V2} R2 (Exposure to Destabilisers) to focus specifically on living and accommodation considerations as a means of reducing potential destabilisation.

R3: Personal Support
No substantial changes.

R4: Treatment/Supervision Response
Continues to capture non-compliance (as per the HCR-20\textsuperscript{V2}) but broadened to include treatment resistance.

R5: Stress or Coping
No substantial changes.

2.3.3 Sub-Items for Complex Risk Factors
Sub-items for complex and multi-faceted risk factors were included to distinguish between the different aspects of an item (Douglas et al., 2013). With specific definitions provided for each sub-item, evaluators are guided in considering each of the content domains within the risk factor. Ten of the 20 risk factors included sub-items. Table 3 outlines the sub-items within these risk factors.
### Table 3

*Risk Factors and Sub-Items on the HCR-20\textsuperscript{V3}*

<table>
<thead>
<tr>
<th>Items</th>
<th>Sub-Items</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Historical</strong> History of Problems with:</td>
<td></td>
</tr>
<tr>
<td>H1 Violence</td>
<td>(H1a) under 12 Years</td>
</tr>
<tr>
<td></td>
<td>(H1b) 12-17 Years</td>
</tr>
<tr>
<td></td>
<td>(H1c) 18+ Years</td>
</tr>
<tr>
<td>H2 Other Antisocial Behaviour</td>
<td>(H2a) under 12 Years</td>
</tr>
<tr>
<td></td>
<td>(H2b) 12-17 Years</td>
</tr>
<tr>
<td></td>
<td>(H2c) 18+ Years</td>
</tr>
<tr>
<td>H3 Personal Relationships</td>
<td>(H3a) Intimate Relationships</td>
</tr>
<tr>
<td></td>
<td>(H3b) Non-intimate Relationships</td>
</tr>
<tr>
<td>H4 Employment</td>
<td>None</td>
</tr>
<tr>
<td>H5 Substance Use</td>
<td>None</td>
</tr>
<tr>
<td>H6 Major Mental Illness</td>
<td>(H6a) Psychotic Disorders</td>
</tr>
<tr>
<td></td>
<td>(H6b) Major Mood Disorders</td>
</tr>
<tr>
<td></td>
<td>(H6c) Other</td>
</tr>
<tr>
<td>H7 Personality Disorder</td>
<td>(H7a) Antisocial, Psychopathic, and Dissocial</td>
</tr>
<tr>
<td></td>
<td>(H8b) Other</td>
</tr>
<tr>
<td>H8 Traumatic Experiences</td>
<td>(H8a) Victimisation/ Trauma</td>
</tr>
<tr>
<td></td>
<td>(H8b) Adverse Childrearing Experiences</td>
</tr>
<tr>
<td>H9 Violent Attitudes</td>
<td>None</td>
</tr>
<tr>
<td>H10 Treatment or Supervision</td>
<td>None</td>
</tr>
<tr>
<td>Response</td>
<td></td>
</tr>
</tbody>
</table>
### Clinical Recent Problems with:

<table>
<thead>
<tr>
<th>Items</th>
<th>Sub-Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1 Insight</td>
<td>(C1a) Into Mental Disorder</td>
</tr>
<tr>
<td></td>
<td>(C1b) Into Violence Risk</td>
</tr>
<tr>
<td></td>
<td>(C1c) Into Need for Treatment</td>
</tr>
<tr>
<td>C2 Violent Ideation or Intent</td>
<td>None</td>
</tr>
<tr>
<td>C3 Symptoms of Major Mental</td>
<td>(C3a) Psychotic Disorders</td>
</tr>
<tr>
<td>Disorder</td>
<td>(C3b) Major Mood Disorders</td>
</tr>
<tr>
<td></td>
<td>(C3c) Other Major Mental Disorders</td>
</tr>
<tr>
<td>C4 Instability</td>
<td>None</td>
</tr>
<tr>
<td>C5 Treatment or Supervision</td>
<td>(C5a) Compliance</td>
</tr>
<tr>
<td>Response</td>
<td>(C5b) Responsiveness</td>
</tr>
</tbody>
</table>

### Risk Management Future Problems with:

<table>
<thead>
<tr>
<th>Items</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R1 Professional Services</td>
<td>None</td>
</tr>
<tr>
<td>R2 Living Situation</td>
<td>None</td>
</tr>
<tr>
<td>R3 Personal Support</td>
<td>None</td>
</tr>
<tr>
<td>R4 Treatment or Supervision</td>
<td>(R4a) Compliance</td>
</tr>
<tr>
<td>Response</td>
<td>(R4b) Responsiveness</td>
</tr>
<tr>
<td>R5 Stress and Coping</td>
<td>None</td>
</tr>
</tbody>
</table>

#### 2.3.4 Inclusion of Relevance Ratings

Relevance ratings are a key change in the HCR-20\textsuperscript{V3}, adding to a far more idiosyncratic assessment that is formulation focused. These ratings allow the assessor to consider the causal importance of risk factors to a person’s violence, in addition to the presence. Relevance in this case refers to the “extent to which the factor is critical to the evaluator’s formulation of
what caused the evaluatee to perpetrate violence and how best to prevent future violence” (Douglas et al., 2013, p.50). Relevance therefore pertains to both perpetration and prevention.

In coding relevance, a risk factor is deemed relevant to an individual’s violence if it: “a) was a material contribution to past violence; b) is likely to influence the person’s decision to act in a violent manner in the future; c) is likely to impair the individual’s capacity to employ non-violent problem solving techniques or to engage in non-violent or non-confrontational interpersonal relations; or d) it is crucial or critical to manage this factor in order to mitigate risk” (Douglas et al., 2013; p. 51). Based on the decision theory of violence, the relevance of a risk factor can also be considered in terms of its functional role: whether the risk factor motivates (“makes violence an attractive or rewarding option”), disinhibits (“makes the evaluatee less likely to be influenced by restraints, prohibitions, or proscriptions against violence”) or destabilizes (“disturbs the evaluatee’s ability to monitor and control decision making”) (Douglas et al., 2013, p.51).

Assessors can further judge the relevance of a risk factor by conducting an anamnestic assessment (see Chapter 1 section 1.4.1) or functional analysis. Functional analysis refers to the application of operant conditioning laws to understand the relationship between stimuli and responses (Skinner, 1953). This assists in determining the relevance of risk factors through identifying the purpose violence serves for the individual (Daffern & Howells, 2009).

Relevance ratings therefore represent a significant change in the traditional HCR-20 framework, emphasising that risk factors are not equally relevant to all persons who possess them (Monahan et al., 2001). By
considering both the presence and relevance of risk factors, nomothetic and idiographic levels of analysis are achieved in conducting the violence risk assessment (Douglas et al., 2013).

### 2.3.5 Inclusion of Item Indicators

Item indicators are a novel feature to the HCR-20\(^{V3}\). They provide a description of possible manifestations of the risk factor (i.e., how the risk factor might present itself). For example, Prior Supervision Failure (H10) has an item indicator of “failure to establish positive working relationships with professionals”. These descriptors assist evaluators in scoring the tool by providing examples of what may constitute the presence of an item.

### 2.3.6 Changes to the Rating System

The 3-point numeric scheme in the HCR-20\(^{V2}\) was removed and replaced with a nominal (alphabetical) scheme in the HCR-20\(^{V3}\) to enhance the clinical utility of the tool (Douglas et al., 2013). The Summary Risk Rating (SRR) is another rating within the HCR-20\(^{V3}\), a concept that has continued since the HCR-20\(^{V2}\). One novel change however is that the HCR-20\(^{V3}\) specifies three SRR’s: Case Prioritisation/ Future Violence, Serious Physical Harm and Imminence of Violence. SRR’s are gross judgments of risk of future violence and need for intervention, being Low, Moderate or High. For Case Prioritisation/ Future Violence, the Low rating indicates that the person is not in need of any special intervention or supervision strategies to manage risk; Moderate ratings indicate that the individual requires some form of management and intervention; and High ratings indicate that there is an urgent need for intervention within the risk management plan (Douglas et al.,
2013). As a SPJ tool, clinical judgment is required in rating presence, but more so relevance and the SRR. These ratings are underpinned by a formulation of the person’s risk of violence into a meaningful whole.

Formulation relates to an understanding of the person’s perpetration of violence, and generation of hypotheses around future violence risk (Douglas et al., 2013). As previously noted, common approaches to formulation include anamnestic assessments and functional analyses. Other broader models such as Weerasekera’s (1996) ‘Four P’ model for integrating information can also be utilised by considering: Predisposing factors (i.e., distal vulnerabilities); precipitating factors (i.e., acute stressors/situational contexts); perpetuating factors (i.e., maintenance) and protective factors (i.e., factors that reduce the impact of risk factors).

2.3.7 Exclusion of the Psychopathy Checklist - Revised (PCL-R)

The PLC-R (Hare, 1991) and PCL:SV (Hare, 1998) are the most commonly used inventories for assessing Psychopathy. Unlike the HCR-20V2, where assessment of psychopathy was required through use of either the PLC-R or PCL:SV, the HCR-20V3 does not require formal completion of these tools. Instead, the Psychopathy item has been replaced with a ‘Personality Disorder’ Item, which captures various personality pathologies (other diagnoses such as Antisocial Personality Disorder can be included in addition to psychopathy). Presence of personality disorder may now be based on diagnostic criteria (i.e., the DSM-V) or alternative measures. The choice to use the PCL-R or PCL:SV to assess for psychopathy is still available to clinicians should they wish.
2.4 HCR-20 Summary

The HCR-20 scheme for violence risk assessment has advanced considerably over the past two decades, and is currently in its third iteration. Much of the justification for revision has stemmed from conceptual developments in violence risk assessment, and the need for the tool to reflect contemporary practice. In the transition from the HCR-20\textsuperscript{V2} to the HCR-20\textsuperscript{V3}, several notable changes have occurred. Despite this, the HCR-20\textsuperscript{V3} remains a SPJ violence risk assessment tool where professional judgment and decision-making occur in a structured manner. The framework has remained reasonably stable across versions, enabling systematic assessment of violence risk. The HCR-20\textsuperscript{V3} continues to embody what Borum (1996, p.950) described as “the promise of (the HCR-20): Its foundation on a conceptual model or scheme for assessing dangerousness and risk; its basis in the empirical literature; its operationally defined coding system…and its practical use”. Green and colleagues (2016) have summarised these changes succinctly:

Between the publication of the previous iteration and the development of the HCR-20\textsuperscript{V3}, the focus of SPJ theory has shifted from the identification of risk factors to the formulation and application of risk management plans developed to address risk factors. Evaluators are instructed to use the item presence and relevance ratings to inform an integrated and individualised narrative, explaining past violence and assisting in proposing plausible scenarios in which examinees would engage in future violence (pp. 50)
2.5 The Violence Risk Appraisal Guide

The authors of the Violence Risk Appraisal Guide (VRAG; Quinsey et al., 2006) aimed to create "an actuarial instrument that predicted violent recidivism for serious offenders" (p. 159). The construction sample was drawn from two previous large-scale studies (Rice, Harris, Lang, & Bell, 1990; Rice, Harris, & Cormier, 1992) involving patients hospitalised at the Mental Health Centre (Oak Ridge Division) between 1965 and 1988 for pretrial psychiatric assessment or treatment (Harris, Rice, & Quinsey, 1993; Webster et al., 1994). Oak Ridge was a maximum-security psychiatric facility located in Penetanguishene (Ontario, Canada). The sample included mentally disordered offenders and those with personality disorder. The sample comprised 685 men, the majority of whom (N = 618) had the opportunity to reoffend during the average 81.5 months follow-up time. During this time, 31% violently recidivated (Quinsey et al., 2006). The outcome variable and definition of violent recidivism used in this case was the presence of a new criminal charge for a violent offence (Quinsey et al., 2006).

Approximately 50 initial predictor variables were included in the study (Quinsey et al., 2006). These variables were selected based on either existing empirical support in violent and/or criminal behaviour literature; relation to rehospitalisation in psychiatric populations; or clinical importance. Overall, these variables encompassed sociodemographic information, childhood problems, adult adjustment, index offence characteristics and psychological assessment information (Quinsey et al., 2006). Variables were coded from institutional files.

Variables were assessed to determine which would statistically discriminate between violent and non-violent recidivists. The 12 predictors
that had the strongest correlation with outcome and therefore, in combination, best predicted violent recidivism were then selected to form the VRAG. Table 4 lists these factors, as well as the correlations between items and violent recidivism in the construction sample.

Table 4

*Items on the Violence Risk Appraisal Guide (VRAG), including Pearson Correlations between Variables and Violent Recidivism*

<table>
<thead>
<tr>
<th>Item</th>
<th>Variable</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Separation from either Biological Parent by age 16</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>Elementary School Maladjustment score</td>
<td>0.31</td>
</tr>
<tr>
<td>3</td>
<td>Alcohol Problems score</td>
<td>0.13</td>
</tr>
<tr>
<td>4</td>
<td>Never Married</td>
<td>0.18</td>
</tr>
<tr>
<td>5</td>
<td>Criminal History Score for Nonviolent Offenses</td>
<td>0.20</td>
</tr>
<tr>
<td>6</td>
<td>Failure on Prior Conditional Release</td>
<td>0.24</td>
</tr>
<tr>
<td>7</td>
<td>Age at Index Offence</td>
<td>0.26</td>
</tr>
<tr>
<td>8</td>
<td>Victim Injury</td>
<td>0.16</td>
</tr>
<tr>
<td>9</td>
<td>Female Victim</td>
<td>0.11</td>
</tr>
<tr>
<td>10</td>
<td>Meets DSM-III criteria for any Personality Disorder</td>
<td>0.26</td>
</tr>
<tr>
<td>11</td>
<td>Meets DSM-III criteria for Schizophrenia</td>
<td>0.17</td>
</tr>
<tr>
<td>12</td>
<td>Psychopathy Checklist Revised (PCL-R) score</td>
<td>0.34</td>
</tr>
</tbody>
</table>

The construction of the VRAG was statistically sophisticated. Each item is scored using a weighting procedure outlined by Nuffield (1982). Within this procedure, the weight is calculated based on how much the individual differs from the base rate. The weight for each predictor variable is calculated by assigning a score of \( \pm 1 \) for each 5% difference from the mean overall violent recidivism rate (i.e., 31%) (Quinsey et al., 2006). To illustrate this, authors provide the following example: The violent recidivism rate for married
offenders was 21% (two 5% increments below 31%) whereas the violent recidivism rate for un-married offenders was 38% (one 5% increment above 31%). Therefore, in scoring the instrument, a married offender will score -2, whereas an offender who had not married will score +1. The VRAG Total score is then derived by adding up item weights. Integer weights range from -5 to +12, and Total scores can range from -26 to +38 (see Appendix D for VRAG scoring rubric).

Total scores are then divided across nine risk categories termed “Bins”. The Bins were created by dividing the range of scores (63 points) into nine equal-sized categories, each with a 7-point range (Quinsey et al., 2006). Each Bin provides a probability estimate of violent recidivism at both 7-year and 10-year intervals. These probabilities are simply the proportion of people in each Bin that had re-offended at each follow-up point. As can be seen in Table 5, the probability estimates for violent recidivism vary from 0% in Bin 1 to 100% in Bin 9. The VRAG is considered to be a pure actuarial tool because all outcomes are statistically driven (Harris & Luigio, 2007). Despite the statistical sophistication of the VRAG construction, the methodology was also limited. The sample attrition rate due to non-release is concerning given that the purpose of the study was to construct an actuarial tool. A larger sample would have improved variability. Furthermore, by only including forensic psychiatric patients in the construction sample, generalisability to criminal offenders may be restricted.
Table 5

VRAG Bin Distribution and Probability of Violent Recidivism at 7-Year and 10-Year Follow-Up Periods

<table>
<thead>
<tr>
<th>VRAG Bin</th>
<th>VRAG Scores Range</th>
<th>7 Years</th>
<th>10 Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>≤ -22</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>-21 to -15</td>
<td>0.08</td>
<td>0.10</td>
</tr>
<tr>
<td>3</td>
<td>-14 to -8</td>
<td>0.12</td>
<td>0.24</td>
</tr>
<tr>
<td>4</td>
<td>-7 to -1</td>
<td>0.17</td>
<td>0.31</td>
</tr>
<tr>
<td>5</td>
<td>0 to +6</td>
<td>0.35</td>
<td>0.48</td>
</tr>
<tr>
<td>6</td>
<td>+7 to +13</td>
<td>0.44</td>
<td>0.58</td>
</tr>
<tr>
<td>7</td>
<td>+14 to +20</td>
<td>0.55</td>
<td>0.64</td>
</tr>
<tr>
<td>8</td>
<td>+21 to +27</td>
<td>0.76</td>
<td>0.82</td>
</tr>
<tr>
<td>9</td>
<td>≥ +28</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note. In the construction sample, the 7-Year base rate of violent recidivism was 31%, the 10-Year base rate of violent recidivism was 43%.

2.5.1 The Violence Risk Appraisal Guide - Revised

The Violence Risk Appraisal Guide - Revised (VRAG-R; Harris et al., 2015) was developed because the authors wanted to combine the VRAG and the Sex Offender Risk Appraisal Guide (SORAG; Quinsey et al., 1998) into a single tool that was applicable for use with both violent offenders and sex offenders. Other practical reasons such as the tool being labour intensive and time consuming were also factors in the revision (Rice, Harris, & Lang, 2013).

The development of the VRAG-R involved a construction sample of 916 male offenders released between 1960 and 1995 and followed-up
between 2003 and 2007. A variable pool of 40+ predictor variables was available, and scored based on institutional files. Again, the 12 predictors with the strongest correlations with violent recidivism were selected to form the VRAG-R. These 12 items represent seven domains: living situation, school performance, substance use, marital status, criminal history, index offence and personality (Brown et al., 2014). Table 6 below displays these variables and their correlations with violent recidivism.

Table 6

*Items on the Violence Risk Appraisal Guide - Revised (VRAG-R), including Pearson Correlations between Variables and Violent Recidivism*

<table>
<thead>
<tr>
<th>Item</th>
<th>Variable</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lived with Both Parents</td>
<td>0.185</td>
</tr>
<tr>
<td>2</td>
<td>Elementary School Maladjustment</td>
<td>0.297</td>
</tr>
<tr>
<td>3</td>
<td>History of Drug or Alcohol Problems</td>
<td>0.222</td>
</tr>
<tr>
<td>4</td>
<td>Marital Status</td>
<td>0.123</td>
</tr>
<tr>
<td>5</td>
<td>Criminal History score for Nonviolent Offences</td>
<td>0.290</td>
</tr>
<tr>
<td>6</td>
<td>Failure on Prior Conditional Release</td>
<td>0.303</td>
</tr>
<tr>
<td>7</td>
<td>Age at Index Offence</td>
<td>0.280</td>
</tr>
<tr>
<td>8</td>
<td>Criminal History score for Violent Offences</td>
<td>0.249</td>
</tr>
<tr>
<td>9</td>
<td>Prior Admissions to Correctional Institutions</td>
<td>0.305</td>
</tr>
<tr>
<td>10</td>
<td>Conduct Disorder</td>
<td>0.295</td>
</tr>
<tr>
<td>11</td>
<td>Sex Offending History</td>
<td>0.213</td>
</tr>
<tr>
<td>12</td>
<td>PCL-R Facet 4 (anti-sociality) score</td>
<td>0.387</td>
</tr>
</tbody>
</table>

Once again, items were weighted based on differences of 5% from the mean overall violent recidivism rate. Integer weights range from -7 to +6, and Total scores range from –34 to +46 (see Appendix E for the VRAG-R scoring rubric). Total scores are derived via summation of weights, and
placed in Bins 1 – 9. These Bins provide the probability estimate of violent recidivism at both 5-year and 12-year follow-up points (see Table 7 below). This is a change from the VRAG, where probabilities of violent recidivism were provided at the 7-Year and 10-Year mark.

Table 7

*Rates of Violent Recidivism in each VRAG-R Category after 5-Years and 12-Years Fixed Periods of Opportunity*

<table>
<thead>
<tr>
<th>VRAG-R Category</th>
<th>VRAG-R Score</th>
<th>5-Years</th>
<th>12-Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt; -24</td>
<td>.09</td>
<td>.15</td>
</tr>
<tr>
<td>2</td>
<td>&gt; -23 to &lt; -17</td>
<td>.12</td>
<td>.24</td>
</tr>
<tr>
<td>3</td>
<td>&gt; -16 to &lt; -11</td>
<td>.16</td>
<td>.33</td>
</tr>
<tr>
<td>4</td>
<td>&gt; -10 to &lt; -4</td>
<td>.20</td>
<td>.42</td>
</tr>
<tr>
<td>5</td>
<td>&gt; -3 to + 3</td>
<td>.26</td>
<td>.51</td>
</tr>
<tr>
<td>6</td>
<td>&gt; +4 to &lt; +11</td>
<td>.34</td>
<td>.60</td>
</tr>
<tr>
<td>7</td>
<td>&gt; +12 to &lt; +17</td>
<td>.45</td>
<td>.69</td>
</tr>
<tr>
<td>8</td>
<td>&gt; +18 to &lt; + 26</td>
<td>.58</td>
<td>.78</td>
</tr>
<tr>
<td>9</td>
<td>&gt; 27</td>
<td>.76</td>
<td>.87</td>
</tr>
</tbody>
</table>

2.6 VRAG-R Key Changes

Table 8 below contrasts the VRAG and VRAG-R in terms of item composition. Items 1 – 7 remained relatively unchanged. Item 3 (History of Alcohol Problems) was expanded to include problems with alcohol and drugs. Five significant changes were made from item 8 onwards. Item 8 ‘Victim Injury’ and item 9 ‘Any Female Victim’ were removed as they did not retain predictive value. They were also removed because previous research had shown that they worked differently amongst sex offenders than general offenders or mentally disordered offenders (i.e., they provide no incremental
value within samples of sex offenders) (Rice & Harris, 2016). Item 10 ‘Meets DSM-III criteria for Personality Disorder’ and item 11 ‘Meets DSM-III criteria for Schizophrenia’ were both removed as they required outdated diagnostic criteria. Therefore, items 8 – 11 were replaced with: Criminal History score for Violent Offences (as per score on the Cormier-Lang Scale; Quinsey et al., 2006); Prior Admissions to Correctional Institutions; Conduct Disorder and Sex Offending. Finally, only Facet 4 of the PCL-R (Antisociality) was required, in contrast to the entire PCL-R being required in the VRAG. This represents a significant change as the PCL-R Total score was the item with the strongest correlation with violence on the VRAG.

Table 8

*Comparison of Items on the VRAG and VRAG-R*

<table>
<thead>
<tr>
<th>Item</th>
<th>VRAG</th>
<th>VRAG-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>Lived with Both Parents</td>
<td>Lived with Both Parents</td>
</tr>
<tr>
<td>Item 2</td>
<td>Elementary School Maladjustment</td>
<td>Elementary School Maladjustment</td>
</tr>
<tr>
<td>Item 3</td>
<td>History of Alcohol Problems</td>
<td>History of Alcohol or Drug Problems</td>
</tr>
<tr>
<td>Item 4</td>
<td>Marital Status</td>
<td>Marital Status</td>
</tr>
<tr>
<td>Item 5</td>
<td>Criminal History Score (non-violent offences)</td>
<td>Criminal History Score (non-violent offences)</td>
</tr>
<tr>
<td>Item 6</td>
<td>Failure on Prior Conditional Release</td>
<td>Failure on Prior Conditional Release</td>
</tr>
<tr>
<td>Item 7</td>
<td>Age at Index Offence</td>
<td>Age at Index Offense</td>
</tr>
<tr>
<td>Item</td>
<td>VRAG</td>
<td>VRAG-R</td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------------------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>Item 8</td>
<td>Victim Injury</td>
<td>Criminal History Score</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Violent Offences)</td>
</tr>
<tr>
<td>Item 9</td>
<td>Any Female Victim</td>
<td>Prior Admissions to</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correctional Institutions</td>
</tr>
<tr>
<td>Item 10</td>
<td>DSM-III - Personality Disorder</td>
<td>Conduct Disorder</td>
</tr>
<tr>
<td>Item 11</td>
<td>DSM-III criteria - Schizophrenia</td>
<td>Sex Offending</td>
</tr>
<tr>
<td>Item 12</td>
<td>Hare PCL-R score</td>
<td>Anti-sociality</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(PCL-R Facet 4)</td>
</tr>
</tbody>
</table>

### 2.7 VRAG Conclusion

The VRAG has evolved over the last 21-years from its initial use in the VPS. It has become one of the most widely used actuarial instruments in violence risk assessment (Davis, 2010; see also Singh et al., 2014). The construction procedures and structure of the VRAG and VRAG-R has remained stable throughout transitions. The VRAG-R represents a combination of the VRAG and SORAG into a single tool applicable to both violent offender and sexual offender populations (Harris et al., 2015). The VRAG-R remains true to an actuarial method of violence risk assessment, where risk factors are selected based on statistical properties, violence risk categories are created based on Total score grouping, and prediction of violence is given as a probability of violence at fixed lengths of opportunity (i.e., 5-year and 12-year follow-up points).
2.8 Contrasts between the HCR-20 and VRAG

The HCR-20\textsuperscript{V3} and VRAG-R represent SPJ and actuarial approaches to violence risk assessment. The tools are now compared in terms of applicability, subjectivity, sensitivity to changes in risk and idiosyncratic assessment.

2.8.1 Applicability

Due to differences in development (i.e., review violence risk assessment literature vs. a construction sample), the breadth of applicability for the HCR-20\textsuperscript{V3} is arguably wider than that of the VRAG-R. The HCR-20\textsuperscript{V3} is applicable to males and females aged 18 and above, within institutional and community contexts, and across correctional, forensic, general psychiatric and civil psychiatric settings (Douglas et al., 2013). In contrast, the VRAG-R is applicable to adult males from correctional and forensic psychiatric populations (Harris et al., 2015).

2.8.2 Subjectivity

Both the HCR-20\textsuperscript{V3} and VRAG-R are subjective in that they depend on human decision-making in identifying the presence of risk factors. The HCR-20\textsuperscript{V3} attracts a greater level of subjectivity through the inclusion of relevance ratings and formulation. Justification of how one arrived at a final risk judgment in the HCR-20\textsuperscript{V3} is less clear. These decision-making points can vary on presence and relevance ratings, but especially on the formulation aspect of the tool, which can vary depending on the theoretical framework adopted. Although the VRAG-R is an actuarial tool, discretion is still required.
However, these processes are more transparent as the summation of scores, resulting bins and probabilities of recidivism are entirely algorithmic.

2.8.3 Changes in Risk

A major point of differentiation in the HCR-20\textsuperscript{V3} and VRAG-R is the inclusion of static and dynamic risk factors. The HCR-20 contains both static and dynamic risk variables, enabling sensitivity to changes and fluctuations in risk. This sits well with conceptual understandings of risk as a fluid construct. McNeil and colleagues (2002) noted that even high-risk individuals are only violent for discrete period/s of time, and the timing of violence is very much influenced by situational and contextual factors (Grove & Colleagues, 2000). Risk is malleable and context based (Davis, 2010). The HCR-20 captures such elements through inclusion of dynamic risk variables, whereas the VRAG-R is context-less.

The inclusion of dynamic risk variables has a dual-function in informing targets for risk management. There is no utility in drawing on static factors for risk management as these factors are not amenable to change. The ultimate purpose of risk assessment is actually risk management (Mullen et al., 2009). A clinician can already glean important inclusions to a risk management plan whilst still in the risk assessment phase.

Indeed, sensitivity to change translates to the need for frequent assessment. The greater the risk, the more frequent assessments need to be. Authors of the HCR-20\textsuperscript{V3} recommend re-assessment every 6 – 12 months; increasing to biweekly or bimonthly for persons with Moderate SRR’s and up to weekly for persons with High SRR’s. This is reflective of the SPJ conceptualisation of violence and a changing entity that demands continuous
reassessment (Douglas et al., 2005). As the VRAG-R does not include any changeable factors, it is best able to indicate an individual’s long-term risk of violence (Rice et al., 2013; McDermott & Holoyda, 2016). Indeed, the VRAG-R’s recidivism probabilities are provided at 5-Year and 12-Year follow-up points (Harris et. al., 2015). Because the VRAG-R comprises only static factors, changes occurring between one assessment and the next are unlikely to result in major outcome disparities.

2.8.4 Idiosyncratic Assessment

Through a blend of Presence ratings, Relevance ratings and formulations, the HCR-20 tailors the risk assessment to the individual, creating a more idiosyncratic assessment. The presence of any other relevant information not already captured by the 20 risk factors is able to be integrated into the HCR-20V3 assessment through the formulation process. In contrast, the VRAG-R is limited to the 12 factors, without any scope for variation on the final risk judgment. Unlike the VRAG-R, items on the HCR-20V3 are not internally weighted because no one item is considered more important than another. However, the Relevance rating option may assist the clinician in determining the factors that are more pertinent to the individual’s perpetration of violence.

2.9 Chapter Conclusion

This chapter provided an overview of the development and revision of the HCR-20 and VRAG, to the currently available versions. The pivotal aspect of the chapter is the appreciation of how the tools differ in their approaches to violence risk assessment (i.e., SPJ and actuarial processes); how they have
transitioned from previous to current versions (i.e., what has changed and what has remained consistent), and how these changes may influence predictive validity (i.e., do administration and content changes lead to changes in overall performance). The following chapter describes the state of current research into the HCR-20 and VRAG, highlighting comparisons in performance from previous to current versions, and comparisons between the HCR-20 and VRAG themselves.
3.1 Introduction

The field of violence risk assessment is fast-growing (Brown et al., 2014) and has seen an explosion of research activity over the past two decades (Douglas, 2014). The current chapter discusses and summarises the empirical research and draws comparisons between the Historical, Clinical Risk Management – 20 (HCR-20) and the Violence Risk Appraisal Guide (VRAG), which respectively represent structured professional judgment (SPJ) and actuarial approaches to violence risk assessment. Studies on the unstructured clinical judgment approach to violence risk assessment are not discussed here, as notable research has already been outlined in Chapter 1, Introduction to Violence Risk Assessment.

This chapter focuses on the psychometric properties (i.e., predictive, concurrent and incremental validity, and inter-rater reliability) of these tools within forensic psychiatric populations, since this is the focus of the current research. The literature will be discussed as per the structuring of hypotheses in the current research (see Chapter 4 section 4.3.3). Because most studies address more than one psychometric property, relevant findings from these studies will be discussed across the applicable sections of the chapter. Because evidence on the concurrent validity and incremental validity of the HCR-20 and VRAG involve intersections between the HCR-20 and VRAG bodies of literature, the concurrent validity and incremental validity for both tools will be discussed together in section 3.3.4 Concurrent Validity, and section 3.3.5, Incremental validity.
3.2 Methodological Issues in Evaluating Violence Risk Assessment Tools

Researchers must consider and resolve several methodological issues in evaluating the accuracy of violence risk assessment tools (Monahan, 1981). There are several basic elements to designing a study for the evaluation of violence risk assessment tools. These include decisions around sample and population, how tools will be administered and scored, how violence and violent recidivism will be defined and measured, and length of follow-up periods (Douglas & Ogloff, 2003). Variations in how these basic elements are determined leads to the variation often seen in study outcomes. As a result, earlier research was reasonably fragmented and lacking in homogeneity. Monahan (1988) identified ‘unsynchronised research efforts’ as one of the problems hindering the development of the violence risk assessment field. Despite these variations, one notable consistency is the tendency for researchers to investigate violence as a binary phenomenon (i.e., did or did not occur) (Mossman, 2013).

Samples can be drawn from a range of populations (correctional, forensic psychiatric or civil psychiatric) and offender types (general offenders, violent offenders, sexual offenders). Violence risk assessment tools may be administered wholly or in part, based on archival data or interviews, in a retrospective, prospective or quasi-prospective manner. Variations in the definition of violence is a key aspect. The definition of violence adopted in different research studies can greatly effect differences in outcomes. Increasing the breadth of the definition of violence (e.g., from acts of property damage to verbal aggression and physical aggression) captures a larger spectrum of offences, and therefore, increases the proportion of those
deemed to be violent recidivists (Quinsey et al., 2006). Some definitions focus on physical harm, and therein also lies variations on what type of offences are considered violent. Some have also extended the definition of violence to include psychological harm, threats to persons and other fear-inducing behaviours (Douglas et al., 2013). As mentioned in the previous chapter, sources used to inform outcome (e.g., official reports, collateral or self-report) can also influence outcomes, and there can be even more variation if charges or convictions are the criterion. Finally, the length or follow-up is important, with immediate short-term and extreme longitudinal studies being reasonably rare (Strub et al., 2014).

3.3 Literature Search

The literature search was conducted on the following research databases: PsychINFO; EBSCOhost; PsycARTICLES; Web of Science and Swinburne University of Technology’s Psychology and Behavioural Sciences Collection. Key words included “violence”, “violent recidivism”, “violence risk assessment”, “risk assessment”, “structured professional judgment”, “actuarial”, “static”, “dynamic”, “HCR-20”, “VRAG” and “predictive validity”. The literature search extended to central texts such as the HCR-20 annotated bibliography (Douglas, Shaffer, Blanchard, Guy, Reeves, & Weir, 2014); ‘The Violence Prediction Scheme’ (Webster et al., 1994); and ‘Violent Offenders Appraising and Managing Risk’ (Quinsey, Harris, Rice, & Cormier, 1998; 2006). Administration manuals for the relevant tools were required: HCR-20 V3: Assessing risk for violence - User Guide (Douglas et al., 2013); and ‘Violent Offenders Appraising and Managing Risk – Third Edition’ (Harris et al., 2015).
3.4 Structured Professional Judgment Assessments: The Historical, Clinical, Risk-Management – 20 Literature Review

The HCR-20 has been drawn upon to assess risk for violence for just over two decades. Since publication of the original version in 1995, the HCR-20 has been empirically evaluated in more than 233 disseminations (Douglas et al., 2014). Given its recent release, limited empirical papers are available on the HCR-20V3. The author is only aware of 12 disseminations that have been released to date. The majority of these have been published in the International Journal of Forensic Mental Health in a special issue on the HCR-20V3 in 2014 (Volume 13, Issue 2). The nine papers within this issue form the foundation of research into the basic theoretical and psychometric properties of the HCR-20V3 (i.e., predictive, concurrent and incremental validity, and inter-rater reliability). One additional external paper has been located. An additional key resource for empirical evaluations of the HCR-20 is the Annotated Bibliography (Douglas et al., 2014). The HCR-20 has received widespread international research attention, being adopted or empirically evaluated in more than thirty-five countries (Douglas et. al., 2014). Few studies have been conducted in Australia involving version 1 or 2 (Abou-Sinna & Luebbers, 2012; Campbell, 2007; Campbell, Shepherd, & Ogloff (2014); Chu, Daffern, & Ogloff, 2013; Chu, Thomas, Ogloff, & Daffern, 2011; Davis, 2010; Green, Carroll, & Brett, 2010; Nanayakkara, O’Driscoll, & Allnutt, 2012; Verbrugge, Goodman-Delahunty, & Frize, 2011). Only a few of these studies have focused on evaluating the HCR-20’s psychometric properties.Version 3 has been evaluated in Canada, UK, Norway, Germany, Sweden, Netherlands, and the USA. Because the HCR-20 was developed in North
America, the majority of its evidence base is derived from American and North American populations (Douglas et al., 2014).

A large body of research has established the psychometric properties of the HCR-20 across versions 1 and 2 (Douglas et al., 2014; Smith et al., 2014). Research indicates that the HCR-20 validly predicts violence across correctional (Belfrage, Fransson, & Strand, 2000; Coid, Yang, Ullrich, Zhang, Sizmur, Roberts, Farrington, & Rogers, 2009; Douglas et al., 1999; Douglas, Yeomans, & Boer, 2005), forensic psychiatric (Grann, Belfrange, Tengström, 2000; Gray, Taylor, & Snowden, 2008; Gray et al., 2003; Strand, Belfrage, Fransson, & Levander, 1999) and civil psychiatric (Douglas et al., 1999; Nicholls, Ogloff, & Douglas, 2004) populations. It has also demonstrated efficacy across community and in-patient settings (Chu et al., 2011; de Vogel & de Ruiter, 2006; Doyle, Dolan, & McGovern, 2002; Wilson, Desmarais, Nicholls, Hart, & Brink, 2013).

In terms of special populations, the HCR-20 has been found to demonstrate predictive validity amongst females (Webster et al., 1995) and males (Douglas et al., 1999; Belfrage & Douglas, 2002). For example, Nicholls and colleagues (2004) explored the utility of the HCR-20 (version 1) for assessing risk of community violence amongst 268 involuntarily hospitalised male and female civil psychiatric patients. Outcome data was retrospectively coded based on hospital files and correctional records, and participants were followed-up for two-years on average. Results indicate that the prediction of violence was comparable between males and females, with the HCR-20 producing moderate to large effect sizes for violence and criminal offending by men (AUC = 0.67-0.75) and women (AUC = 0.66-0.83). The study was limited however by a retrospective design, a small sample size and
hence, statistical power for certain analyses (e.g., community based female physical violence). Other studies have also demonstrated equivalent predictive validity in males and female samples on HCR-20 item scores, scale scores and Total scores (Strand & Belfrage, 2001). Garcia-Mansilla, Rosenfeld and Cruise (2011) adopted data from the MacArthur Study of Mental Disorder and Violence to examine the predictive validity of the HCR-20\textsuperscript{V2} in a sample of 827 male and female civil psychiatric patients. The HCR-20\textsuperscript{V2} was retrospectively coded, and participants were interviewed every 10-weeks for one-year post-release. Results indicated no main effect for gender, but noted that the HCR-20 was slightly but non-significantly better for evaluating future risk for violence in men than women. The study was limited by a sample that was homogenous regarding psychiatric symptoms, which is likely to have impacted upon the Clinical scale. Furthermore, an abbreviated version of the HCR-20\textsuperscript{V2} was used, whereby only 14 of the 20 items were scored, based on approximations from the MacArthur data. Albeit limited research, the HCR-20\textsuperscript{V3} has also been found to show predictive validity within female samples (Green et al., 2016), and amongst another study showing that the performance of the HCR-20\textsuperscript{V3} was not moderated by gender (Strub et al., 2014). These results are consistent with a non-gendered perspective on violence risk assessment and suggest that risk factors are relevant to both sexes (Nicholls et al., 2004; Simourd & Andrews, 1994).

The HCR-20 has also demonstrated predictive validity amongst intellectually disabled populations (Gray, Fitzgerald, Taylor, MacCulloch, & Snowden, 2007; Morrissey, Hogue, Mooney, Allen, Johnston, Hollin, Lindsay, & Taylor, 2007; Lindsay, Hogue, Taylor, Steptoe, Mooney, O'Brien, Johnston, & Smith, 2008; Verbrugge et al., 2011) sex offenders (Stadtland, Hollweg,
Kleindienst, Dietl, Reich, & Nedopil, 2005) and ethnic minorities (Snowden, Gray, & Taylor, 2010).

3.4.1 Predictive Validity

The predictive validity of the HCR-20 has been evaluated at the item, scale, Total score and Summary Risk Rating (SRR) level. The vast majority of the existing body of research clearly indicates that the HCR-20 is related to violence (Douglas et al., 2014), with relatively few exceptions (e.g., Vojt, Thomson, & Marshall, 2013). Vojt and colleagues (2013) examined the predictive validity of the $HCR^{V2}$ in a sample of 109 mentally disordered male offenders from a maximum security forensic hospital. Risk assessments formed part of normal clinical practice (i.e., not conducted exclusively for research purposes). Outcome data pertained to any violent incidents (actual or attempted) and was sourced from hospital files and incident reporting systems, and official reconviction data. Self-report data was also sought, with limited responses however. As sampling was conducted on a rolling basis, follow-up time varied greatly from 1 month to > 36 months. Results indicated that, regardless of time and setting (inpatient or community), the $HCR^{V2}$ was a poor predictor of ‘all incidents’ (AUC = 0.50) and ‘minor incidents’ (AUC = 0.54), however it did predict serious violent incidents (AUC = 0.86), however this represented only three of 234 total incidents recorded, which is also a relatively low number of incidents. Researchers suggested that this may be due to lack of a confidential enquiry approach, whereby patients who declined participation were possibly the chronically violent patients, whom the research was targeting.
A few notes of caution for consumers of the research: Although the HCR-20 is designed to be used as an SPJ instrument (i.e., with Low, Moderate or High outcome ratings), the use of Total scores in research is essentially more of an actuarial undertaking. Research has suggested that the HCR-20 has moderate-to-high predictive validity for both the mechanical use of the Total score and SRR’s (Douglas, Guy, Reeves, & Weir, 2008). While the actuarial use of the HCR-20 is more common in research, of more relevance to clinical utility is the SRR’s, because the tool is adopted in practice in this manner.

3.4.1.1 Research in Forensic Psychiatric Populations

The predictive validity of the HCR-20 within forensic psychiatric populations has been demonstrated since version 1. Wintrup (1996) explored the predictive validity of the HCR-20 (version 1) in a sample of 80 male forensic psychiatric patients discharged from a maximum security forensic psychiatric institution. The HCR-20 was coded based on in-patient reports (e.g., medical, legal, psychological, psychiatric and social work reports), and outcome data were obtained from official police reports. Patients were followed-up for an average of eight years, during which the base rate of offending was 70%. Regression analyses showed that the HCR-20 Total score was predictive of various indices of violence. Correlations between HCR-20 Total scores and violence was $r = 0.20$ ($p < .01$). Interestingly, the HCR-20 was more strongly linked to forensic hospital readmissions ($r = 0.35$) and number of psychiatric hospitalisations ($r = 0.45$) than to post-release violence. However, it was unknown whether these readmissions and
rehospitalisations involved violence which would have resulted in being charged with a violent crime.

In a Swedish study, Strand and colleagues (1999) investigated the predictive validity of the HCR-20<sup>V2</sup> in discriminating between violent recidivists and non-recidivists in a sample of 40 forensic psychiatric patients. Participants were discharged from hospital between 1985 and 1994, and retrospectively followed-up for a period of 3-12 years. The HCR-20<sup>V2</sup> was scored using patient hospital files and forensic psychiatric reports. Recidivists scored significantly higher on the HCR-20<sup>V2</sup> than non-recidivists. The mean Total score was 22.39 for the non-recidivism group, and 30.77 for the recidivism group. All participants with HCR-20<sup>V2</sup> Total scores of 34 and greater recidivated violently. The HCR-20<sup>V2</sup> Total scores demonstrated high predictive validity (AUC = 0.80). The HCR<sup>V2</sup> “worked perfectly” in predicting violence for those who scored very high, however it was imperfect in predicting violence for the “middle-group scorers” and also some low scoring individuals (p.74).

Gray and colleagues (2008) used a pseudo-prospective method to evaluate the predictive validity of the HCR-20<sup>V2</sup> within a sample of 887 forensic psychiatric patients in the UK. The HCR-20 was completed using only pre-discharge data. Participants were followed-up for 6-months, 1-year, 2-years and 5-years. Offending behavior was obtained from official records (‘violent offences’ and ‘any offences’). Results indicated that the HCR-20<sup>V2</sup> Total score was a good predictor of post-release violence and general recidivism. For violent offences AUC’s ranged from 0.70 – 0.76, varying as a function of follow-up. For general recidivism, AUC’s ranged from 0.69 – 0.75. All AUC’s for HCR-20 Total score were significant at the .001 level. Predictive
accuracy tended to be highest for shorter follow-up periods, and gradually declined as follow-up periods extended. For example, AUC’s for violent and general offending were 0.76 and 0.75 at the six-month follow-up point, which reduced to 0.71 and 0.69 at the 2-year follow-up point, and eventually 0.70 and 0.69 at the 5-year follow-up point. This was a statistically significant trend. The authors concluded that the HCR-20$^{V2}$ was a good predictor of both general and violent offending in forensic psychiatric samples.

Dolan and Blattner (2010) evaluated the predictive validity of the HCR-20$^{V2}$ in forensic psychiatric patients transferred from high to medium levels of security prior to community release in the UK. The study differed from traditional follow-up studies by evaluating the HCR-20$^{V2}$’s predictive value in relation to decisions to transfer patients between institutions of varying levels of security. The HCR-20$^{V2}$ was scored from case file data. Community follow-up lasted for 6 years on average. Successful outcomes were defined as “successful rehabilitation from the MSU [medium secure unit] to the community with no adverse events” (p. 4). Failure was defined as “1) direct return to the HSPH [High Security Psychiatric Hospital], 2) return to the HSPH after discharge to the community, and 3) reconviction for a serious offence after discharge to the community” (p.4. The HCR-20$^{V2}$ Total score was a robust predictor of outcomes (AUC = 0.86). The study was novel in taking the level of security at the point of discharge point into account, but was also limited by a small sample size.

The predictive validity of the HCR-20 can be assessed in regards to both Total scores and SRR. Douglas, Ogloff and Hart (2003) explored the predictive validity of structured clinical risk ratings on the HCR-20$^{V2}$. This represents an important question in the literature because it evaluates the
validity of the HCR-20$^{\text{V2}}$ in its intended clinical use (i.e., using SRR’s as opposed to Total scores). Using a pseudo-prospective design, the sample included 100 forensic psychiatric (that is, persons found Not Guilty by Reason of Mental Impairment, (NGRMI)) who were discharged to community from a maximum security facility. Participants were discharged between 1996 and 1997, and followed-up in the community until 2001. Violent outcomes were file-based and measured through several sources, including official criminal records and clinical files regarding re-admissions to hospitals (including forensic psychiatric hospitals). Three categories of violent outcomes were considered: Any violence, physical violence and non-physical violence. AUC values ranged from 0.67 to 0.70 for the HCR-20$^{\text{V2}}$ Total score, and 0.68 – 0.74 for the SRR’s. The highest observed AUC was for SRR’s for physical violence (AUC = 0.74, $p < .01$). These results indicated that the HCR-20$^{\text{V2}}$’s SRR’s significantly predicted post-release community violence, performing better than Total scores across all three outcome categories. When AUC values were converted to Cohen’s $d$, this indicated moderate to large effect sizes. Survival analyses demonstrated that participants deemed High risk were indeed more likely to be violent, and to perpetrate violence sooner than those who were not. The study was limited by a pseudo prospective design and reliance on file review to complete assessments.

3.4.1.2 Civil Psychiatric Populations and Mixed Samples

The HCR-20 also has demonstrated predictive validity in civil psychiatric populations. Douglas and colleagues (1999) evaluated the predictive validity of the HCR-20 (version 1 and 2) in a sample of 193 civil psychiatric patients in Canada (mixed gender sample). They conducted a
retrospective file-review study, and followed participants for an average of 626 days. Four types of violence were considered: Any violence, physical violence, threatening behaviour and criminal violence (i.e., being arrested or convicted of a violent criminal offence). Results indicated that the HCR-20 was predictive of violence across all four categories: Any violence (AUC = 0.76), physical violence (AUC = 0.76), threatening behaviour (AUC = 0.77) and criminal violence (AUC = 0.80). All AUC’s were significant at the $p \leq 0.5$ level.

They also found that participants scoring above the HCR-20 median score were six times more likely to commit violent acts in the community (across the three violence categories: Any violence, physical violence and threatening behaviour). For the fourth category (criminal violence), those scoring above the HCR-20 median score were 13 times more likely to commit violent acts in the community. The odds of violence were also provided for the HCR-20 scales (see Table 9). As HCR-20 scores increased, so did the proportion of violence across the four outcome types. For those who scored below the median of the HCR-20, 80% remained violence free. Of those who scored above the median, only 40% remained violence free. This is similar to results reported by Douglas and Webster (1999) who found that persons scoring above the median on the HCR-20 was approximately five times more likely than those scoring below to have been charged with a violence offence in the past. Although the study was limited by a small sample and relatively restricted follow-up period (less than two years), it supported the use of the HCR-20 within civil psychiatric populations.
Table 9

Odds of Violence based on four Categories of Violent Outcomes and HCR-20 Scores derived from Douglas, Ogloff, Nicholls and Grants (1999)

<table>
<thead>
<tr>
<th>Recidivism Type</th>
<th>HCR-20</th>
<th>H-Scale</th>
<th>C-Scale</th>
<th>R-Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any Violence</td>
<td>5.86**</td>
<td>3.63**</td>
<td>1.87*</td>
<td>3.34**</td>
</tr>
<tr>
<td>Physical Violence</td>
<td>6.14**</td>
<td>3.92**</td>
<td>1.26</td>
<td>4.44**</td>
</tr>
<tr>
<td>Threatening Behaviour</td>
<td>5.79**</td>
<td>3.74**</td>
<td>2.08*</td>
<td>3.20**</td>
</tr>
<tr>
<td>Criminal Violence</td>
<td>13.25**</td>
<td>6.77**</td>
<td>2.25</td>
<td>3.14*</td>
</tr>
</tbody>
</table>

* Significant at p ≤ .05; Significant at p ≤ .001.

Garcia-Mansilla and colleagues (2011) utilised data from the MacArthur Study of Mental Disorder and Violence to retrospectively score a component of the HCR-20V2 (the Historical and Clinical scales) in a sample of 827 civil psychiatric patients. Due to restrictions in the availability of data, the R-Scale and Item 10 of the H-Scale ('prior supervision failure') were excluded. Over a 20-week follow-up, incidents of community violence (specifically, physical harm) were recorded based on official records, self-report and collateral information. Given the problems related to relying on self-report data only (i.e., false reporting, low base rates), drawing on several sources of outcome data was a strength of the study. The HC-14 Total score AUC was 0.66, p < .001. This study demonstrated that the HCR-20V2 still produced acceptable levels of accuracy when the R-Scale was excluded.

Guy’s (2008) meta-analytic review of SPJ approaches to assessing risk of violence (as a form of anti-social behaviour) involved a more comprehensive study with mixed samples that reported outcomes for both Total scores and SRR's. The review was based on 113 disseminations (104 independent samples). The HCR-20V2 was assessed in addition to other SPJ
tools, such as the Structured Assessment of Violence Risk in Youth (SAVRY, Borum, Bartel, & Forth, 2003). The mean AUC across SPJ instruments was 0.66. The mean AUC was 0.68 for both Total scores and final summary judgments.

Across the range of outcomes, AUC’s for Total scores and final risk judgments respectively were: Antisocial behaviour (AUC = 0.68; AUC = 0.69), violent behaviour (AUC = 0.74; 0.71), physical violence (AUC = 0.68; AUC = 0.77) and sexual violence (AUC = 0.60; AUC = 0.65). The HCR-20 V2’s performance on gross measures for violent outcomes were as follows: Total score AUC = 0.73, SRR AUC = 0.76. For physical violence outcomes (including sexual violence), the AUC’s were 0.67 and 0.79 for Total scores and SRR’s respectively. The largest effect size was found for the HCR-20 SRR with physical violence (including sexual violence), AUC = 0.79. Results indicated that both the HCR-20 V2 Total score (AUC = 0.73 - 0.67) and HCR-20 V2 SRR (AUC= 0.76 - 0.79) predicted violent behaviour and physical violence (including sexual violence). Furthermore, SRR’s tended to perform better than Total scores, and were found to add incremental validity.

3.4.1.3 Australian-based Research

Research by Campbell (2007) remains the only known study in Australia that has evaluated the predictive validity of the HCR-20 V2 in a follow-up community setting amongst the forensic psychiatric patients. The sample comprised of 116 mentally disordered offenders (MDO’s) from the Thomas Embling Hospital, a secure forensic mental health facility in Victoria. Recidivism data were recorded for both general and violent recidivism, based on official police records noting first post-release convictions. Participants
were followed-up in the community post-discharge for an average of 343 days. During this time, 45.7% of the sample were reconvicted, 25% for a violent reconviction. Results indicated that the HCR-20\textsuperscript{V2}, as coded at discharge, produced moderate to strong AUC’s for both types of reconviction: Total score for general recidivism (AUC = 0.76, \( p < .05 \)); Total score for violent recidivism (AUC = 0.68, \( p < .05 \)). Significant prediction for violent recidivism was found across the three different categories of violence (any violence, fear-inducing violence and physical violence), all within the AUC = 0.64 – 0.68 range.

Campbell (2007) also found that participants scoring above the median on the HCR-20\textsuperscript{V2} were significantly more likely to be reconvicted of an offence than those scoring below. In her sample, participants scoring above the median score were more than six times more likely to be reconvicted (OR = 6.68, \( p < .01 \)). Across the three categories of violence, those scoring above the median demonstrated a clear increase in the likelihood of violence: Any violence (OR = 3.89, \( p < .01 \)), fear-inducing violence (OR = 3.92, \( p < .01 \)) and physical violence (OR = 2.55, \( p = .05 \)). These results further supported the internationally established relationship between the HCR-20 and violent recidivism within a community context amongst forensic psychiatric populations in Australia.

3.4.1.4 The Historical Clinical Risk Management - 20 (Version 3)

The IJFMH special issue on the Historical Clinical Risk Management - 20 (Version 3; HCR-20\textsuperscript{V3}) provides the core preliminary evidence base for the HCR-20\textsuperscript{V3}. Two articles in this edition focused specifically on predictive validity: Doyle, Power, Coid, Kallis, Ullrich and Shaw (2014) and Strub,
Douglas and Nicholls (2014). In addition, de Vogel and colleagues (2014) provided initial commentary and data on clinical beta-testing in the draft phases of the revision. External research by Green and colleagues (2016) further added to this body of research, by demonstrating that the HCR-20\(^{V3}\) scales predicted classification of violence in institutional settings amongst forensic psychiatric patients.

Pilot research into the HCR-20\(^{V3}\) by de Vogel and colleagues (2014) examined the predictive validity of the draft version of the HCR-20\(^{V3}\) and also compared it to the performance of the HCR-20\(^{V2}\). Using a retrospective file review, researchers followed-up on 86 forensic psychiatric patients convicted of non-sexual violent offences and discharged from the Van der Hoeven Kliniek (a forensic psychiatric hospital in the Netherlands) between 1990 and 2006. The predictive validity outcomes were reported at one-year, two-year and three-year follow-up periods, with violent recidivism as the outcome. Total score comparisons at one-year, two-year and three-year follow-up points were as follows: HCR-20\(^{V3}\) (AUC = 0.77, 0.75 and 0.67) and HCR-20\(^{V2}\) (AUC = 0.80, 0.74 and 0.67). All AUC’s were significant at the \(p < .05\) level. AUC’s were then compared based on a method described by DeLong, DeLong and Clarke-Pearson (1988), and were found to be not significantly different from each other. SRR’s (Case Prioritisation/ Future Violence) for the HCR-20\(^{V3}\) were also provided at one-year, two-year and three-year follow-up points (AUC’s = 0.72, 0.67 and 0.64 respectively). Interestingly, when a 5-point SRR scale was used (Low, Low-Moderate, Moderate, Moderate-High and High) as opposed to the conventional 3-point scale, larger AUC’s were found at one-year, two-year and three-year follow-up points (AUC’s = 0.82, 0.74 and 0.71 respectively). Indeed when compared, AUC’s produced by the
5-point scale were significantly better than AUC’s produced from the 3-point scale at the one-year follow-up point.

A valuable addition to the study was a survey on user-satisfaction and clinical value of the final version of the HCR-20\textsuperscript{V3}. 192 practitioners who attended HCR-20\textsuperscript{V3} training workshops completed the survey. The vast majority of responses were positive and indicated that the new tool was clinically useful, and added value to risk assessment and management. This study represents one of the first studies to compare the HCR-20\textsuperscript{V2} to the HCR-20\textsuperscript{V3}, and provide insights into clinical utility. However, the study was limited by a retrospective design and a small and homogenous sample. Obviously, the study also provides initial data on the draft version of the tool.

Doyle and colleagues (2014) evaluated the HCR-20\textsuperscript{V3}’s predictive validity and association with frequency of community violence in a sample of 387 forensic psychiatric patients in England and Wales. The study utilized a prospective confidential inquiry approach in which the consent of participants was not sought based on the notion that non-consenting persons are actually those at greater risk of non-compliance and anti-sociality, and hence, violence. Patients were sourced from 32 medium-security units, from where they had been discharged between 2010 and 2011 to community and non-forensic placements (e.g., supported accommodation, open rehabilitation/psychiatric units – which were included under the ‘community’ umbrella). Participants were followed-up for 12-months post-discharge. During this time, the HCR-20\textsuperscript{V3} was completed at 6-month and 12-month time points, based on clinical records and interviews with social supervisors and/or care coordinators who were familiar with the patient. Violence included sexual assaults, acts of battery and assaultive acts or threats made with a weapon;
verbal threats or abuse were not included. It was hypothesised that participants with higher Total and scale scores would be significantly more likely to exhibit violence at both 6-month and 12-month follow-up points. Incidents of violence were recorded through clinical and police records, and collateral (interviews with staff).

Outcomes indicated that 14% of the sample had committed a violent act by 6-months, and 22.5% had committed a violent act by 12-months. The HCR-20 V3 scale scores and scale scores were both strongly correlated with the frequency of violence. Correlations between post-discharge violence and the HCR-20 Total and scale scores as 6-months were as follows: HCR-20 V3 Total $r = 0.231$, H-Scale $r = 0.144$, C-Scale $r = 0.218$ and R-Scale $r = 0.181$. At 12-months, the correlation between post-discharge violence and the HCR-20 Total and scale scores were as follows: HCR-20 V3 Total $r = 0.243$, H-Scale $r = 0.139$, C-Scale $r = 0.243$ and R-Scale $r = 0.194$. All correlations were significant at the $p > .001$ level, except for the H-Scale scores, which were significant at the $p < .01$ level. As can be seen, the HCR-20 V3 had the strongest association with violence at both follow-up points, and the C-Scale had an identical association at the 12-month point.

The HCR-20 V3 Total score and scale scores all significantly predicted violence at both follow-up points. AUC values for the HCR-20 V3 Total scores at the 6-month follow-up were 0.73 and 0.70 at 12-months. Furthermore, participants in the violent group scored significantly higher on the HCR-20 V3 Total score and scale scores at both follow-up points. Those scoring higher than the median ($Md = 23$) were 2-5 times more likely to be violent at 6-month and 12-month follow ups than those scoring below the median. The authors concluded that the HCR-20 V3 was associated with frequency of violence and
moderately predictive of post-discharge violence (at levels comparable to the HCR-20\textsuperscript{V2}), and discriminated well between violent and non-violent participants.

Strub and colleagues (2014) focused on the predictive validity of the HCR-20\textsuperscript{V3} SRR’s and Relevance ratings in a mixed sample of 106 civil psychiatric patients and criminal offenders transitioning from institution to community in Canada. Using a prospective method, data collection involved interviewing participants prior to their release into the community or once selected in the community, and at five subsequent follow-up points (maximum). Correctional and psychiatric files were also reviewed at these points. Violence was defined as per the official HCR-20 operational definition. Outcomes were provided for the short term (4-6 weeks) and long term (6-8 months).

The base rate of violence in the short-term was 15.8%, and 34% in the long-term. In the short-term, 2% of those rated Low risk were violent, compared to 16% of those rated Moderate and 44% of those rated High. In the long-term, 16% of those rated Low were violent, compared to 36% of those rated Moderate and 67% of those rated High.

Comparisons of HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} at both follow-up points revealed that both were significantly correlated with violence: at 4 – 6 weeks HCR-20\textsuperscript{V2} $r = 0.36$, $p < .0009$ and HCR-20\textsuperscript{V3} $r = 0.38$, $p < .0006$. At 6-8 months: HCR-20\textsuperscript{V2} $r = 0.46$, $p = 0.0007$; and HCR-20\textsuperscript{V3} $r = 0.46$, $p = 0.0005$. For SRR’s correlations with violence were $r = 0.43$ ($p = .00003$) at 4 - 6 weeks, and $r = 0.41$ ($p = .00006$) at 6-8 months.

Predictive validity results are displayed in Table 10. As can be seen, AUC’s ranged from 0.63 – 0.81 overall and were significant across all
predictors (Presence ratings, Relevance ratings and SRR’s) and follow-up periods. The highest AUC was observed for the SRR. Across the scales of the HCR-20\textsuperscript{V3} Presence ratings were always more predictive than Relevance ratings although values across the three scale and Total scores were reasonably similar. Findings indicated that SRR’s, Presence ratings and Relevance ratings were all predictive of community violence in both the short term and long term. The study is important because of its truly prospective design and approximately equal numbers of males and females.

Table 10

<table>
<thead>
<tr>
<th>HCR-20\textsuperscript{V3} Indices</th>
<th>AUC at 4-6 weeks</th>
<th>AUC at 6-8 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Presence</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-Scale</td>
<td>0.76***</td>
<td>0.73***</td>
</tr>
<tr>
<td>C-Scale</td>
<td>0.71**</td>
<td>0.71***</td>
</tr>
<tr>
<td>R-Scale</td>
<td>0.74**</td>
<td>0.75***</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V3} Presence Total</td>
<td>0.78***</td>
<td>0.77***</td>
</tr>
<tr>
<td><strong>Relevance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-Scale</td>
<td>0.72**</td>
<td>0.67**</td>
</tr>
<tr>
<td>C-Scale</td>
<td>0.67*</td>
<td>0.69***</td>
</tr>
<tr>
<td>R-Scale</td>
<td>0.66*</td>
<td>0.63*</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V3} Relevance Total</td>
<td>0.71**</td>
<td>0.68**</td>
</tr>
<tr>
<td><strong>SRR</strong></td>
<td>0.81***</td>
<td>0.73***</td>
</tr>
</tbody>
</table>

*Note. Based on the combined psychiatric and offender sample. * $p < .05$, ** $p < .01$, *** $p < .001$. \[\]
3.4.1.5 Conclusion

From a plethora of studies available on the HCR-20, a few select studies focusing on predictive validity of the tool across versions 1 to 3 have been discussed. These papers represent an extensive body of literature establishing the predictive validity of the HCR-20 across iterations (Douglas et al., 2014). Although evidence supports its utility in predicting violent recidivism across a range of populations and settings, the main focus herein is its performance within forensic psychiatric populations within community settings, as this is the context of the current research. Only two Australian studies are available evaluating the performance of version 2 of the HCR-20 in forensic psychiatric settings including community follow-up. Results are generally consistent, indicating the HCR-20’s ability to predict future violence. The observed AUC’s tend to be in the acceptable to good range. Table 11 below provides a summary of the results of the aforementioned studies.

Table 11
Summary of HCR-20, HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} Predictive Validity Studies

<table>
<thead>
<tr>
<th>Authors</th>
<th>Sample</th>
<th>Country</th>
<th>Version</th>
<th>AUC/ ( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nicholls, Ogloff,</td>
<td>Civil</td>
<td>Canada</td>
<td>HCR-20\textsuperscript{V1/2}</td>
<td>0.67 - 0.75\textsuperscript{4}</td>
</tr>
<tr>
<td>&amp; Douglas (2004)</td>
<td>Psychiatric</td>
<td></td>
<td></td>
<td>0.66 – 0.83\textsuperscript{5}</td>
</tr>
<tr>
<td>Wintrup (1996)</td>
<td>Forensic</td>
<td>Canada</td>
<td>HCR-20\textsuperscript{V1}</td>
<td>( r = 0.20 )</td>
</tr>
<tr>
<td>Strand, Belfrage,</td>
<td>Forensic</td>
<td>Sweden</td>
<td>HCR-20\textsuperscript{V2}</td>
<td>0.80</td>
</tr>
<tr>
<td>Fransson, &amp; Levander</td>
<td>Psychiatric</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1999)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Sample</td>
<td>Country</td>
<td>Version</td>
<td>AUC/ r</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------------</td>
<td>--------------</td>
<td>----------</td>
<td>------------</td>
</tr>
<tr>
<td>Gray, Taylor, &amp; Snowden (2008)</td>
<td>Forensic</td>
<td>UK</td>
<td>HCR-20(^{v2})</td>
<td>0.70 – 0.76</td>
</tr>
<tr>
<td>Dolan &amp; Blattner (2010)</td>
<td>Forensic</td>
<td>UK</td>
<td>HCR-20(^{v2})</td>
<td>0.86</td>
</tr>
<tr>
<td>Douglas, Ogloff, &amp; Hart (2003)</td>
<td>Forensic</td>
<td>Canada</td>
<td>HCR-20(^{v2})</td>
<td>0.67 – 0.70(^{2})</td>
</tr>
<tr>
<td>Douglas, Ogloff, Nicholls, &amp; Grant (1999)</td>
<td>Civil</td>
<td>Canada</td>
<td>HCR-20 &amp; HCR-20(^{v2})</td>
<td>0.76 - 0.80</td>
</tr>
<tr>
<td>Garcia-Mansilla, Rosenfeld, &amp; Cruise (2011)</td>
<td>Civil</td>
<td>USA</td>
<td>HC(^{v2})-15(^{b})</td>
<td>0.66</td>
</tr>
<tr>
<td>Guy (2008)</td>
<td>Mixed</td>
<td>Intl. Sample</td>
<td>HCR-20(^{v2})</td>
<td>0.73(^{2})</td>
</tr>
<tr>
<td>Campbell (2007)</td>
<td>Forensic</td>
<td>Australia</td>
<td>HCR-20(^{v2})</td>
<td>0.64 - 0.68</td>
</tr>
<tr>
<td>de Vogel, van den Broek, &amp; de Vries (2014)</td>
<td>Forensic</td>
<td>Netherlands</td>
<td>HCR-20(^{v2})</td>
<td>0.67 – 0.80</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>HCR-20(^{v3})</td>
<td>0.67 – 0.77(^{2})</td>
</tr>
<tr>
<td>Robbé (2014)</td>
<td></td>
<td></td>
<td></td>
<td>0.64 – 0.72(^{3})</td>
</tr>
<tr>
<td>Doyle, Power, Coid, Kallis, Ullrich, &amp; Shaw (2014)</td>
<td>Forensic</td>
<td>UK</td>
<td>HCR-20(^{v3})</td>
<td>0.70 – 0.73</td>
</tr>
<tr>
<td>Authors</td>
<td>Sample</td>
<td>Country</td>
<td>Version</td>
<td>AUC/ r</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------</td>
<td>-----------</td>
<td>-------------</td>
<td>--------</td>
</tr>
<tr>
<td>Strub, Douglas,</td>
<td>Mixed Civil</td>
<td>Canada</td>
<td>HCR-20V3</td>
<td>0.78²</td>
</tr>
<tr>
<td>&amp; Nicholls</td>
<td>&amp; Forensic</td>
<td></td>
<td></td>
<td>0.81³</td>
</tr>
<tr>
<td>(2014)</td>
<td>Psychiatric</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. ¹ HCR-20 refers to the original HCR-20 (Version 1); ² Based on Total scores; ³ Based on SRR’s; ⁴ sample exclusively of males; ⁵ sample exclusively of females. ⁶ Actually represents H-14 because ‘prior supervision item’ was excluded from H-Scale. Intl = International study.

### 3.4.2 Strength of Scales

Although the HCR-20 as a whole (across Total scores and summary ratings) has established predictive validity, there is some variability amongst studies regarding the strength of scales. In an overview of the research, Otto and Douglas (2010) noted that AUC’s for HCR-20 scales ranged from 0.40 - 0.85 for the H-Scale, 0.46 – 0.80 for the C-Scale and 0.48 to 0.88 for the R-Scale. The general consensus is that the predictive validity of the HCR-20 is strongly underpinned by the Historical scale, the static baseline measure of violence (Douglas et al., 2014; Webster et al., 1994). However, Smith and colleagues (2014) have noted that while the contribution of the H-Scale has been considered superior in psychiatric settings, the C-Scale and R-Scale have improved upon the prediction of static factors alone (see Campbell et al., 2009; Doyle et al., 2006; McDermott, Edens, Quanbeck, Busse, & Scott, 2008).

Early research by Wintrup (1996) on the HCR-20 (version 1) indicated that correlations between the Historical, Clinical and Risk Management scales and violence was \( r = 0.25, r = -0.02, \) and \( r = .08 \) respectively, with only the
Historical scale being significant. Douglas and Webster (1999) also found that, regarding the HCR-20 version 1, the Historical scale was very strongly related to various indices of past violence, whereas the Clinical scale related less strongly and less consistently. The combined Historical Clinical scale was intermediate.

In Campbell’s (2007) Australian research on the HCR-20\textsuperscript{V2}, the Historical scale significantly predicted post-release general reconviction (AUC = 0.78, \( p < .05 \)) and violent reconviction (AUC = 0.66, \( p < .05 \)). This was for ‘any violence’, whereas the H-Scale failed to significantly predict outcomes in the more specific physical violence and fear-inducing violence categories. The Risk Management scale significantly predicted post-release reconvictions (AUC = 0.71, \( p < .05 \)) and post-release violence (AUC = 0.67, \( p < .05 \)), across all three different types of violence (AUC’s ranging from 0.64 – 0.67, all significant). The Clinical scale failed to significantly predict general nor violent recidivism, with AUC’s in the near chance prediction range (AUC’s = 0.56 – 0.59, all non-significant). The authors noted that the non-significance of the Clinical scale may be related to its dynamic nature and coding upon discharge, when clinical symptoms are probably most stable. Given the nature of the sample and setting (forensic psychiatric patients, some of whom being in acute care), the homogeneity of the sample regarding clinical scale items (e.g., symptoms of mental illness) may have affected the ability of the C-Scale to emerge as significant in the analyses.

Garcia-Mansilla and colleagues (2011) examined the historical and Clinical scales of the HCR-20\textsuperscript{V2} in a civil psychiatric sample. Over a 20-week follow-up identifying incidents of community violence, the predictive accuracy of the Historical scale was significantly higher than for the Clinical scale. The
Historical scale AUC was 0.68, \( p < .001 \). This is actually based on nine items of the Historical scale, because H10 (Prior Supervision Failure) was excluded. The Clinical scale was much worse, and in fact was no better than chance (AUC = 0.54, \( p = 0.17 \)). The performance of the HC-14 scale was intermediate (AUC = 0.66, \( p < .001 \)). These results are consistent with findings by Douglas and Webster (1999) on the HCR-20 Version 1. The authors rationalised that, although the Clinical scale demonstrated poor predictive validity, this may be a reflection of it being scored at intake as opposed to discharge, where mental state may differ. The authors also noted that the impacts of archival data collection is greater on current clinical items than historical items.

Other researchers have also observed poorer predictive validity for the Clinical scale. In the previously mentioned study by Gray and colleagues (2008), the predictive validity of the HCR-20\(^{V2}\) was evaluated in a sample of forensic psychiatric patients in the UK. In addition to providing the AUC’s for the HCR-20\(^{V2}\) Total score over a staggered 0.5 Year – 5 Year period, the authors also reported AUC’s at a scale level. Table 12 below summarises the outcomes, which were reported for general recidivism and violent recidivism.

Table 12

<table>
<thead>
<tr>
<th>Recidivism Type</th>
<th>Follow-up Period</th>
<th>AUC: H-Scale</th>
<th>C-Scale</th>
<th>R-Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violent</td>
<td>0.5 Year</td>
<td>0.77**</td>
<td>0.61</td>
<td>0.69*</td>
</tr>
<tr>
<td></td>
<td>1 Year</td>
<td>0.76**</td>
<td>0.61</td>
<td>0.68**</td>
</tr>
<tr>
<td></td>
<td>2 Years</td>
<td>0.71**</td>
<td>0.54</td>
<td>0.65**</td>
</tr>
<tr>
<td></td>
<td>5 Years</td>
<td>0.68**</td>
<td>0.57</td>
<td>0.63**</td>
</tr>
<tr>
<td>Recidivism Type</td>
<td>Follow-up Period</td>
<td>AUC: H-Scale</td>
<td>C-Scale</td>
<td>R-Scale</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------</td>
<td>-------------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>General</td>
<td>0.5 Year</td>
<td>0.75**</td>
<td>0.55</td>
<td>0.68**</td>
</tr>
<tr>
<td></td>
<td>1 Year</td>
<td>0.69**</td>
<td>0.51</td>
<td>0.69**</td>
</tr>
<tr>
<td></td>
<td>2 Years</td>
<td>0.69**</td>
<td>0.51</td>
<td>0.67**</td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>0.70**</td>
<td>0.52</td>
<td>0.66**</td>
</tr>
</tbody>
</table>

*Note. H-Scale = Historical Scale, C-Scale = Clinical Scale, R-Scale = Risk Management Scale. Significant at \( p < .001 \) level. Significant at \( p < .05 \) level.*

As can be seen from Table 12, the Historical scale demonstrated superior or equal predictive validity at each follow-up point, for both general and violent recidivism. The Historical scale had the highest AUC (AUC = 0.68 - 0.77), being the highest for the violent recidivism sample. This pattern of results closely matched the pattern of AUC’s for HCR-20\(^{V2}\) Total scores (see previous description). This was followed by the Risk Management scale (AUC = 0.63 - 0.69) and the Clinical scale (AUC = 0.51 - 0.61). The Clinical Scale was the only scale that did not significantly predict general or violent recidivism. The general trend was for AUC’s to decline as time extended, with few exceptions. This is consistent with the HCR-20\(^{V2}\) being conceptualised as a dynamic tool (i.e., as time progresses, variations to changeable factors reduce the relevance of the original assessment and hence accuracy of the tool). The only exception to this is the Historical scale, which represents stable historical factors. It is then somewhat surprising then that the Historical scale demonstrated gradual declines in accuracy over time. In another study by the same authors, Snowden, Gray and Taylor (2010) also found that the pattern of prediction for HCR-20\(^{V2}\) scales went from Historical scale (AUC = 0.70), to Risk Management scale (AUC = 0.69) to Clinical scale (AUC = 0.54),
the Historical scale being the strongest predictor, whilst the Clinical scale being the weakest.

Other research has indicated the slight superiority of the Historical scale, and equal outcomes for Clinical and Risk Management scales. For example, in Yang, Wong and Coid’s (2010) meta-analysis, little difference between AUC’s for the HCR-20\textsuperscript{V2}’s three scales was found: Historical scale = 0.67, Clinical scale = 0.66, and Risk Management scale 0.66.

Whilst a large proportion of the research indicates the dominance of the Historical scale, still other research has indicated the superiority of the Clinical or Risk Management scale. Douglas and colleagues (1999) (who evaluated the predictive validity of the HCR-20 (version 1 and version 2) reported Scale-level predictive validity indices for a sample of 193 civil psychiatric patients in Canada. A summary of these outcomes is provided in Table 13 below.

Table 13

*AUC’s for Violent Outcomes based on HCR-20 Scale Scores derived from* Douglas, Ogloff, Nicholls and Grant (1999)

<table>
<thead>
<tr>
<th>Recidivism Type</th>
<th>H-Scale</th>
<th>C-Scale</th>
<th>R-Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any Violence</td>
<td>0.72</td>
<td>0.63</td>
<td>0.73</td>
</tr>
<tr>
<td>Physical Violence</td>
<td>0.71</td>
<td>0.59</td>
<td>0.77</td>
</tr>
<tr>
<td>Threatening Behaviour</td>
<td>0.73</td>
<td>0.66</td>
<td>0.71</td>
</tr>
<tr>
<td>Criminal Violence</td>
<td>0.78</td>
<td>0.63</td>
<td>0.74</td>
</tr>
</tbody>
</table>

*Note.* H-Scale = Historical Scale, C-Scale = Clinical Scale, R-Scale = Risk Management Scale. All AUC’s are significant at the $p \leq .05$ level.
Although all scales were significantly predictive of post-release violence, the Historical scale and Risk Management scale produced the largest AUC’s, which were comparable. For the comprehensive outcome, ‘Any violence’ the R-Scale produced the largest AUC, closely followed by the Historical scale. Across all violent outcomes, the Clinical scale had the lowest AUC.

The superiority of the Risk Management scale, and similar performance by other Scales, was also detected in Guy’s (2008) comprehensive meta-analysis of SPJ approaches to assessing risk of violence. The performance of HCR-20\textsuperscript{V2} scales for violence was: Historical scale AUC = 0.70, Clinical scale AUC = 0.69, and Risk Management scale AUC = 0.71. For physical violence outcomes (including sexual violence), the Historical scale demonstrated superiority. AUC’s were 0.66, 0.63 and 0.63 for the Historical, Clinical and Risk Management scales respectively.

In Strand, Belfrage, Fransson and Levander's (1999) study, recidivists scored significantly higher on the HCR-20\textsuperscript{V2} than non-recidivists, particularly on items on the Clinical scale and Risk Management scale. No scale level AUC’s are provided, but the significance level is reported. At the scale level, the Historical scale demonstrated low levels of predictive validity (Historical scale Total score \( p = 0.54 \)). In contrast, the Clinical scale and Risk Management scale demonstrated high predictive validity (\( p 's < 0.001 \)). Specifically, all items on both scales had statistically significant predictive power (except for C3, psychiatric symptoms). Results suggested that it was the clinical and risk management factors, rather than historical factors that contributed to the predictive validity obtained. The authors acknowledged the outcome of the Historical scale as having low predictive validity being contrary to past research. They noted the importance of Historical variables in
offending populations generally, but queried whether clinical and risk management factors may hold special importance in forensic psychiatric populations. Campbell and colleagues (2014) suggested that the homogeneity of Strand and colleagues sample may have contributed to the H-Scale not discriminating between recidivists and non-recidivists, as 70% of the sample had violent index offences, which are captured on the H-Scale. Alternatively, clinical symptoms may see more variation, and therefore, greater distinction.

Dolan and colleagues (2010) evaluated the predictive validity of the HCR-20\textsuperscript{V2} in forensic psychiatric patients in the UK. They found the predictive validity of the HCR-20\textsuperscript{V2} was strongly underpinned by the Clinical and Risk Management scales. AUC’s were as follows: Historical scale (AUC = 0.59, p = 0.16), Clinical scale (AUC = 0.91, p < .01), and Risk Management scale (AUC = 0.86, p < .01). This is in contrast to Gray and colleagues (2008) who found the Historical scale and Risk Management scale to be most predictive in patients discharged from medium security care. What is even more surprising, under the notion of static risk factors as the best predictors of long-term violence risk, is that Gray and colleagues’ (2008) study had a two-year follow-up, whereas Dolan and Blattner’s (2010) study had a mean follow-up period of six-years, yet the Clinical scale was still most predictive. The predictive validity of the Clinical and Risk Management scales over the Historical scales has also been found in in-patient forensic psychiatric settings (Chu et al., 2011), particularly within short-term follow-ups (i.e., 1-3 months).

Furthermore, the superior predictive validity of the Clinical scale relative to other scales has been reported. In their community follow-up study
of forensic psychiatric patients, Douglas and colleagues (2003), found the HCR-20\(^{V2}\)’s Clinical scale to be a significant predictor of violence (AUC = 0.68, \(p < .05\)), which surpassed the performance of the Historical scale (AUC = 0.63, non-significant) and Risk Management scale (AUC = 0.53, non-significant).

In HCR-20\(^{V3}\) research, both Strub and colleagues (2014) and Doyle and colleagues (2014) found the Clinical scale to be the strongest predictor of post-release violence. Doyle and colleagues (2014) found that all three scales significantly predicted violence at both 6-month and 12-month follow-up points, but the Clinical scale has the largest AUC. AUC values at 6-months were as follows: Historical scale = 0.63; Clinical Scale = 0.74; and Risk Management scale = 0.67. AUC values at 12-months were as follows: Historical scale = 0.63; Clinical Scale = 0.70; and Risk Management scale = 0.63. At both follow-up points, the Clinical Scale was the best predictor, whereas the Historical scale was the worst predictor. In a mixed civil psychiatric and forensic sample, Strub and colleagues (2014) found that the HCR-20\(^{V3}\) Historical scale was more predictive in the short-term (4-6 weeks), whereas the Clinical scale was more predictive in the long term (6-8 months).

In a sample of 99 forensic psychiatric patients, Hogan and Olver (2016) found that the HCR-20\(^{V3}\) scales were all significantly predictive of inpatient aggression. Presence scores on the Historical scale produced a medium effect size (AUC = 0.64, \(p < .05\)), whereas the Clinical and Risk Management scales produced large effect sizes (Clinical scale AUC = 0.76, \(p < .001\); Risk Management scale AUC = 0.76, \(p < .001\)). Interestingly, the Relevance ratings followed a very similar pattern, with lower AUC’s on Historical scale and identical AUC’s for both the Clinical scale and Risk
Management scale: Historical scale (AUC = 0.64, \( p < .05 \)), Clinical scale (AUC = 0.72, \( p < .001 \)) and Risk Management scale (AUC = 0.72, \( p < .001 \)).

3.4.2.1 Conclusion

The Historical scale is generally regarded as the most predictive scale (Douglas et al., 2014). A substantial body of evidence supports this assertion, however this is often dependent on the setting, population and follow-up periods. Other research has demonstrated comparable performance by the Historical and Risk Management scales; the Clinical and Risk Management scales; and occasionally the Clinical scale in isolation. The superiority of scales varies greatly within the body of research, across different settings and populations. Despite this, one generally consistent finding across research is the utility of each of the HCR-20’s scales in predicting violence. Table 14 summarises research on HCR-20’s scales.

Table 14

*Summary of Studies on the HCR-20’s Scale-level Predictive Validity*

<table>
<thead>
<tr>
<th>Authors</th>
<th>Sample</th>
<th>Country</th>
<th>Version</th>
<th>AUC/ ( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wintrup (1996)</td>
<td>Forensic</td>
<td>Canada</td>
<td>HCR-20(^1)</td>
<td>H: ( r = 0.20 )</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td></td>
<td>C: ( r = -0.02 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: ( r = 0.08 )</td>
</tr>
<tr>
<td>Campbell (2007)</td>
<td>Forensic</td>
<td>Australia</td>
<td>HCR-20(^{1/2})</td>
<td>H: 0.66-0.78</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td></td>
<td>C: 0.56-0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.67-0.71</td>
</tr>
<tr>
<td>Authors</td>
<td>Sample</td>
<td>Country</td>
<td>Version</td>
<td>AUC/ r</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------------</td>
<td>-----------</td>
<td>----------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Garcia-Mansilla, Civil</td>
<td>USA</td>
<td>HC²⁻¹⁵</td>
<td>H: 0.68</td>
<td></td>
</tr>
<tr>
<td>Rosenfeld, &amp; Cruise (2011)</td>
<td>Psychiatric</td>
<td></td>
<td>C: 0.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HC: 0.66</td>
<td></td>
</tr>
<tr>
<td>Gray, Taylor, &amp; Snowden (2008)</td>
<td>Forensic</td>
<td>UK</td>
<td>HCR⁻²⁻²</td>
<td>H: 0.68-0.77</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>C: 0.54-0.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: 0.63-0.69</td>
<td></td>
</tr>
<tr>
<td>Snowden, Gray, &amp; Taylor (2010)</td>
<td>Forensic</td>
<td>UK</td>
<td>HCR⁻²⁻²</td>
<td>H: 0.70</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>C: 0.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: 0.69</td>
<td></td>
</tr>
<tr>
<td>Yang, Wong, Coid (2010)</td>
<td>Mixed Sample</td>
<td>Intl.</td>
<td>HCR⁻²⁻²</td>
<td>H: 0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C: 0.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: 0.66</td>
<td></td>
</tr>
<tr>
<td>Douglas, Ogloff, Nicholls, &amp; Grant (1999)</td>
<td>Civil</td>
<td>Canada</td>
<td>HCR⁻²⁻²</td>
<td>H: 0.72</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td>HCR⁻²⁻²</td>
<td>C: 0.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: 0.73</td>
<td></td>
</tr>
<tr>
<td>Guy (2008)</td>
<td>Mixed Sample</td>
<td>Intl.</td>
<td>HCR⁻²⁻²</td>
<td>H: 0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C: 0.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: 0.71</td>
<td></td>
</tr>
<tr>
<td>Strand, Belfrage, Fransson, &amp; Levander (1999)</td>
<td>Forensic</td>
<td>Sweden</td>
<td>HCR⁻²⁻²</td>
<td>H: p = 0.54</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>C: p &lt; .001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>Dolan &amp; Blattner (2010)</td>
<td>Forensic</td>
<td>UK</td>
<td>HCR⁻²⁻²</td>
<td>H: 0.59</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>C: 0.91</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: 0.86</td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Sample</td>
<td>Country</td>
<td>Version</td>
<td>AUC/ r</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td>Douglas, Ogloff, &amp; Hart (2003)</td>
<td>Forensic</td>
<td>Canada</td>
<td>HCR-20(^{v2})</td>
<td>H: 0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.53</td>
</tr>
<tr>
<td>Doyle, Power, Coid, Kallis, Ullrich, &amp; Shaw (2014)</td>
<td>Forensic</td>
<td>UK</td>
<td>HCR-20(^{v3})</td>
<td>H: 0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 0.70-0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.63-0.67</td>
</tr>
<tr>
<td>Hogan &amp; Olver (2016)</td>
<td>Forensic</td>
<td>Canada</td>
<td>HCR-20(^{v3})</td>
<td>H: 0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.76</td>
</tr>
</tbody>
</table>

*Note.* \(^1\) No Scale level AUC’s are provided, but the significance level is reported. Intl = International study.

### 3.4.3 Inter-rater Reliability

Inter-rater reliability focuses on whether raters agree on ratings. This is particularly important given the clinical judgment features of the HCR-20. Inter-rater reliability can be considered for several aspects of the tool, including Items, scales, Total scores and Summary Risk Ratings (SRR’s). Novel elements of the HCR-20\(^{v3}\) such as the Relevance ratings can also be assessed for inter-rater reliability. Statistically, inter-rater reliability is usually evaluated through correlations or the use of the Intraclass Correlation Coefficient (ICC), which is a measure of agreement and consistency between raters, and can be reported based on single-rater ratings (ICC\(_1\)) or averaged ratings (ICC\(_2\)) (see Chapter 5 Statistical Methods, Section 5.6.7 for a more detailed description). The focus of the current section is the inter-rater
reliability for the HCR-20\textsuperscript{v3}, as relevant to the current research (i.e., Presence scale-level and Total score ratings, and Relevance ratings).

Research has consistently revealed good inter-rater reliability for the various aspects of the HCR-20 (Douglas & Weir, 2003; Douglas et al., 2014). Early research into the inter-rater reliability of the HCR-20 (Version 1 and Version 2) within forensic psychiatric and civil psychiatric samples indicated moderate to high inter-rater reliability and internal consistency (e.g., Belfrage, 1998; Dernevik, 1998; Douglas et al., 1999; Wintrup, 1996).

Douglas and colleagues (2003) looked at the inter-rater reliability of Scale scores, Total scores and SRR’s of the HCR-20\textsuperscript{v2} using a sample of 100 forensic psychiatric patients. Two raters scored half the sample ($n = 50$) for reliability purposes. At the Item-level, ICC’s ranged from 0.41 – 0.89 for the H-Scale, 0.34 – 0.69 for the C-Scale and 0.01 – 0.54 for the R-Scale. Whilst eight of 10 ICC’s on the H-Scale were greater than 0.70, the range of R-Scale scores was clearly below acceptable levels.

ICC values for Scale Scores were as follows: H-Scale (ICC\textsubscript{1} = 0.90, $p \leq .001$; ICC\textsubscript{2} = 0.94, $p \leq .001$); C-Scale (ICC\textsubscript{1} = 0.79, $p \leq .001$; ICC\textsubscript{2} = 0.88, $p \leq .001$); and R-Scale (ICC\textsubscript{1} = 0.47, $p \leq .001$; ICC\textsubscript{2} = 0.64, $p \leq .001$). The greatest level of agreement was found for the H-Scale, followed by the C-Scale and R-Scale.

ICC values for the Total scores was ICC\textsubscript{1} = 0.85 ($p < .001$) and ICC\textsubscript{2} = 0.92 ($p < .001$). Raters agreed on 70% of cases for the SRR. There were no extreme polarities (i.e., low and high combinations). A one-way random effects model revealed inter-rater reliability for SRR’s ranged from good to substantial. ICC\textsubscript{1} was 0.61 ($p \leq .001, 95\% \text{ CI} = 0.41$ to 0.76). ICC\textsubscript{2} was .76 ($p \leq .001, 95\% \text{ CI} = 0.58$ to 0.86). Inter-rater agreement was found to be greater
for Total scores than SRR’s. These observed Total score results are similar to Snowden and colleagues (2010), who reported an ICC of 0.80 (95% CI = 0.54-0.91) for the HCR-20 Total score in a forensic psychiatric setting, as rated by four psychologists across 19 cases. Despite the promising outcomes observed in Douglas and colleagues research (2003), it is noted that the study was limited by a pseudo-prospective design and ratings based exclusively on in-patient files.

Strong agreement for the HCR-20V² was also found in Campbell’s (2007) research, across Total scores (ICC = 0.93) and scales (H-Scale = 0.95; C-Scale = 0.86; R-Scale = 0.86). This is similar to reliability outcomes in Doyle and Dolan’s (2006) research, where the HCR-20V²’s C-Scale and R-Scale ICC’s ranged from 0.83 – 0.85. Good inter-rater reliability has also been found in other studies within forensic psychiatric populations (e.g., Strand et al., 1999).

de Vogel and colleagues (2014) provided initial research from the Netherlands on the inter-rater reliability of the first draft of the HCR-20V³. The research also focused on comparisons to the HCR-20V². Three raters rated twenty-five cases (29.1%) to derive inter-rater reliability data. Total scores yielded an ICC of 0.84. In comparison, the HCR-20V² Total scores produced an ICC of 0.83, suggesting that the two tools were comparable in predicting violent recidivism. The HCR-20V³ SRR’s yielded an ICC of 0.72, which was considered ‘good’. Notably, this study employed the first draft of the HCR-20V³, and was also lacking in reporting no item-level nor Scale-level ICC’s.

Kötter and colleagues (2014) also conducted initial research into the HCR-20V³ in Germany. Their involvement in the refinement of the tool involved clinical beta-testing of draft 1 and an evaluation of the inter-rater
reliability of draft 2 of the HCR-20\textsuperscript{V3} (i.e., the penultimate draft). Beta-testing resulted in qualitatively based commentary on the administration, clarity and reasoning of risk factors, clinical utility, relevance in risk management and improvements in risk communication.

Regarding the inter-rater reliability aspect of the research, five psychologists with no experience in conducting SPJ risk assessments each scored 30 case vignettes, based on actual patients at the Vitos Haina Forensic Psychiatric Hospital, an institution for MDO’s in Hessen. Although raters were inexperienced, they each held at least a degree in psychology. These vignettes represented an all-male sample, but varied on diagnoses (psychotic disorders, personality disorders or intellectual disability – some with comorbid diagnoses) and levels of risk. Ratings were based on discharge summaries in hospital files, and hence the R-scale was coded in a ‘community/out’ setting. ICC’s were calculated for individual items, sub-items, scales and SRR’s. H-Scale ICC’s ranged from 0.06 - 0.99; four of nine items had ICC’s greater than 0.70, as did 10/15 sub-items. C-Scale ICC’s ranged from 0.43 - 0.92. Two of five ICC’s were greater than 0.70, as were seven of the sub-items. R-Scale ICC’s ranged from 0.52 - 0.86; again two of five ICC’s were greater than 0.70, as were two sub-items. The average ICC’s for H-Scale, C-Scale and R-Scale were 0.65, 0.66 and 0.73, respectively. This range is considered ‘good’ to ‘substantial’ (based on ICC ranges by Fleiss (1981) and Landis and Koch (1977)). For SRR’s, the level of agreement ranged from ‘excellent’ to ‘almost perfect’ (ICC = 0.86, 95% CI = 0.78-0.92).

These preliminary results indicate that the HCR-20\textsuperscript{V3} can be reliably rated. Although these results are based on a draft version of the HCR-20\textsuperscript{V3}, they are still reflective of the final tool’s psychometric properties as the
differences between the draft and final version were essentially minor wording changes. Given the inexperience of raters, these results are also promising. In comparison to other inter-rater reliability studies such as Douglas and colleagues (2003), H-Scale ICC’s were lower but still falling within a good range. The greatest improvement has been seen in the R-Scale, which is probably reflective of the fact that the R-Scale was subjected to the most substantial changes in the revision process.

Douglas and Belfrage (2014) also conducted research into the inter-rater reliability of the HCR-20\textsuperscript{V3}. Three experienced clinicians from the Sundsvall Forensic Psychiatric Hospital in Sweden each conducted 3 independent ratings of 32 forensic psychiatric patients. Inter-rater reliability data were computed for Presence ratings (risk factors and sub-items), Relevance ratings and SRR’s. Ratings were based on psychiatric file records and interviews with the patients. All raters had access to identical information. Both ICC\textsubscript{1} (single rater) and ICC\textsubscript{2} (group of raters) were computed, with ICC\textsubscript{2} values being generally higher because it represents averaged ratings. Interpretation of ICC’s are once again based on Fleiss (1981) and Landis and Koch’s (1977) categories. Outcomes were computed for both institutional (In) and community (Out) for the R-Scale, Total scores and SRR’s. Table 15 summarises the results.

At the item-level, the average ICC\textsubscript{1}’s for the H-Scale ranged from substantial-excellent, whereas the C-Scale ranged from good to substantial, and R-Scale ranged from poor-fair. The average item-level reliability for Relevance ratings were 0.79 for H-Scale, 0.59 for C-Scale, 0.56 for R-Scale (In) and 0.61 for R-Scale (Out). For the HCR-20\textsuperscript{V3} Total Relevance score, they were 0.60 (In) and 0.61 (Out).
Scale-level analyses were based on the summation of risk factor scores (not including sub-items). Results indicated that inter-rater reliability was ‘excellent’ to ‘almost perfect’ for the all scale-level Presence ratings, except one (R-Scale (In)), which was in the ‘good’ range. Inter-rater reliability was generally lower for the R-Scale. Relevance ratings were in the ‘substantial’ to ‘good’ range.

Inter-rater reliability was good for SRR’s across both In and Out contexts. Paired ratings revealed that, across community and institutional contexts, 86.15% of pairings were in perfect agreement (e.g., Low-Low), 13.20% disagreed by one category (e.g., Low-Moderate, Moderate-high), and 0.65% (one case) disagreed with a two category discrepancy (i.e., Low-High).

As an interesting addition, authors correlated the ICC\(_1\) values for Presence and Relevance ratings, to assess if the reliability of risk factor Presence was associated with the reliability of risk factor Relevance. The association was 0.31 (statistic unspecified). This suggested “more reliably rated items, in terms of Presence, are also rated more reliably in terms of Relevance” (p. 136).

Findings indicated that the inter-rater reliability of the HCR-20\(^{V3}\) SRR and Presence of risk factors was excellent. For Relevance ratings, the reliability of the majority of items ranged from good to excellent. These findings support the inter-rater reliability of the HCR-20\(^{V3}\) across a range of different agreements: items, sub-items, scales, Total scores and SRR’s. In general, there was a strong level of agreement amongst raters at the item level, scale level and SRR.

A novel aspect of this study was the provision of greater insight into how Relevance scores operate, which is important given that it is a new
aspect in violence risk assessment. A clear limitation of the study was that the same raters completed both versions.

Table 15

*Intraclass Correlation Coefficients for the HCR-20V3 derived from Douglas and Belfrage (2014)*

<table>
<thead>
<tr>
<th>Scale/ Judgment</th>
<th>Presence</th>
<th></th>
<th></th>
<th>Relevance</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICC₁</td>
<td>ICC₂</td>
<td>ICC₁</td>
<td>ICC₂</td>
<td></td>
</tr>
<tr>
<td>H-Scale</td>
<td>0.94</td>
<td>0.98</td>
<td>0.62</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>C-Scale</td>
<td>0.86</td>
<td>0.95</td>
<td>0.60</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>R-Scale (In)</td>
<td>0.69</td>
<td>0.87</td>
<td>0.74</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>R-Scale (Out)</td>
<td>0.75</td>
<td>0.90</td>
<td>0.69</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>HCR-20 Total (In)</td>
<td>0.94</td>
<td>0.98</td>
<td>0.76</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>HCR-20 Total (Out)</td>
<td>0.94</td>
<td>0.98</td>
<td>0.80</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>SRR (In)*</td>
<td>0.81</td>
<td>0.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRR (Out)*</td>
<td>0.75</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. N = 23 for SRR analyses.

Doyle and colleagues (2014) evaluated the inter-rater reliability of the HCR-20V3 in a sample of 387 forensic psychiatric patients in England and Wales. It was hypothesised that the HCR-20V3 Total scores and scale scores could be reliably coded across different raters. Four raters each independently rated 20 randomly selected participants. Notably, raters were members of staff who knew the patient well, but researchers did not control or match for the type of professional (e.g., psychiatrist, nurse, psychologist). ICC’s were as follows: HCR-20V3 Total score = 0.92, H-Scale = 0.91, C-Scale = 0.90 and R-Scale = 0.93. These results suggest a very good level of inter-
rater reliability for both Presence ratings Total scores and scale scores. Interestingly, inter-rater reliability was greater for the R-Scale than the H-Scale.

Smith and colleagues (2014) evaluated the inter-rater reliability of the HCR-20\textsuperscript{V3} in a sample of 84 male inmates from a pre-trial correctional facility (i.e., remand facility) in the USA. 15 cases (18%) were randomly selected and independently rated by two researchers. ICC’s for Presence ratings were as follows: H-Scale (ICC = 0.92), C-Scale (0.67), R-Scale (In) (ICC = 0.68) and R-Scale (Out) (ICC = 0.88). ICC’s for Relevance ratings were as follows: H-Scale (ICC = 0.85), C-Scale (0.77), R-Scale (In) (ICC = 0.48) and R-Scale (Out) (ICC = 0.67). Amongst Presence ratings, inter-rater reliability was higher for the H-Scale and R-Scale (Out), whereas amongst Relevance ratings, inter-rater reliability was higher for the H-Scale and C-Scale. The R-Scale (In) demonstrated the lowest ICC, and is clearly below acceptable levels. The three SRR types (Future Violence, Serious Physical Harm and Imminent Violence) were compared for both institutional and community contexts, resulting in six SRR categories. The overall agreement for SRR’s was moderate (weighted kappa = 0.60, based on Shrout (1998) standards). Across the six SRR’s inter-rater reliability ranged from fair (weighted kappa = 0.43) to substantial (weighted kappa = 0.85). Only one of 90 comparisons had a low-high discrepancy.

More recent research by Howe, Rosenfeld, Foellmi, Stern and Rotter (2016) evaluated the HCR-20\textsuperscript{V3}’s inter-rater reliability for items and SRR’s, relative to the HCR-20\textsuperscript{V2}. Researchers compared HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} ratings for 64 civil psychiatric inpatients. Four raters each completed Versions 2 and 3 approximately 4-weeks into the patients hospital stay, based on
medical records and, where possible, brief interviews with the treating team. It was hypothesised that good to excellent inter-rater reliability would be found for the HCR-20\textsuperscript{V3}. ICC’s were calculated between two raters on 35 cases. For Version 2, ICC’s were found to be good to excellent for scale scores, Total scores and SRR’s: H-Scale (ICC = 0.84), C-Scale (ICC = 0.70), R-Scale (ICC = 0.60), Total score (ICC = 0.80) and SRR (ICC = 0.66). For Version 3, ICC’s were moderate to excellent for Scale scores, Total scores and SRR’s: H-Scale (ICC = 0.71), C-Scale (ICC = 0.55), R-Scale (ICC = 0.48), Total score (ICC = 0.64) and SRR (Case Prioritisation) (ICC = 0.51). The authors also noted that SRR’s had lower ICC’s than Total scores, suggesting “strong agreement about the number of risk factors present, [but] less agreement about how the item ratings inform SRR’s” (p. 409). The authors highlighted that lower SRR inter-rater reliability may be due to the restricted variance of SRR’s. Inter-rater reliability was stronger for Presence than Relevance ratings, which may be a reflection of the Relevance ratings as a new component in the HCR-20\textsuperscript{V3}.

This study was limited by variable times between ratings (i.e., ratings occurred on average at 4-weeks, but varied between 14 days post admission to post-discharge). The order of ratings were not counter-balanced, and Version 2 was always completed before Version 3. Ratings were also completed by graduate students, with limited experience in risk assessment. This may be a particularly important issue in relation to Relevance ratings, which may benefit from clinical experience.

Finally, Green and colleagues (2016) found that inter-rater reliability for the HCR-20\textsuperscript{V3} was excellent (Kappa = 0.89, \( p < .001 \)). The HCR-20\textsuperscript{V3} Total score yielded an ICC\textsubscript{2} of 0.94 (95% CI = 0.90 – 0.97) for males, and an ICC\textsubscript{2}
of 0.90 (95% CI = 0.72- 0.96) for females. Inter-rater reliability was excellent for males across all scales: H-Scale ICC\textsuperscript{2} = 0.91, C-Scale ICC\textsuperscript{2} = 0.88, and R-Scale ICC\textsuperscript{2} = 0.87. Similar results were observed for females: H-Scale ICC\textsuperscript{2} = 0.80, C-Scale ICC\textsuperscript{2} = 0.89, and R-Scale ICC\textsuperscript{2} = 0.92. The percentage of agreement was also high across Items.

### 3.4.3.1 Conclusion

The HCR-20 has generally produced good to excellent inter-rater reliability at the scale, Total score and SRR levels, across versions and settings. Most studies report ICC’s of .80 or greater (Douglas & Reeves, 2010). Studies surveyed herein suggest greater agreement for Total scores than SRR’s (e.g., Douglas et al., 2003; de Vogel et al., 2014; Douglas et al., 2014; Howe et al., 2016), however limited studies are available, particularly on the HCR-20\textsuperscript{V3}. One consistent finding across the research is that there are limited polarities (i.e., conflicting Low/ High judgments). Most studies indicate greater agreement for the Historical scale, which is probably due to its static nature (Douglas et al., 2010). Greater variability in ratings has been noted for the R-Scale, which may be associated with uncertainty in contextual issues in discharge/release planning. Table 16 below provides a summary of the studies discussed herein. The ICC\textsubscript{1} index is used.
Table 16

Summary of Inter-rater Reliability Outcomes for the HCR-20

<table>
<thead>
<tr>
<th>Authors</th>
<th>Sample</th>
<th>Country</th>
<th>Version</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>Item-level</td>
<td>H: 0.41-0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 0.34-0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.01-0.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Scale-level</td>
<td>H: 0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 0.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total score: 0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SRR: 0.61</td>
</tr>
<tr>
<td>Snowden, Gray, &amp; Taylor (2010)</td>
<td>Forensic</td>
<td>UK</td>
<td>HCR-20V2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>Total score: 0.80</td>
<td></td>
</tr>
<tr>
<td>Campbell (2007)</td>
<td>Forensic</td>
<td>Australia</td>
<td>HCR-20V2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>Scale-level</td>
<td>H: 0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total score: 0.93</td>
</tr>
<tr>
<td></td>
<td>&amp; Forensic</td>
<td></td>
<td>Scale-level</td>
<td>C: 0.85</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td></td>
<td>R: 0.83</td>
</tr>
<tr>
<td>de Vogel, van den Broek, &amp; de Vries Robbé (2014)</td>
<td>Forensic</td>
<td>Netherlands</td>
<td>HCR-20V2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>Total score: 0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HCR-20V3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total score: 0.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SRR:     0.72</td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Sample</td>
<td>Country</td>
<td>Version</td>
<td>ICC</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------</td>
<td>-------------</td>
<td>----------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Kötter, Franqué,</td>
<td>Forensic</td>
<td>Germany</td>
<td>HCR-20$^{V3}$</td>
<td></td>
</tr>
<tr>
<td>Bolzmacher, Eucker,</td>
<td>Psychiatric</td>
<td></td>
<td>Scale-level</td>
<td>H: 0.65</td>
</tr>
<tr>
<td>Holzinger, &amp;</td>
<td></td>
<td></td>
<td></td>
<td>C: 0.66</td>
</tr>
<tr>
<td>Müller-Isberner (2014)$^1$</td>
<td></td>
<td></td>
<td></td>
<td>R: 0.73</td>
</tr>
<tr>
<td>Douglas &amp; Belfrage (2014)</td>
<td>Forensic</td>
<td>Sweden</td>
<td>HCR-20$^{V3}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>Scale-level</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Presence</td>
<td>H: 0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Relevance</td>
<td>H: 0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total score:</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SRR:</td>
<td>0.75-0.81</td>
</tr>
<tr>
<td>Smith, Kelley, Rulseh,</td>
<td>Correctional</td>
<td>USA</td>
<td>HCR-20$^{V3}$</td>
<td></td>
</tr>
<tr>
<td>Sörman, &amp; Edens (2014)</td>
<td></td>
<td></td>
<td>Scale-level</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Presence</td>
<td>H: 0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.68-0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Relevance</td>
<td>H: 0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 0.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R: 0.48-0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SRR:</td>
<td>0.43-0.85$^2$</td>
</tr>
<tr>
<td>Authors</td>
<td>Sample</td>
<td>Country</td>
<td>Version</td>
<td>ICC</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------------</td>
<td>---------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>Howe, Rosenfeld,</td>
<td>Civil</td>
<td>USA</td>
<td>HCR-20\textsuperscript{V2}</td>
<td></td>
</tr>
<tr>
<td>Foellmi, Stern, &amp;</td>
<td>Psychiatric</td>
<td></td>
<td>H: 0.84</td>
<td></td>
</tr>
<tr>
<td>Rotter (2016)</td>
<td></td>
<td></td>
<td>C: 0.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R: 0.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total score: 0.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SRR: 0.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HCR-20\textsuperscript{V3}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scale-level</td>
<td>H: 0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C: 0.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R: 0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total score: 0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SRR: 0.51\textsuperscript{3}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green, Schneider,</td>
<td>Insanity</td>
<td>USA</td>
<td>HCR-20\textsuperscript{V3}</td>
<td></td>
</tr>
<tr>
<td>Griswold, Belfi,</td>
<td>Acquittees</td>
<td></td>
<td>Total score 0.90-0.94</td>
<td></td>
</tr>
<tr>
<td>Herrera, &amp; DeBlasi</td>
<td></td>
<td></td>
<td>Scale-level H:0.80-0.91</td>
<td></td>
</tr>
<tr>
<td>(2016)</td>
<td></td>
<td></td>
<td>C:0.88-0.89</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R:0.87-0.92</td>
<td></td>
</tr>
</tbody>
</table>

Note. \textsuperscript{1} Average ICC’s across Items. \textsuperscript{2} weighted Kappa. \textsuperscript{3} Case Prioritisation SRR.

### 3.4.4 Concurrent Validity

Concurrent validity refers to the relationship of the HCR-20\textsuperscript{V3} to existing violence risk assessment instruments. Because such comparisons are often made between the HCR-20 and VRAG, this section of the chapter will also address the concurrent validity of the VRAG and VRAG-R.
Early research on the HCR-20 (Version 1) by Wintrup (1996) indicated that the HCR-20 and PCL-R shared identical correlations with post-release violence in a sample of 80 forensic psychiatric patients ($r = .20, p < .10$). AUC’s produced by the HCR-20 have also fallen in similar ranges to AUC’s produced by other violence risk assessment tools (see section 3.7 for a more detailed discussion). For example, the HCR-20 has produced similar AUC’s to both the PCL-R/SV (Chu et al., 2011) and the VRAG (Doyle et al., 2012; Doyle et al., 2002; Grann et al., 2000). Comparable AUC’s between the VRAG and PCL-R have also been observed in other studies (Doyle & Dolan, 2006). Yang and colleagues (2010) have shown that similar AUC’s were produced by all three tools, the HCR-20, VRAG and PCL-R.

In one of the first studies on the concurrent validity of the HCR-20, Douglas and Webster (1999) conducted a retrospective file review of 75 maximum-security offenders in Canada. They compared the historical and clinical component of the HCR-20 (version 1) (i.e., HC-15) with the VRAG and the PCL-R. The HC-15 and VRAG were compared to various indexes of past violence. Results showed that the HC-15, particularly the H-10 scale was more strongly correlated with past violence than the VRAG. Results indicated a fair degree of overlap between the HCR-20 and the two other instruments. The H-Scale significantly correlated with the VRAG ($r = 0.62$) and the PCL-R ($r = 0.50$). The C-Scale was not as strongly correlated with the VRAG, $r = 0.26$, nor the PCL-R, $r = 0.43$. The HC-15 scale produced large correlations with the VRAG ($r = 0.55$) and PCL-R ($r = 0.59$). The authors concluded that although the HCR-20 shares variance with other instruments, it also possesses independent variance.
Warren, South, Burnette, Rogers, Friend, Bale and Patten (2005) investigated the concordance between the HCR-20\textsuperscript{V2} and the PCL-R in a sample of 132 female offenders. The sample was young (60% < 32 years-old) and mostly represented women of minority status. The majority of participants (83%) had at least one violent crime conviction. The tools were completed based on interviews and file-reviews. Results indicated that the HCR-20\textsuperscript{V2} Total score and PCL-R Total score were significantly correlated positively correlated with a large effect size ($r = 0.80$, $p = .001$). Significant positive correlations were also found between the PCL-R Factor 1 and 2 scores and each of the HCR-20\textsuperscript{V2} scale scores. The authors concluded that there was a “high degree of overlap” (p. 285) between the two tools. The study was limited however by a small sample size and attrition rate (due to transfers to lower-level security prisons), which could indicate that the remaining sample comprised of females with more violent crimes and longer periods of incarceration. This may also affect generalisability to community-based women or those residing in less secure settings.

Regarding more recent research on the HCR-20\textsuperscript{V3}, comparisons have tended to be drawn between the HCR-20\textsuperscript{V3} and the HCR-20\textsuperscript{V2}, as opposed to other violence risk assessment tools. de Vogel and colleagues (2014) compared the HCR-20\textsuperscript{V2} and the draft version of the HCR-20\textsuperscript{V3} in a sample of 86 male forensic psychiatric patients convicted of violent offences and discharged from the Van der Hoeven Kliniek in the Netherlands between 1990 and 2006. They found that the HCR-20\textsuperscript{V2} Total scores and Total scores on the draft version of the HCR-20\textsuperscript{V3} were significantly correlated (0.93, $p < .001$).
Bjorkly and colleagues (2014) also looked at the concurrent validity of the HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} in a sample of 20 forensic psychiatric patients from medium-security units in Norway. Two experienced psychiatric nurses conducted the assessments. The tools were completed based on information contained in patient’s files, observations and consultations with colleagues. The HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} Total Presence Scores correlated significantly ($r = 0.58$, $p = 0.007$). A paired sample t-test estimated the differences between versions on presence scores. Significant differences in scores were found for the Total score, H-Scale and C-Scale, but not for the R-Scale. Cronbach’s alpha was used to evaluate the internal consistency of the tools across Total scores and scale scores. On both versions, internal consistency ranged from moderate for the C-Scale to good for the H-Scale, R-Scale and Total scores. These results reflect the common underlying dimensions between the two tools, but also differences, which seemed to stem predominantly from clinical items. The authors concluded that the results were promising given the substantial changes in transitions between versions, and concluded that the concurrent validity of the two versions was “significant and solid” (p. 240), despite the study being limited by a small sample size ($N = 20$) and convenience sampling.

Strub and colleagues (2014) also evaluated that concurrent validity between the HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} in a sample of 106 psychiatric patients and criminal offenders transitioning from institution to community in Canada. The HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} were coded by the same raters. Correlations between the Total scores, scales and SRR’s on both instruments were as follows: H-Scale $r = 0.89$, C-Scale $r = 0.76$, R-Scale $r = 0.81$, Total scores $r = 0.91$, and SRR $r = 0.98$. All correlations were significant at the $p < .001$ level,
and in the moderate to large effect size range. The HCR-20\textsuperscript{V3} was highly correlated with the HCR-20\textsuperscript{V2}, which provides evidence for the “continuity of concept” goal in revision (Douglas et al., 2013, p.18), and indicates that the “underlying risk content is similar” (p. 156).

Howe and colleagues (2016) (described earlier) also sought to evaluate the concurrent validity by association between the HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} corresponding scales and SRR’s. It was hypothesised that strong associations would be found between versions. Two raters independently rated 35 cases. The correlation between the HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} Total scores were moderate and significant ($r = 0.62$, $p < .001$). Correlations between scales were as follows: H-Scale ($r = 0.68$, $p < .001$), C-Scale ($r = 0.48$, $p = .004$), R-Scale ($r = 0.46$, $p = .005$). Scale correlations were all significant, but C and R-scales were clearly lower than H-Scale associations. SRR correlations were provided for each type: Case Prioritisation ($r = 0.40$, $p = .02$), Serious Physical Harm ($r = 0.39$, $p = .02$), and Imminence of Violence ($r = 0.58$, $p < .001$). Thus, imminence of violence was the most reliably rated SRR. Notably, when comparisons were made between ratings on both tools computed by the same rater, associations were substantially stronger. Given the strong and significant correlations for scales and SRR’s, it is apparent that there is considerable similarity between the two versions and that item level changes have not greatly impacted overall scores and clinical judgments.

Douglas and Belfrage (2014) evaluated the concurrent validity of the HCR-20\textsuperscript{V3} relative to the HCR-20\textsuperscript{V2} across three raters who each evaluated 32 forensic psychiatric patients at Sundsvall Forensic Psychiatric Hospital in Sweden. Assessments were completed through file review and interviews with patients. Findings indicated significant strong correlations between
Version 2 and Version 3 Presence ratings: H-Scale $r = 0.87$, C-Scale $r = 0.76$, R-Scale (In) $r = 0.67$, R-Scale (Out) $r = 0.82$, Total score (In) $r = 0.85$, and Total score (Out) $r = 0.90$.

The author is only aware of one study that has addressed the concurrent validity of the VRAG-R. Preliminary evidence by Rice and colleagues (2013) demonstrated that the ICC between the VRAG and VRAG-R categories was 0.83 (95% CI = 0.81 – 0.85), indicating that the two tools produce similar case rankings.

### 3.4.4.1 Conclusion

Given that both the HCR-20 and VRAG purport to measure risk of violence, one would expect acceptable levels of concurrent validity. These studies represent a select few from an abundance of research. Common themes across the body of research relate to greater associations being observed between static components of measures (e.g., the Historical scale and the VRAG-R). Research has supported this notion (Douglas et al., 1999; Strub et al., 2014). Regarding the HCR-20, it would also be expected that Versions 2 and 3 correlate, especially given the “continuity of concept” goal in revision (Douglas et al., 2013, p.18). Several studies comparing the two versions in correctional, forensic psychiatric, and civil psychiatric populations have supported this (Bjorkly et al., 2014; Douglas et al., 2014; Howe et al., 2016; Strub et al., 2014), although differences are also noted. For example, although Bjorkly and colleagues (2014) found significant correlations between versions for both Presence and Relevance ratings, there were significant differences in Presence ratings for H-Scale, C-Scale and Total scores. There were also significant differences in Relevance ratings for all three scales and
Total scores. Investigation on the clinical application of the HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} using a case illustration also indicated differences, where the HCR-20\textsuperscript{V3} showed improved method in item clarity, differentiation of presence from relevance, and construction of risk management plans.

Levels of agreement for HCR-20\textsuperscript{V3} SRR’s and Total scores vary, with some studies suggesting that there is greater agreement for SRR’S (e.g., Strub et al., 2014), whereas others indicate greater levels of agreement for Total scores (e.g., Howe et al., 2016). The evidence base regarding the HCR-20\textsuperscript{V3} remains too limited for any firm conclusions to be drawn.

Evidence on the concurrent validity of the VRAG and VRAG-R was also reported in this section of the thesis because it represents intersections with the HCR-20 and other violence risk assessment tools. Initial research has suggested good associations between the VRAG and VRAG-R. Table 17 provides a summary of the concurrent validity studies discussed herein.

To date, the evaluation of concurrent validity for the HCR-20\textsuperscript{V3} and VRAG-R has only been extended to comparisons with their predecessors, the HCR-20\textsuperscript{V2} and original VRAG. A more convincing assessment of concurrent validity would stem from comparing the HCR-20\textsuperscript{V3} and VRAG-R to other independent established violence risk assessment tools.
Table 17

Concurrent Validity between the HCR-20 and other Comparative Violence Risk Assessment Instruments

<table>
<thead>
<tr>
<th>Authors</th>
<th>Sample</th>
<th>Country</th>
<th>Instruments</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Douglas &amp; Webster (1999)</td>
<td>Correctional</td>
<td>Canada</td>
<td>H-Scale &amp; VRAG</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C-Scale &amp; VRAG</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HC-15 &amp; VRAG</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H-Scale &amp; PCL-R</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C-Scale &amp; PCL-R</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HC-15 &amp; PCL-R</td>
<td>0.59</td>
</tr>
<tr>
<td>Warren, South, Burnette, Rogers,</td>
<td>Correctional</td>
<td>USA</td>
<td>HCR-20V2 &amp; PCL-R</td>
<td>0.80</td>
</tr>
<tr>
<td>Friend, Bale, &amp; Patten (2005)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de Vogel, van den Broek, &amp; de Vries (2014)</td>
<td>Forensic</td>
<td>Netherlands</td>
<td>HCR-20V2 &amp; HCR-20V3</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>Total scores</td>
<td>0.93</td>
</tr>
<tr>
<td>Bjorkly, Eidhammer, &amp; Selmer (2014)</td>
<td>Forensic</td>
<td>Norway</td>
<td>HCR-20V2 &amp; HCR-20V3</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>Total scores</td>
<td>0.58</td>
</tr>
<tr>
<td>Strub, Douglas, &amp; Nicholls (2014)</td>
<td>Sample</td>
<td>Canada</td>
<td>HCR-20V2 &amp; HCR-20V3</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H-Scale</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C-Scale</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R-Scale</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total scores</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SRR</td>
<td>0.98</td>
</tr>
<tr>
<td>Authors</td>
<td>Sample</td>
<td>Country</td>
<td>Instruments</td>
<td>Correlation</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------</td>
<td>---------</td>
<td>----------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Howe, Rosenfeld,</td>
<td>Civil</td>
<td>USA</td>
<td>HCR-20\textsuperscript{V2} &amp; HCR-20\textsuperscript{V3}</td>
<td></td>
</tr>
<tr>
<td>Foellmi, Stern, &amp; Rotter (2016)</td>
<td>Psychiatric</td>
<td></td>
<td>H-Scale</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C-Scale</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R-Scale</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total scores</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SRR</td>
<td>0.40\textsuperscript{2}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.39\textsuperscript{3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.58\textsuperscript{4}</td>
</tr>
<tr>
<td>Douglas &amp; Belfrage (2014)</td>
<td>Forensic</td>
<td>Sweden</td>
<td>HCR-20\textsuperscript{V2} &amp; HCR-20\textsuperscript{V3}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>H-Scale</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C-Scale</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R-Scale (In)</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R-Scale (Out)</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total score (In)</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total score (Out)</td>
<td>0.90</td>
</tr>
</tbody>
</table>

*Note.* Pearsons Correlations. \textsuperscript{1} HCR-20 (Version 1). \textsuperscript{2} Future Violence/ Case Prioritisation SRR \textsuperscript{3} Serious Physical Harm SRR \textsuperscript{4} Imminence of Violence SRR.

### 3.4.5 Incremental Validity

The question of incremental validity pertains to within instrument comparisons, as well as between instrument comparisons. Within the HCR-20, it has been questioned whether judgments such as the SRR’s add to numerical scores, and whether Relevance ratings add to Presence ratings.
Regarding incremental predictive validity within the HCR-20, a body of research had indicated that SRR’s add incremental validity to numerical Total scores (for a more comprehensive review, see Douglas, Hart, Groscup, & Litwack, 2014; Douglas & Reeves, 2010; Guy, 2008). Douglas and colleagues (2003) looked at the incremental validity of the HCR-20\textsuperscript{V2} associated with SRR’s and numerical scores. A cox regression was run to assess the incremental validity associated with scores and clinician final ratings, using ‘any violence’ as the outcome. Scale scores were entered into the model first, followed by SRR’s. The scores produced a significant model fit (notably, the C-Scale was the only significant predictor in the model). The addition of SRR’s resulted in significant improvement in model fit. These summary ratings had the strongest relationship to violence, above and beyond actuarial scores. Therefore, summary ratings added incremental validity to the HCR-20\textsuperscript{V2} Total scores. Additional analyses were conducted to include other covariates (e.g., PCL-R scores, gender, violent index offence). Results revealed that SRR’s were the only significant predictors in the model – suggesting that these summary ratings add incremental validity to HCR-20\textsuperscript{V2} actuarial scores even above other relevant predictors. The authors concluded that these findings support the SPJ model for assessing risk, which is essential given its intended use. Specifically regarding the HCR-20\textsuperscript{V2}, the clinical judgment component adds meaningfully to actuarial methods of assessing risk.

The introduction of Relevance ratings to the HCR-20\textsuperscript{V3} created a new element for the assessment of incremental validity within the HCR-20. Strub and colleagues (2014) focused on the incremental validity of the Presence and Relevance ratings, and then SRR’s relative to Presence and Relevance
ratings in a Canadian sample of 106 psychiatric patients and criminal offenders transitioning from institution to the community. Results indicated strong significant associations between Presence and Relevance ratings, where H-Scale $r = 0.73$, C-Scale $r = 0.74$, R-Scale $r = 0.63$, and HCR-20$^\text{V3}$ Total $r = 0.75$. The authors conducted a series of logistic regression analyses to see if Relevance ratings added incremental validity to Presence ratings, at both follow-up points. Presence ratings were entered into the model in Block 1, followed by Relevance ratings in Block 2. The overall model was significant but only Presence ratings were significant in the final model. Results indicated that Relevance ratings did not add incremental validity to the Presence ratings on scales or Total Scores. A secondary analysis using the interaction between Presence and Relevance ratings also indicated no improvement in incremental validity.

Regarding the incremental validity of SRR’s relative to Presence and Relevance ratings, further logistic regression analyses were performed, again at both time points. Incremental validity was tested for the HCR-20$^\text{V3}$ Presence total, as well as for scales individually. In both the short and long-term, SRR’s added incremental validity to Relevance ratings. Regarding Presence ratings, in the short-term, the SRR’s consistently added incrementally to the Presence ratings, with only one exception. In the long-term, SRR’s failed to add incrementally to the Presence ratings. Authors noted that the methodological aspect of comparing coding ranges of $0 – 40$ for Presence ratings and $0 – 2$ for SRR’s may influence outcomes as different metric ranges are being compared. The authors therefore trichotomised Presence ratings into three categories based on splits at the $33^{\text{rd}}$ and $66^{\text{th}}$ percentile where scores $0 - 19 = 0; 20 - 26 = 1; \text{and } 27 - 40 = 2$. Following this
conversion, the analysis was re-run (Presence ratings in Block 1; Relevance ratings in Block 2), and SRR’s were found to add incrementally to Presence ratings. The authors also ran logistic regression analyses to determine if outcomes were moderated by gender, however gender failed to emerge as significant in any models. In conclusion, SRR’s added incremental validity to Presence ratings, which is consistent with past research. SRR’s also added incremental validity to Relevance ratings, which is a novel finding. Even if Relevance ratings do not demonstrate incremental validity, they are still purportedly important to facilitating formulation and tailored idiographic assessment. The method of trichotomisation is a minor methodological variation that accounts for reduced variability in SRR’s and categorical and continuous pairings and may contribute to findings of no incremental validity between scores and summary judgments in some studies.

Smith and colleagues (2014) evaluated the incremental validity of the HCR-20V3 in a sample of 84 male inmates from a pre-trial correctional facility (i.e., remand facility) in the USA. Researchers wanted to investigate the relationship between Presence ratings and SRR’s, and whether any detected differences might be associated with the SRR’s conceptually representing both Presence and Relevance ratings. Therefore, they examined incremental validity afforded by the interaction between Presence and Relevance ratings, above Presence ratings alone. Completion of the HCR-20V3 was based on file review and patient interview. Because the study was conducted at the pre-trial stage, and there was uncertainty regarding the R-Scale ‘In/Out’ disposition, the R-Scale was coded for both In (institutional) and Out (community). An interaction term was created by multiplying each items Presence score with their Relevance score.
Stepwise Discriminant Function Analyses were run with interactions and Presence ratings to see if this interaction predicted SRR’s. Results indicated that only interaction terms remained in the model and uniquely predicted SRR’s, suggesting that SRR’s were a function of the interaction between Presence and Relevance ratings, rather than just a reflection of the sum of Presence scores. This makes sense conceptually because SRR’s are decided following formulation, which is based on a consideration of the presence and relevance of factors. Overall, SRR’s tended to be lower for institutional settings than community settings. Results supported previous research indicating that HCR-20 Total scores and SRR’s are not isomorphic, and that SRR’s have incremental validity over scale scores. This study was however limited in that not all raters attended the official training, instead training was subsequently provided to the rest of the research team by attendees. Also, participants may have presented in a socially desirable way or have concerns about how their disclosures could impact future care and monitoring.

More recent research by Howe and colleagues (2016) (described earlier) also sought to evaluate how the HCR-20V3’s Presence and Relevance ratings are associated with SRR’s, by exploring how the addition of Relevance ratings affected clinical judgments. It was hypothesised that the Presence and Relevance rating interaction would predict SRR’s (i.e., case prioritisation) more strongly than Relevance ratings in isolation. Interestingly, the study adopted a 5-point SRR rating (low, low-moderate, moderate, moderate-high, high) to identify subtle differences in risk. Regression analyses were computed where Presence was entered into Block 1, and Relevance in Block 2, and the Presence × Relevance interaction in Block 3.
This analysis was run for each risk factor. Out of the 20 items, the Presence × Relevance interaction provided a unique contribution in only four models.

Relevance ratings demonstrated fewer significant correlations with SRR’s than Presence ratings, but in several cases, correlations between the interaction terms and SRR’s were stronger than for presence ratings alone. This is consistent with Smith and colleagues (2014) who found that Relevance ratings contributed unique variance to the prediction of SRR’s beyond that accounted for by Presence ratings alone.

Between instrument comparisons relate to whether the HCR-20 adds predictive validity above and beyond other risk assessment tools, such as the VRAG, and vice versa. Because addressing the incremental validity of the HCR-20 also relates to the VRAG, the incremental validity of the VRAG-R will also be discussed in this section of the Chapter. The incremental validity of the Psychopathy Checklist – Revised (PCL-R) or Psychopathy Checklist: Screening Version (PCL:SV) is also important since the assessment of psychopathy was previously required in the HCR-20\textsuperscript{V2} and in the VRAG (although it is still partially required in the VRAG). Although the PCL-R and PCL:SV are not violence risk assessment tools per se, research indicates that they are related to violence and have high accuracy in the prediction of violence (for a discussion on the PCL-R and risk for violence, see Dematteo, Edens, & Hart, 2009; Hare, 2006; Hart, Hare, & Forth, 1994; Jaber & Mahmoud, 2015).

The decision to exclude the PCL-R from the HCR-20\textsuperscript{V3} was empirically grounded. Several studies indicated that the HCR-20 demonstrated incremental predictive validity above the PCL-R, but that the PCL-R did not add to the predictive accuracy of the HCR-20 (see below, Campbell, 2007;
Douglas et al., 1999; Douglas et al., 1999; Guy, Douglas, & Hendry, 2010). Early research on the HCR-20 (Version 1) suggested that the HCR-20 accounts for a larger proportion of variance in violence than the PCL-R, and that the PCL-R does not add incremental validity to the HCR-20 (Wintrup, 1996).

The previously mentioned study by Douglas and colleagues (1999) demonstrated that, 1) the HCR-20 (Versions 1 and 2) produced consistently greater AUC’s than the PCL:SV, a difference which was statistically significant (as indicated by the AccuROC program; Vida, 1997); 2) and that the HCR-20 added incremental validity to the PCL:SV (even when H7 Psychopathy was removed), however the converse was not true.

Doyle and Dolan (2006) assessed the incremental validity of static and dynamic measures in a mixed sample of 129 forensic psychiatric and civil psychiatric populations in the UK. The study utilised a prospective design, carried out across multiple sites. Patients were interviewed pre-discharge using a semi-structured interview schedule, and staff with knowledge of the patient were also interviewed as a source of collateral information. Based on this data, assessment tools were scored at the point of discharge. Of the 129 participants, 112 (86.8%) were available for community follow-up. The sample was almost entirely comprised of white males (93%), 70% of whom were diagnosed with serious mental illness. The total follow-up period was 24-weeks, averaging around 5.5 months. The target outcome was community violence, defined as “any acts that include battery that resulted in physical injury; sexual assaults; assaultative acts that involved the use of a weapon; or threats made with a weapon in hand” (p.521; as per the definition in the MacArthur Study of Mental Disorder and Violence, Monahan et al., 2001).
Outcomes were identified through official records, self-report (interviews with participants during the follow-up period) and collateral input.

The authors specifically wanted to examine the incremental validity of dynamic measures, the HCR-20\textsuperscript{V2}’s C-Scale and R-Scale were added to a battery of measures, including the PCL:SV, H-Scale (excluding PCL:SV) and VRAG (excluding PCL:SV), amongst other tools (i.e., The Novaco Anger Scale (NAS; Novaco, 2003) and the Barratt Impulsiveness Scale (BIS; Barratt, 1994)). Using hierarchical logistic regression with baseline measures entered into Block 1, this initial model was significant, correctly classifying 86% of the sample. However, the predictive accuracy was attributed to the NAS and BIS. The C-Scale and R-Scale were added in Block 2, the percentage of correctly classified people increased to 88%, which was significant. In this final model, only the C-Scale and R-Scale predicted community violence. The authors concluded that the HCR-20\textsuperscript{V2}’s C and R scales added significant incremental validity to the baseline static historical measures, indicating that the addition of dynamic factors improved prediction of community violence. While the sample was described as being heterogeneous, which has advantages for generalisability, it is noted that a more homogenous sample would have stabilised variability and improved control of confounding factors.

Douglas and Webster (1999) used hierarchical regression to compare the HCR-20 (Version 1) and VRAG. The HCR-20 was entered into the model first, then the VRAG. When the VRAG was added into the model there was no change in the amount of variance in violence accounted for (i.e., the HCR-20 accounted for all variance). When the VRAG was entered first, followed by HCR-20 total, the amount of variance accounted for increased
from $R^2 = 0.04$ to $R^2 = .19$. The authors concluded that the HCR-20 added incremental validity and improved the performance of the VRAG, but the VRAG did not add to the HCR-20.

Campbell (2007) used logistic regression to investigate the incremental validity associated with the HCR-20$^2$ and the PCL-R. At the univariate level, the HCR-20 Total score was significantly associated with violent reconviction ($p = 0.007$), whereas the PCL-R was not ($p = 0.233$). When the tools were entered into a multivariate model together, the HCR-20 continued to be significantly associated with violent reconviction, but the PCL-R demonstrated a weak association. The HCR-20 added to the predictive strength of the PCL-R (even when item H7, Psychopathy, was removed), but the PCL-R did not add to the predictive strength of the HCR-20.

Edens, Skeem and Douglas (2006) compared the VRAG and PCL:SV in a sample of civil psychiatric patients ($N = 695$), a subset of data from the MacArthur Violence Risk Assessment Study. The study used a modified 10-item version of the VRAG (including the PCL:SV), based on the lack of suitable proxies for two items (Failure on Prior Conditional Release and Victim Injury) from the MacArthur database (see Harris, Rice, & Camilleri (2004) for more detail). Participants were interviewed in hospital, and twice during follow-up. Outcomes were detected through patient self-report, collateral report (e.g., family member, friend or associated professional) and official records. Violence was defined as “battery that resulted in physical injury (ranging from bruises to death), sexual assaults, assaultive acts that involved the use of a weapon, or threats made with a weapon in hand” (p. 370). Outcomes were reported at 20-week and 50-week follow-up points, at which the base rates of violence were 18.6% and 27.5%, respectively.
The correlation between the modified VRAG and violence was $r = 0.22$. This was drastically reduced after controlling for the PCL:SV ($r = 0.04$), indicating a point of almost no correlation. Conversely, the correlation between the PCL:SV and violence ($r = 0.31$) remained reasonably unchanged after controlling for the modified VRAG ($r = 0.29$). Similar outcomes were found at the 50-week follow-up point (VRAG $r = 0.05$; PCL:SV $r = 0.30$).

AUC’s for the 10-item VRAG, PCL:SV, and 9-item VRAG (scored without the PCL:SV item) were provided at the 20-week point. AUC’s were as follows: 0.73, 0.78, and 0.64 (all significant at the $p < .001$ level). A similar pattern of results were found at the 50-week follow up. To address incremental validity, researchers ran two linear regression analyses, whereby the 9-item VRAG (without PCL:SV item) was used to predict PCL:SV scores, and the opposite computation in the second analysis. The AUC for the modified VRAG was 0.58, whereas the AUC for the PCL:SV was 0.75. Again, similar results were found at the 60-week point. In conclusion, the VRAG’s strong correlation with community violence is almost non-existent when the PCL:SV is take into account. Assessment of incremental validity demonstrated that the VRAG did not improve upon the predictive validity of the PCL:SV, and accounted for little variance secondary to the PCL:SV. However, the PCL:SV accounted for considerable variance secondary to the VRAG. Essentially, the validity of the VRAG is greatly underpinned by the PCL:SV.

Guy and colleagues (2010) conducted a meta-analysis into the role of psychopathy in evaluating violence risk using the HCR-20$^{V2}$. Practically, they investigated the inclusion of the PCL-R or PCL:SV in HCR-20$^{V2}$ assessments.
of violence and anti-social behaviour. They asked three specific questions: “(1) does the predictive accuracy of the HCR-20 drop meaningfully if the PCL-R/psychopathy item is removed? (2) What if the HCR-20 is completed without the PCL-R having been rated in the first place? (and) (3) Does the HCR-20 add incrementally to the PCL-R?” (p. 552). The initial sample pool comprised 55 studies, representing 53 independent samples. Two distinct sample groups were used: Aggregate findings from published and unpublished research (sample 1) and raw datasets for multivariate analyses (sample 2). Sample 2 represented forensic psychiatric, correctional and civil psychiatric populations, and had a high proportion of males (approximately 80%). Sample 1 was used to look at the association between violence and the HCR-20 and HCR-19 (i.e., HCR-20 without H7, Psychopathy). A subset of this sample was also used to make direct comparisons between the HCR-20 and PCL-R, based on studies that had used both tools. Sample 2 comprised 14 raw datasets and provided AUC’s for the prediction of violence for: The HCR-20, HCR-19 (HCR-20 excluding H7, Psychopathy), H7 and the PCL-R. A subset of this sample was also used to evaluate the incremental validity of the HCR-20 and PCL-R. Two outcome categories were used: Violent and ‘any’ recidivism. Violence was based on the official operational definition of the HCR-20.

Regarding the association between violence and the HCR-20 and HCR-19, results indicated that the HCR-19 performed similarly to the full HCR-20 in predicting violence (mean AUC’s = 0.69 and 0.70, respectively). For ‘any’ anti-social behaviour, the mean AUC’s for the HCR-20 and HCR-19 were 0.69 and 0.71, respectively. The predictive accuracy of the HCR-20 did not change significantly through exclusion of H7. For both outcomes, there
was no statistically significant variability across effect sizes, meaning that the exclusion of H7 did not significantly change the predictive accuracy of the HCR-20.

Direct comparisons between the HCR-20 and PCL-R based on 34 samples revealed that AUC’s for violence were similar and did not differ significantly. Mean AUC’s for the prediction of violence were 0.67 and 0.69 for the PCL-R and HCR-20 respectively. The exclusion of H7 did not reduce the HCR-20’s predictive accuracy. Similar results were observed for anti-social behaviour as the outcome. The authors expanded on this by looking at five studies that had completed the HCR-20 without H7. AUC’s for HCR-19 ranged from 0.62 – 0.81 (mean = 0.72). These values are comparable to effect sizes for HCR-20 assessments including H7, and indicate that the observed predictive validity of the HCR-20 is not based on the PCL-R. The mean AUC was 0.73 for the HCR-20. The mean AUC was 0.73, and remained stable at 0.73 even when H7 was removed.

Researchers investigated the incremental validity of the HCR-20 and PCL-R by conducting multivariate analyses on seven raw datasets that incorporated both tools. The mean AUC’s were 0.76 for the HCR-19 and 0.67 for the PCL-R. Hierarchical logistic regression was run with H-19 entered into Block 1, resulting in a significant model fit. The PCL-R was entered into Block 2. No significant improvements in the model fit were observed. The analysis was then re-run with the ordering of variables reversed. A significant improvement in model fit was observed. Only the HCR-19 possessed unique validity. The probability of violence for every one-point increase in the HCR-19 while controlling for the PCL-R was 23%. In contrast, the probability of violence for every one-point increase in the PCL-R while controlling for the
HCR-20 was -1%. Therefore, the HCR-20 added incremental validity to the PCL-R, but the converse did not. Authors concluded that the predictive validity of the HCR-20 was not dependent on the PCL-R, however the importance of considering psychopathy as part of a comprehensive risk assessment was still stressed.

Three main findings emerged: 1) when the HCR-20 \(^{V2}\) and PCL-R were evaluated for violent outcomes, they produced similar AUC’s (HCR-20 \(^{V2}\) = 0.69, PCL-R = 0.67); 2) when H7 (psychopathy) was removed from the HCR-20 \(^{V2}\), predictive accuracy did not decrease significantly; and 3) when both tools were incorporated into a regression model, only the HCR-20 \(^{V2}\) (with H7 removed) demonstrated unique predictive validity. The PCL-R did not contribute significantly to the prediction of violence over and above the HCR-20. However, the HCR-20 (with H7 removed) provided incremental validity over and above the PCL-R in predicting violence. These findings contrast with de Vogel, de Ruiter, Hildebrand, Bos and van de Ven (2004) who found that, within their forensic psychiatric sample, removal of H7 eliminates the prediction advantage of the HCR-20 over the PCL-R. These conflicting outcomes may be due to methodological differences: (i.e., meta-analysis vs. single empirical study; sample size ((8 - 1,657) vs.120); or populations (mixed sample vs. pure forensic psychiatric sample; which may have impacted the presence of psychopathology like Psychopathy in the sample).

Despite research suggesting that the superiority of the HCR-20 over the PCL-R, the decision to exclude the formal PCL-R ratings only emerged in version 3. Smith and colleagues (2014) evaluated the differences on the HCR-20 \(^{V3}\) and HCR-20 \(^{V2}\) regarding the exclusion of the PCL-R in a sample of
84 male inmates from a pre-trial correctional facility (i.e., remand facility). They hypothesised that participants would have higher scores for H7 (Personality Disorder) on the HCR-20^V3 than the HCR-20^V2, partly due to the less stringent criteria for assessing personality pathology in Version 3. Results on Presence ratings for H7 revealed that more participants received a score of 2 'Definitely Present' on the HCR-20^V3 (33.30%) than those who would have received a score of 2 on the HCR-20^V2 (8.30%). Although this study is not focused on incremental validity, it does provide insight into psychopathy Item-level changes between versions.

3.4.5.1 Conclusion

In conclusion, incremental validity has been studied both for the HCR-20 (relating to scores and SRR’s, as well as Presence and Relevance ratings in the HCR-20^V3), and between tools such as the VRAG and PCL-R/SV. Within the HCR-20, several studies have indicated that SRR’s do add to numerical scores. Research has also indicated, in the HCR-20^V3, SRR’s reflect an interaction between Presence and Relevance ratings. Between tools, the general trend appears to be that the HCR-20 adds predicative validity to the PCL-R, but in several instances, the converse is not true. This trend has been the basis for the exclusion of the PCL-R from the HCR-20^V3.

3.5 Actuarial Assessments: The Violence Risk Appraisal Guide

Literature Review

Like the HCR-20, a wealth of research is available on the VRAG. Key research resources are the three editions of the official manual, 'Violent Offenders: Appraising and Managing Risk' (Quinsey et al., 1998; Quinsey et al., 2006; Harris et al., 2015). These editions contain comprehensive
summaries of VRAG research, development procedures and administration guidelines. Furthermore, an online archive of research has been initiated but is still in the process of generation (see www.violentoffendersthree.com).

The VRAG was developed alongside the SORAG (Quinsey et al., 1998) and there have been approximately 60 studies (across 10 countries) of the VRAG and SORAG separately (Harris et al., 2015). Whilst there is an abundance of research demonstrating the validity and reliability of the VRAG (Quinsey et al., 2006), only one study is available on the VRAG-R (Rice et al., 2013). Furthermore, this study is the validation paper as part of the initial construction of the VRAG-R, and was released prior to the final VRAG-R publication (Harris et al., 2015).

Whilst validation is important for all tools that have undergone revision, it is particularly important in actuarially developed tools. As Monahan and colleagues (2005) explained, “as a rule, models constructed by using procedures that rely on associations between variables in a particular sample are apt to lose predictive power when applied to new samples. This “shrinkage” is due to capitalisation on chance associations in the original construction sample” (p. 810) (see also Silver, Smith, & Banks, 2000).

The current section of the literature review focuses on the predictive validity and inter-rater reliability of the VRAG and VRAG-R in forensic psychiatric settings, as relevant to the current research. The concurrent and incremental validity of the VRAG/-R has been addressed earlier on in this Chapter, as it represents cross-sections with the HCR-20 and PCL-R/PCL:SV bodies of literature. As a cautionary note, unlike the HCR-20 that reports both numerical and summary risk ratings, the VRAG reports Total scores, which are subsequently placed into further numerical categories, termed ‘Bins’. Both
of these outcomes are actuarial undertakings in both research and clinical contexts.

3.5.1. Predictive Validity

There have been approximately 60 replications on the predictive validity of the VRAG within correctional and psychiatric populations (Rice, Grant, & Lang, 2013). In their official summary, authors of the VRAG-R (Harris et al., 2015) note that the VRAG has demonstrated “large predictive effects”, with a mean AUC of 0.72 (based on 26 non-overlapping studies) (p. 150). These outcomes are observed when 1) the tool is used within its intended sample (i.e., male offenders), 2) for the intended outcome (i.e., violent community recidivism), and 3) when the tool is used correctly (i.e., full completion and limited variations in re-offending opportunities). Whilst the VRAG is intended to predict violent recidivism, 13 independent studies have indicated that the VRAG demonstrates predictive validity for general recidivism as well (mean AUC = 0.75). Thus, similar levels of predictive accuracy have been noted for violent and general recidivism (Harris et al., 2015).

The VRAG has consistently demonstrated predictive validity amongst forensic and correctional populations (e.g., Coid et al., 2009; Hastings, Krishnan, Tangney, & Stuewig, 2011; Kroner & Loza, 2001; Mills, Jones, & Kroner, 2005; Rice et al., 1995; Snowden, Gray, Taylor, & MacCulloch, 2007; Yessine & Bonta, 2006). The range of observed AUC’s tend to lie in the 0.64 to 0.80 upper limit. Relatively few studies have challenged this assertion (e.g., Loza, Villeneuve, & Loza-Fanous, 2002 – who found that the AUC of the VRAG in a sample of 124 Canadian offenders for violent recidivism was 0.54,
non-significant). Predictive validity has also been demonstrated amongst forensic psychiatric populations (e.g., Harris, Rice, & Cormier, 2002; Rice, 1997). The VRAG is not applicable to civil psychiatric populations because it was constructed from forensic samples, and the presence of a forensic history is assumed in the tool’s items (e.g., Item 5-8). Despite this, research has suggested that it does demonstrate validity within civil psychiatric populations (Harris et al., 2004). However, it remains that predictive validity outcomes are generally better if the demographic characteristics are similar between replication and construction samples (Singh, Grann, & Fazel, 2011).

The VRAG had demonstrated validity in both in-patient settings (Doyle et al., 2002; Hastings et al., 2011; Snowden, Gray, Taylor, & Fitzgerald, 2009; Vitacco, Gonsalves, Tomony, Smith, & Lishner, 2012) and community settings (Snowden et al., 2007; Doyle & Dolan, 2006). A summary of research on institutional violence indicated that the mean AUC of the VRAG is at least 0.70 (Harris et al., 2015).

In terms of special populations, the VRAG has demonstrated predictive validity amongst offenders with Intellectual Disability (Gray et al., 2007; Lindsay et al., 2008; Verbrugge et al., 2011) juvenile offenders (Rice et al., 2013) and ethnic minorities (Snowden et al., 2010). Although the original VRAG was not intended for use amongst sex offenders (hence, the SORAG was developed), research has shown that the VRAG can predict violent recidivism amongst sex offenders (Barbaree, Seto, Langton, & Peacock, 2001; Dempster, 1998; Langton, Barbaree, Seto, Peacock, Harkins, & Hansen, 2007; Urbaniok, Noll, Grunewald, Steinbach, & Edrass, 2006).

Unlike the HCR-20, which is applicable to both males and females, the VRAG and VRAG-R were both developed from an all-male construction
sample. Empirical support for the use of the VRAG amongst females remains scant. Of the limited research available, some research has not supported the use of the VRAG amongst females (e.g., Harris et al., 2002; Hastings et al., 2011), whilst other research has (e.g., Coid et al., 2009; Harris et al., 2004). Research by Snowden and colleagues (2010) indicated that the VRAG was more accurate amongst males. Given the limited research base, use of the VRAG with females is cautioned against, especially given that it was not designed for use amongst female populations. Studies that support the use of the VRAG amongst females are consistent with a non-gendered perspective on violence risk assessment, which suggests that risk assessment tools developed and validated for males may also be valid for use amongst females (Simourd & Andrews, 1994).

3.5.1.1 Forensic Psychiatric Populations

Snowden and colleagues (2007) evaluated the performance of the VRAG in a large sample of mentally disordered offenders in the UK. They adopted a pseudo-prospective design based on file review (at point of discharge). Violence was defined as per the official Home Office definition of violence against persons, including criminal damage endangering life, robbery, kidnap, rape and indecent assault. VRAG scores were obtained for 421 participants, from an initial sample of 996. Participants were discharged from medium secure units between 1992 and 2001, and followed-up for a minimum of 2-years. Outcome data were obtained from official records. The base rate of general recidivism during this time was 32.5%, and the average number of offences was 2.59. The average VRAG score was 3.29 (ranging from -24 to 36). Results indicated that the VRAG had good predictive validity
for both general and violent recidivism, producing large effect sizes: General recidivism (AUC = 0.74), violent recidivism (AUC = 0.77), both significant at the p < .0001 level. Thus, VRAG predicted violent reconvictions with similar accuracy as general reconvictions.

The researchers also provided AUC’s at fixed time-points. These are displayed below in Table 18. As can be seen, the highest AUC was observed for the VRAG for violent recidivism at the 6-month time point. The pattern in AUC’s indicates that AUC’s generally reduce as follow-up time increases. In shorter-periods, excellent predictive validities for both offence categories were observed (AUC > 0.84). The authors then compared AUC’s based on Hanley and McNeil’s (1983) method for comparing AUC’s for instruments scores on the same individuals. Three hundred and twenty individuals were available for this comparison. The general recidivism AUC’s were comparable to violent recidivism AUC’s.

Table 18

AUC’s for the VRAG at fixed follow-up points derived from Snowden, Gray, Taylor and MacCulloch (2007)

<table>
<thead>
<tr>
<th>Outcome Category</th>
<th>Fixed Follow-up Period</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Recidivism</td>
<td>6 Months</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>1 Year</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>2 Years</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>3 Years</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>5 Years</td>
<td>0.74</td>
</tr>
<tr>
<td>Outcome Category</td>
<td>Fixed Follow-up Period</td>
<td>AUC</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Violent Recidivism</td>
<td>6 Months</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>1 Year</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>2 Years</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>3 Years</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>5 Years</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Note. All AUC’s significant at the $p < .0001$ level. AUC’s have not been rounded up.

The authors noted that their observed reconviction rates at 5-years were lower than the reconviction rates reported in the VRAG construction sample at 7-years. Their conviction rates were approximately 50% of the expected rate. Thus, the VRAG seemed to overestimate the probability of violence, however the time-frames for comparison are different, which may influence findings.

Harris, Rice and Cormier (2002) conducted an 8-year follow-up study of data obtained in a comprehensive survey of all forensic patients from secure psychiatric units in Ontario, Canada in the 1990’s. The VRAG was completed as part of this, resulting in a prospective evaluation. The VRAG was coded from institutional files. Of this original cohort, 467 forensic patients were involved in this follow-up (including insanity acquittees and persons deemed unfit to stand trial). Most participants were males with schizophrenia who had a history of violent offending. Violence was defined as “any criminal charge for a violent offence against persons (assault, assault causing bodily harm, wounding, attempted homicide, homicide, kidnapping, forcible confinement, armed robbery, and all “hands-on sex offences”) that occurred
subsequent to the index offence. Also included were any actions that resulted in patients being returned to maximum-security…that, in the judgment of the research assistants, would otherwise have resulted in such a criminal charge” (p. 383). Outcomes are reported for the 406 of the 467 that had the opportunity to recidivate. The mean length of opportunity was 85.2 months (i.e., just under 7-years), and the average VRAG score was 2.80. Results indicated that those who reoffended had higher VRAG scores than those who did not reoffend. Accuracy was reported at a fixed follow-up point of 5-years (i.e., excluding those with less than 5-years opportunity, and considering those who recidivated after 5-years to have ‘survived’ during follow-up). 133 participants met this criteria (for whom the 5-year base rate of offending was 39%). The VRAG’s AUC was 0.80, a large effect size indicating high-accuracy. The authors noted that their obtained recidivism rates did not differ significantly from the 7-year predicted probabilities in the VRAG norms.

Few studies have found that the VRAG is a poor predictor of violence. For example, van Heesch, Jeandarme, Pouls and Vervaeke (2016) evaluated the validity of the VRAG amongst forensic psychiatric patients discharged from medium-security units in Belgium. The sample consisted of 191 Flemish patients, who were discharged to either the community or other less secure settings. The average follow-up period was 2.4 years, during which the base rate of violence was 19.90%. The comprehensive outcome category ‘violent offence’ incorporated both new reconviction(s) for violent offending (including sexual offences) based on official records, as well as all other violent incidents reported to authorities. Results indicated that the VRAG predicted outcomes at a level similar to chance, failing to discriminate between violent and non-violent recidivists. The AUC for the VRAG Total score was 0.46.
(95% Confidence Interval 0.37 – 0.56) and 0.46 (95% Confidence Interval 0.36 – 0.55) for the VRAG Bins. Thus, the VRAG failed to significantly predict any of the outcomes. Analysis of other discriminatory indices indicated that only one-third of recidivists were assessed as being high risk (sensitivity = 21.05– 33.33%). Low-risk judgments were better: Approximately two-thirds of non-recidivists were assessed as being low risk (specificity = 69.93– 71.98%). The VRAG performed poorly in predicting recidivism: (Positive Predictive Power\(^2\) (PPV) = 5.56 – 14.81%), but predicted non-recidivism well (Negative Predictive Power\(^3\) (NPV) = 78.10– 95.62%). The study was limited however by a lack of sufficient file information, resulting in items having to be dropped for 46.07% of participants. This did not exceed the threshold of four missing items however.

3.5.2 Inter-rater reliability

Within forensic settings, research has indicated that the VRAG demonstrates excellent inter-rater reliability, with excellent ICC’s in the .90’s range (e.g., Douglas et al., 2005; Gray et al., 2007; Hastings et al., 2011; Kroner & Mills, 2006; Snowden et al., 2010). Within a mixed forensic psychiatric and civil psychiatric sample, Doyle and Dolan (2006) detected an ICC of 0.99, between three raters across seven cases. Hastings and colleagues (2011) have also reported that the VRAG scores are correlated with the number of violent offences \(r = .37\).

Within forensic psychiatric samples, Thomson, Davidson, Brett, Steele and Darjee (2008) reported an ICC of 0.70; whereas van Heesch and

\(^2\) See Chapter Five, Statistical Methods, section 5.6.1.4 for an explanation on Positive Predictive Power and Negative Predictive Power.
colleagues (2016) have reported an ICC of 0.91, based on comparisons for 27 participants. Snowden and colleagues (2010) reported an ICC of 0.95 (95% CI = 0.86-0.98) for the VRAG in a forensic psychiatric setting, as rated by four psychologists across 19 cases. In the only known study on the VRAG-R, Rice and colleagues (2013) reported an ICC of .987 (single rater, absolute agreement) for the VRAG-R Total score, based on ratings by two raters for 10 participants.

3.5.3 Use of the VRAG in Australia

There is a paucity of research on the VRAG in Australia. This is likely a reflection of limited use of the tool in practice. As mentioned previously, the HCR-20 is the most commonly used violence risk assessment tool in Australia. Despite this, Chu and colleagues (2011) have evaluated the VRAG in an in-patient setting in Victoria, looking at predictive accuracy in the short-term (< 1 month) and medium-term (< 6 months). They defined any inpatient aggression as comprising of interpersonal violence (e.g., physical assault or throwing objects intending to injure) and verbal threat. Results indicated that the VRAG demonstrated poor to modest predictive accuracy of inpatient aggression at any follow-up within the 6 months. At 1-month, 3-month and 6-month follow-up points, all VRAG AUC’s were non-significant, and in the 0.53 - 0.58 range.

Verbrugge and colleagues (2011) have also conducted Australian-based research on both the HCR-20V2 and VRAG, with a focus on validating an Intellectual Disability (ID) Supplement for the HCR-20. Using a retrospective file review method, a sample of 59 community-based violent offenders with an ID were followed-up for a minimum of two years. Results
indicated that the VRAG significantly predicted both general recidivism (AUC = 0.92, \( p < .01 \), 95% CI: 0.79 – 1.00) and violent recidivism (AUC = 0.79, \( p < .001 \), 95% CI: 0.66 – 0.93). Notably, the HCR-20\(^2\) also yielded AUC’s within a similar range for general recidivism (AUC = 0.94, \( p < .001 \), 95% CI: 0.83 – 1.00) and violent recidivism (AUC = 0.80, \( p < .001 \), 95% CI: 0.63 – 0.96).

Both tools demonstrated good levels of accuracy in predicting general and violent recidivism in offenders with ID. Other research into the VRAG within Australia (e.g., Davis, 2010) has evaluated the psychometric properties, but not the predictive validity of the tool per se. Overall, there is a limited evidence-base within Australia.

### 3.5.4 Violence Risk Appraisal Guide – Revised (VRAG-R)

At the time of writing, there are no known studies evaluating the VRAG-R, except for the initial construction and validation paper by Rice and colleagues (2013). Rice and colleagues (2013) evaluated the accuracy of the original VRAG, and conducted the first known validation of the draft VRAG-R in a sample of 1261 male offenders (some of whom were participants in the original VRAG construction sample). The study was essentially a pilot study of the development of the VRAG-R, as well as a separate validation study. The sample was therefore divided into 961 participants for the construction study, and 300 for the validation sample. However, only 957 participants were used for the construction study because four had no opportunity to reoffend.

The sample composition included both violent offenders and sex offenders. This was because the VRAG-R was designed to be a combination of the VRAG with the SORAG. Participants were released from prison or
psychiatric facilities between 1960 and 1995, and followed-up again for this revised version, between 2003 and 2007. Researchers reported fixed follow-up durations (ranging from 6 months to 49 years). Participants were considered to have had the opportunity to reoffend if they were released directly into the community, or transferred to open ward psychiatric facilities, minimum security correctional facilities or half-way houses, or had unescorted day-passes. Institutional files were used to score the VRAG, as well as 30 additional variables that had shown promise as predictors of violent recidivism in previous studies. Outcomes relating to the number of offences, and the severity and imminence of violence were recorded. Participants were considered to have recidivated if they had a new criminal charge or committed an institutional act of violence that, according to the judgment of raters, would have resulted in criminal charges. Violent offences included “all assaults, sexual assaults, armed robbery, forcible confinement, threatening with a weapon, and pointing a firearm”, but excluded “possession of a weapon, robbery or arson” (p. 954).

The base rate of violent recidivism was 7% at 6 months, and 88% at 49 years. Results indicated that the original VRAG score correlated significantly with violent recidivism ($r = 0.43$, $p < .001$). The VRAG scores also predicted violent recidivism (AUC = 0.75, 95%CI = 0.72 – 0.77, $N = 1261$). This was almost identical to the AUC reported in the VRAG’s original construction (AUC = 0.76) (Quinsey et al., 2006). The VRAG Bins predicted violent recidivism (AUC = 0.74, 95%CI = 0.717 – 0.772). The VRAG-R produced an AUC of 0.76 (95% CI = 0.73 – 0.79, $N = 957$) from the development sample. The VRAG-R produced an AUC of 0.75 (95% CI = 0.69 – 0.80, $N = 300$) from the validation sample. Based on the entire sample ($N =$
1261), the VRAG-R produced an AUC of 0.76 (95% CI = 0.73 – 0.78). The study also contained a subgroup analysis, where the VRAG-R also predicted violent recidivism amongst sex offenders in isolation (AUC = 0.70, 95% CI = 0.62 – 0.76, N = 186) and juvenile offenders (AUC = 0.74, 95% CI = 0.66 – 0.83, N = 137). The authors concluded that the VRAG-R predicted violent recidivism with high levels of predictive accuracy. Although this study was conducted prior to the official release of the VRAG-R, the item composition is identical to that of the published VRAG-R (Harris et al., 2015)

3.6 Conclusion

The VRAG has established predictive validity across a range of settings and populations. Amongst forensic psychiatric populations, predictive accuracy seems to fall in the good to excellent range (i.e., AUC’s ranging from 0.60 to 0.85; Quinsey et al., 2006). Observed inter-rater reliability has also been excellent for both the VRAG and VRAG-R. Unfortunately, only one study is available to inform VRAG-R psychometrics.

In addition to providing a summary of the tool’s use, a review of a few studies that focus on the validity of the VRAG in a forensic psychiatric population for community violence outcomes has been conducted. As the VRAG is specifically designed for use in the forensic population, research is more limited in other populations and comparison to the HCR-20 for example. A large proportion of VRAG validation studies have been evaluated in conjunction with the HCR-20. It is beneficial to discuss outcomes that have been derived from the same samples simultaneously. Comparisons of the HCR-20 and VRAG are discussed below.
3.7 Comparisons between the Historical, Clinical, Risk-Management – 20 and the Violence Risk Appraisal Guide

Several studies and meta-analyses that have addressed the performance of the HCR-20, VRAG and/or PCL-R in the same sample are discussed. In terms of associations with violence, Douglas and Webster (1999) conducted a retrospective file review of 75 maximum-security offenders in Canada, and compared indices of past violence (i.e., number of past charges) to the combined H-Scale and C-Scale (i.e., HC-15; version 1) and VRAG. Results showed that the HC-15 was more strongly associated with violence than the VRAG ($r = 0.44$, $p = .001$; and $r = 0.20$, $p > 0.01$, respectively). The H-Scale had a stronger association with violence ($r = 0.50$, $p = .001$) than the C-Scale ($r = 0.30$, $p = .01$); however, both were significant. The HCR-20 Total score had a stronger association with the number of past violent offence charges ($r = 0.44$) than the PCL-R ($r = 0.41$) and VRAG ($r = 0.20$).

In a Swedish study, Grann and colleagues (2000) compared the predictive validity of the HCR-20’s H-Scale and VRAG in a sample of 404 male forensic psychiatric patients. Index forensic psychiatric reports were used for data collection, in addition to other corroborating data. Using retrospective follow-up and violent reconviction within 2 years of release as the outcome variable, two different offender groups were evaluated: Violent offenders with personality disorder, and violent offenders with Schizophrenia. In the combined sample, both the VRAG and the H-Scale significantly predicted violent reconviction with moderate sized effects. The VRAG AUC was 0.68 (95% CI = 0.63 - 0.73) and the H-Scale AUC was 0.71 (95% CI = 0.66 -0.76). The sample was then divided into a personality disorder group
and a Schizophrenia group. In both groups, the H-Scale performed better than the VRAG. In the personality-disordered cohort, the AUC’s were 0.71 and 0.68 for the H-Scale and VRAG respectively. In the schizophrenia cohort, the AUC’s were 0.66 and 0.60 respectively. This differences between AUC’s in his study was small and non-significant, with overlapping AUC confidence intervals. In addition to comparing the tools, the authors drew attention to the robust nature of historical data in predicting violence, and questioned whether clinical and risk management factors may be more important in populations where major mental illness is evident.

In a similar study, Doyle and colleagues (2002) explored the predictive validity of the HCR-20\textsuperscript{V2}, VRAG, and PCL:SV amongst 87 forensic psychiatric patients in a UK in-patient setting. Patients were observed for a period of three months, during which episodes of in-patient aggression were recorded. Tools were scored using relevant clinical and forensic reports and admission information. Only the H-Scale of the HCR-20\textsuperscript{V2} was completed. Violence was defined as per the official HCR-20\textsuperscript{V2} definition, “actual, attempted or threatened harm to a person” (Webster et al., 1997, p. 24). Results indicated that violent participants had significantly higher scores on the PCL:SV, VRAG and H-10. AUC’s for ‘any’ violence were as follows: PCL:SV Total score (AUC = 0.76, \( p < .001 \)), H-Scale Total score (AUC = 0.70, \( p < .01 \)), and VRAG Total score (AUC = 0.71, \( p < .01 \)). Therefore, the PCL:SV performed better than both the VRAG and the Historical scale in predicting violence, however AUC’s were not statistically compared.

Similar research by Chu and colleagues (2011) looked at the predictive validity of static and dynamic violence risk assessment tools (namely, the HCR-20\textsuperscript{V2}, PCL-R and VRAG, amongst others such as the Short-Term
Assessment of Risk and Treatability (START; Webster et al., 2009) and the Level of Service Inventory–Revised: Screening Version (LSI-R:SV; Andrews & Bonta, 1998) in an in-patient forensic psychiatric setting in Australia. The sample comprised 66 patients (83.3%, male). Tools were retrospectively coded from patient files. Outcomes were considered in terms of both short term (up to 1 month) and medium term (between 1 and 6 months), and were divided into interpersonal violence (e.g., physical harm, throwing of objects); verbal threat (e.g., threat to kill or cause harm); and ‘any’ in-patient aggression (a combination of the two previous categories). The results are displayed below in Table 19 for ‘any’ in-patient aggression as outcome. ‘Any’ in-patient aggression is the more comprehensive outcome measure that captures contemporary definitions of violence including threats to harm or behaviour that induces fear. AUC’s for the HCR-20 ranged from 0.59 – 0.78 across the observation periods. The HCR-20\textsuperscript{V2} produced larger AUC’s than the VRAG across all outcome categories and follow-up periods. For interpersonal violence, the HCR-20\textsuperscript{V2} produced larger AUC’s in the short-term follow-up (i.e., < 1 month) but not in the medium term (i.e., 3 and 6 months). The VRAG failed to predict any outcome category to a significant level. Notably, the START was the only measure to significantly predict outcomes at the 6-month follow-up. Overall, the START was superior to other tools in the short-to-medium term but the HCR-20 had the highest predictive accuracy for interpersonal violence over 1 month, (AUC = .78, p < .01). These results supported the use of dynamic measures over static measures in assessing in-patient aggression, in a forensic psychiatric setting, in the short-to-medium term.
Table 19

*Predictive Validity AUC’s for Any In-patient Aggression across Short-term and Medium-term derived from Chu, Thomas, Ogloff and Daffern (2011)*

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Up to 1 month AUC</th>
<th>Up to 3 months AUC</th>
<th>Up to 6 months AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCR-20</td>
<td>0.72*</td>
<td>0.78**</td>
<td>0.59</td>
</tr>
<tr>
<td>VRAG</td>
<td>0.53</td>
<td>0.58</td>
<td>0.54</td>
</tr>
<tr>
<td>PCL-R</td>
<td>0.67</td>
<td>0.72*</td>
<td>0.61</td>
</tr>
<tr>
<td>LSI-R:SV</td>
<td>0.57</td>
<td>0.60</td>
<td>0.43</td>
</tr>
<tr>
<td>START</td>
<td>0.74*</td>
<td>0.83**</td>
<td>0.74*</td>
</tr>
</tbody>
</table>

*Note. I/P denotes in-patient. * p < .05, ** p < .01.*

Using a prospective design, Doyle and colleagues (2012) evaluated the predictive validity of the HCR-20_V2 and VRAG, amongst other tools (e.g., PCL:SV), in predicting community violence. Participants (N = 114) represented a UK acute civil psychiatric sample. The tools were completed based on interviews with participants, file review and collateral report. The definition of violence included situations where threats were made in the presence of a weapon. Participants were followed-up for 20-weeks post-discharge. The base rate of violence during the follow-up period was 25.4%. Researchers used the Receiver Operating Characteristic (ROC) to evaluate the predictive validity of tools. Both tools significantly predicted post-discharge violence (HCR-20_V2 AUC = 0.68, p < .01; VRAG AUC = 0.65, p < .05). The predictive validity of the HCR-20_V2 and VRAG remained, even when controlling for age and gender. However, a high false positive rate (65%) was observed at the median split. The predictive validity of the PCL:SV was also included, and performed similarly to others, however this was non-significant (AUC = 0.62, p > 0.05). In addition to validity, researchers also
noted that the frequency of violence was significantly associated with the
HCR-20$^{V2}$ ($r = 0.435$, $p < .05$), but not with the VRAG ($r = 0.318$, $p = 0.092$).

Guy’s (2008) meta-analysis, although focused on the performance of
SPJ tools, also included a small component on the comparison of SPJ and
actuarial tools, including the VRAG. The HCR-20$^{V2}$’s performance on gross
measures for violent outcomes were as follows: Total score AUC = 0.73, SRR
AUC = 0.76. For physical violence outcomes (including sexual violence), the
AUC’s were 0.67 and 0.79 for Total scores and SRR’s respectively. The
largest effect size was found for the HCR-20 SRR with physical violence
(including sexual violence) as outcome, AUC = 0.79. The AUC for the VRAG
regarding violent behaviour and physical violence (including sexual violence)
ranged from 0.68 – 0.71. Results indicated that SPJ and actuarial methods
performed at comparable levels.

Campbell and colleagues (2009) meta-analysis involved 88
prospective studies (predominantly conducted in North America) within
correctional and forensic psychiatric samples. The sample represented a
large adult cohort (232,790 individuals for institutional violence, and 40,944
for community violence), the vast majority of whom were male. Several
violence risk assessment tools were compared: HCR-20, VRAG and PCL-R,
amongst others. Base rates of violence were 25.84% for institutional violence
and 21.73% for community violence. Average correlations with violent
recidivism were as follows: HCR-20 ($r = 0.25$), VRAG ($r = 0.27$) and PCL-R ($r
= 0.24$).

Due to variations in statistics adopted in different studies (e.g., $t$, $\chi^2$, $p$,
AUC), $Z^*$ was the metric used to indicate the magnitude of relationships.
Overall, the results indicated that actuarial tools preformed better in predicting
institutional violence, but SPJ instruments performed better at predicting community violence. The HCR-20 produced the largest effect size for predicting institutional violence, followed by the VRAG and PCL-R. For community violence, the VRAG produced the largest effect size, followed by the PCL-R and then the HCR-20. However, little variations and discrepancies in effect sizes were found amongst tools. For institutional violence, 95% Confidence Intervals overlapped for all three tools, indicating that they were not statistically different. For community violence, the HCR-20 and PCL-R 95% Confidence Intervals also overlapped. However, the HCR-20 and VRAG Confidence Intervals did not overlap, indicating that they represented different population parameters. Because the sample was characterised by a mostly male low-to-moderate risk cohort, generalisability to female and high-risk populations was limited.

These findings are somewhat contrasted by Yang and colleagues’ (2010) meta-analysis. Yang and colleagues evaluated nine commonly used tools (including the HCR-20\textsuperscript{V2}, VRAG, PCL-R and PCL:SV, amongst others) to determine which tools demonstrated predictive validity for violence. The meta-analysis drew on 28 independent studies that were truly prospective or quasi-prospective in nature. Criteria for study inclusion were: 1) post-1999 studies that involved 2) assessment of more than one risk assessment instrument within the same sample; 3) reporting of violent outcomes; with 4) enough statistical content for effect sizes to be computed. The majority of studies were based in the UK and Canada, across correctional and forensic psychiatric populations. Both institutional and community follow-ups were included. The mean follow-up period was 43.8 months, during which the detected base rate for violence was 24.9%.
Effect sizes for the predictive validity of the tools (whilst controlling for study features) fell in the moderate range (AUC’s = 0.56 – 0.71). The efficacy of the HCR-20\textsuperscript{V2}, VRAG, PCL-R and PCL:SV in predicting violent outcomes was as follows: HCR-20\textsuperscript{V2} (AUC = 0.71), VRAG (AUC = 0.68), PCL-R (AUC = 0.65) and PCL:SV (AUC = 0.68). These meta-regressions revealed that both the HCR-20 and VRAG produced larger effect sizes than the PCL-R. When heterogeneity of the samples and correlations between instruments were taken into account, only the HCR-20 Total score yielded a larger effect size than the PCL-R Total score. This advantage remained even after controlling for country of origin, design, length of follow-up and gender. The authors concluded that, overall, there were no large discrepancies in effect sizes or clinically significant meaningful differences. Although the HCR-20\textsuperscript{V2} predicted violence better than the psychopathy tools, this finding is largely confounded by the use of the PCL-R in scoring the HCR-20\textsuperscript{V2} (i.e., the PCL-R is embedded in the HCR-20\textsuperscript{V2}). Furthermore, although the VRAG predicted violence better than the psychopathy tools, controlling for study features and correlations between instruments resulted in this difference being attributed to chance. The authors rationalised these findings by suggesting that risk factors from different tools draw on the same “pools of variance that reflect a long-standing pattern of dysfunctional and aggressive interpersonal interactions and antisocial and unstable lifestyle that are common to many perpetrators of violence” (p. 759). Overall, the clinical utility of these findings is also limited by the exclusion of SRR’s.

Another meta-analysis by Singh and colleagues (2011) also looked at the predictive validity of nine risk assessment tools for general offending, including the HCR-20 (Versions 1 and 2) and VRAG as well as the PCL-R.
The review was based on 68 studies representing 88 independent samples \((N = 25,980)\), across 13 countries. The VRAG had one of the highest median AUC’s \((AUC = 0.74)\), compared to an AUC of 0.70 for the HCR-20 and 0.66 for the PCL-R. In fact, of the nine tools, the PCL-R produced the lowest overall predictive validity. This research contrasts with Yang and colleagues (2010), as substantial differences in the predictive validity of tools was observed. The study was limited by not screening for the methodological quality of included studies.

In their combined UK forensic psychiatric and civil psychiatric sample, Dolan and Doyle (2006, mentioned previously) examined the predictive validity of the VRAG, HCR-20\(^{V2}\) and PCL:SV. Following discharge from in-patient care, 112 participants were followed-up for 24-weeks in the community. The target outcome was community violence, as detected through official records, self-report and collateral information. The average time to follow-up was approximately 5.5 months, during which the base rate of violence was 19%. AUC’s were as follows: VRAG \((AUC = 0.657)\); HCR-20\(^{V2}\) Total score \((AUC = 0.797)\); H-Scale \((AUC = 0.675)\); PCL:SV \((AUC = 0.687)\), all significant at the \(p < .05\) level. Therefore, all tools were predictive of community violence, however the largest AUC was observed for the HCR-20\(^{V2}\). The authors found that the HCR-20\(^{V2}\) continued to be significantly predictive of post-discharge violence when confounding variables such as age, gender, length of in-patient stay and forensic status were accounted for, and the most robust predictor of community violence overall.

Snowden and colleagues (2010) compared the predictive validity of the VRAG and HCR-20\(^{V2}\) in a sample of forensic psychiatric patients from medium-security facilities in the UK. Unlike other studies where the sample is
predominantly Caucasian males, this study focused on the predictive validity of these tools within ethnic minorities. For the HCR-20, 1016 participants were available for analyses, compared to 579 for the VRAG. The study methodology was pseudo-prospective, with tools being completed at discharge, based on file review. Outcome data were also sourced from official records. The authors conducted a fixed 2-year follow-up period, during which the base rate of violence was 12%. The observed AUC’s for the VRAG ranged from 0.74 – 0.79, indicating large effect sizes across white and minority groups. AUC’s did not differ significantly across groups (based on Hanley & McNeil’s (1982) method). Observed AUC’s for the HCR-20 ranged from 0.66 - 0.72, with no statistical difference between groups. Results indicated that within forensic psychiatric ethnic minorities, the performance of the VRAG was superior to that of the HCR-20, however both were significant.

Ho, Thomson and Darjee (2009) evaluated the predictive validity of the VRAG, HCR-20\textsuperscript{V2} (historical component) and PCL:SV within mentally disordered offenders discharged from a medium-security psychiatric unit in Scotland. The sample comprised 96 patients (the vast majority of whom were male, with co-morbid psychiatric disorders), who were discharged from the unit between 2001 and 2004. Approximately half of the sample was discharged to the community, the rest to other facilities. However, the vast majority (> 80%) had community access during follow-up. Violence was defined according to the official HCR-20\textsuperscript{V2} definition (Webster et al., 1997). Participants were followed-up for 2-years, during which the base rate of violence was 4.2% for serious violent offences, and 40.6% for minor violent offences.
The highest AUC was observed for the VRAG (AUC = 0.68), followed by the PCL:SV (AUC = 0.63) and then the H-Scale (AUC = 0.61). This indicated moderate accuracy across tools in predicting any form of violence. This pattern of performance was also observed for minor violent incidents (AUC’s = 0.70, 0.64 and 0.62 for the VRAG, PCL:SV and H-Scale respectively). The VRAG and H-Scale performed similarly in predicting serious violence, AUC’s = 0.74 and 0.74, respectively. This was superior to the PCL:SV’s prediction of serious violent incidents, AUC = 0.66. Thus, the VRAG was the best predictor of any violence and minor violence. The H-Scale was slightly superior for serious violence, although only marginally. The study was limited however by its reliance on archival data collection and a small sample size.

Thomson and colleagues (2008) also compared the VRAG, HCR-20\textsuperscript{V2} (H-Scale) and PCL-R in a sample of 135 forensic psychiatric patients with schizophrenia in a high-security facility in Scotland. Tools were scored retrospectively from file review, and outcomes were informed by official criminal records and recorded incidents of aggression over an 8-10 year period. Some patients remained in hospital during this time, whilst 54 were discharged. The average follow-up period was 8.74 Years. The violent recidivism base rate was 5%.

Reported AUC’s for violence recidivism were as follows: VRAG (AUC = 0.80), H-Scale (AUC = 0.79), PCL-R (AUC = 0.83), all significant at the $p < .01$ level except for the H-Scale, which was significant at the $p < .05$ level. Reported AUC’s for general recidivism were as follows: VRAG (AUC = 0.76), H-Scale (AUC = 0.76), PCL-R (AUC = 0.73), all significant at the $p < .01$ level. Outcomes for serious incidents (i.e., “any aggressive incident resulting in
death or injury requiring hospital treatment, any sexual incident involving contact with the victim, or any fire-setting” p. 176.) were poor. The AUC’s for the VRAG, H-Scale and PCL-R were 0.55, 0.53 and 0.54, all non-significant. The authors concluded that the tools demonstrated moderate to high predictive accuracy for both general and violent recidivism, but were ineffective at predicting serious offences.

Tengström (2001) compared the performance of the VRAG and the HCR-20\textsuperscript{V2}'s H-Scale for predicting violent recidivism in a sample of 106 Swedish male forensic psychiatric patients with schizophrenia and histories of violent offending. Tools were retrospectively scored, based on file review. The authors reported that the definition of violence adopted closely matched the official VRAG definition, but the definition was not specified. Participants were followed-up for a period of 7-years (mean follow-up period being 86 months), during which 29% of the sample violently recidivated. Results for predictive validity analyses were as follows: H-Scale AUC = 0.76 (95% CI = 0.66 – 0.83), VRAG AUC = 0.68 (95% CI = 0.65 – 0.82). Therefore, both the H-Scale and VRAG demonstrated moderate predictive validity for violent outcomes, however the H-Scale revealed better accuracy. These results should be interpreted cautiously because the VRAG was heavily modified (e.g., items dropped or not scored as per the official administration instructions).

One study by Hogan and Olver (2016) has evaluated and compared the predictive validity of the HCR-20\textsuperscript{V3}, VRAG-R and PCL-R for inpatient aggression. Using archival review, pre-treatment and post-treatment ratings were completed for 99 forensic inpatients admitted to a maximum security facility in Canada between January 2005 and July 2008. Results indicated that, for pre-treatment scores, the HCR-20\textsuperscript{V3} Total Presence scores, Total
Relevance scores, and SRR’s (Case Prioritisation/Future Violence) demonstrated significant predictive validity (AUC’s = 0.76, 0.70 and 0.68 respectively), whereas the VRAG-R Total scores (AUC = 0.60, p > .05) and PCL-R Total scores (AUC = 0.63, p > .05) did not. The study was limited however by an archival methodology and a lack of clear and specific timeframes around outcome behaviours. Table 20 provides a summary of the studies discussed herein.

Table 20

*Studies comparing the Predictive Validity of the HCR-20\(^V2\), VRAG and PCL-R/SV.*

<table>
<thead>
<tr>
<th>Authors</th>
<th>Sample</th>
<th>Country</th>
<th>Version</th>
<th>ROC-AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grann, Belfrage, &amp; Tengstrom (2000)</td>
<td>Forensic</td>
<td>Sweden</td>
<td>H(^V2)-Scale</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>VRAG</td>
<td>0.68</td>
</tr>
<tr>
<td>Doyle, Dolan, &amp; McGovern (2002)</td>
<td>Forensic</td>
<td>UK</td>
<td>H(^V2)-Scale</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>VRAG</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL:SV</td>
<td>0.76</td>
</tr>
<tr>
<td>Chu, Thomas, &amp; Daffern (2011)</td>
<td>Forensic</td>
<td>Australia</td>
<td>HCR-20(^V2)</td>
<td>0.59-0.72</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>VRAG</td>
<td>0.53-0.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL-R</td>
<td>0.61-0.72</td>
</tr>
<tr>
<td>Doyle, Carter, &amp; Shaw, &amp; Dolan (2012)</td>
<td>Civil</td>
<td>UK</td>
<td>HCR-20(^V2)</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>VRAG</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL:SV</td>
<td>0.62</td>
</tr>
<tr>
<td>Guy (2008)</td>
<td>Mixed Sample</td>
<td>Intl.</td>
<td>HCR-20(^V2)</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SRR</td>
<td>0.76</td>
</tr>
<tr>
<td>Authors</td>
<td>Sample</td>
<td>Country</td>
<td>Version</td>
<td>ROC-AUC</td>
</tr>
<tr>
<td>------------------------------</td>
<td>----------------</td>
<td>---------</td>
<td>--------------</td>
<td>---------</td>
</tr>
<tr>
<td>Campbell, French, &amp; Gendreau</td>
<td>Mixed</td>
<td>Intl.</td>
<td>HCR-20</td>
<td>0.22¹</td>
</tr>
<tr>
<td>&amp; Gendreau (2009)</td>
<td>Sample</td>
<td></td>
<td>VRAG</td>
<td>0.27¹</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL-R</td>
<td>0.32¹</td>
</tr>
<tr>
<td>Yang, Wong, &amp; Coid (2010)</td>
<td>Mixed</td>
<td>Intl.</td>
<td>HCR-20 V²</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Sample</td>
<td></td>
<td>VRAG</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL-R</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL:SV</td>
<td>0.68</td>
</tr>
<tr>
<td>Singh, Grann, &amp; Fazel (2011)</td>
<td>Mixed</td>
<td>Intl.</td>
<td>HCR-20 &amp; V²</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Sample</td>
<td></td>
<td>VRAG</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL-R</td>
<td>0.66</td>
</tr>
<tr>
<td>Doyle &amp; Dolan (2006)</td>
<td>Civil &amp;</td>
<td>England</td>
<td>H10</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Forensic</td>
<td></td>
<td>HCR-20 V²</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Psychiatri</td>
<td></td>
<td>VRAG</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL:SV</td>
<td>0.69</td>
</tr>
<tr>
<td>Snowden, Gray, &amp; Taylor (2010)</td>
<td>Forensic</td>
<td>UK</td>
<td>HCR-20</td>
<td>0.66 – 0.72</td>
</tr>
<tr>
<td></td>
<td>Psychiatri</td>
<td></td>
<td>VRAG</td>
<td>0.74 – 0.79</td>
</tr>
<tr>
<td>Ho, Thompson, &amp; Darjee (2009)</td>
<td>Forensic</td>
<td>Scotland</td>
<td>H-Scale V²</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Psychiatri</td>
<td></td>
<td>VRAG</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL:SV</td>
<td>0.63</td>
</tr>
<tr>
<td>Thomson, Davidson, &amp; Brett, Steele, &amp; Darjee (2008)</td>
<td>Forensic</td>
<td>Scotland</td>
<td>VRAG</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Psychiatri</td>
<td></td>
<td>H-Scale</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL-R</td>
<td>0.83</td>
</tr>
<tr>
<td>Tengström (2001)</td>
<td>Forensic</td>
<td>Sweden</td>
<td>H-Scale</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Psychiatri</td>
<td></td>
<td>VRAG</td>
<td>0.68</td>
</tr>
<tr>
<td>Authors</td>
<td>Sample</td>
<td>Country</td>
<td>Version</td>
<td>ROC-AUC</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------</td>
<td>---------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>Hogan &amp; Olver (2016)</td>
<td>Forensic</td>
<td>Canada</td>
<td>HCR-20$^{\text{V3}}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Psychiatric</td>
<td></td>
<td>Total score</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SRR</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>VRAG-R</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PCL-R</td>
<td>0.63</td>
</tr>
</tbody>
</table>

*Note. Intl = International study. † Effect sizes are based on $Z^*$ statistic, which represents $r$ weighted by sample size.*

### 3.8 Conclusion

Multiple research studies have shown that the predictive validity of SPJ and actuarial approaches are comparable. Within this, the performance of the HCR-20 and VRAG are also comparable, although in some studies the predictive validity of the VRAG is superior, and in other studies the predictive validity of the HCR-20 or its constituent scales, is superior. This literature review has discussed research on the HCR-20 and VRAG, as tools representative of the SPJ and actuarial approaches to violence risk assessment. Whilst this review has been comprehensive, it is not an exhaustive review of available research on the HCR-20 and VRAG. For the purposes of the current research, this literature review has focused on specific psychometric properties (i.e., predictive, concurrent and incremental validity, and inter-rater reliability), with an emphasis on studies in forensic psychiatric populations. Overall, the established validity and reliability of the HCR-20 and VRAG is evident, albeit there are minor variations in study methodologies.
One clear outcome of this literature review is that, whilst there is a plethora of research available on the HCR-20 (Versions 1 and 2) and the VRAG, there is relatively limited research available on the HCR-20$^{V3}$ and the VRAG-R, and none in Australia. All known studies of the current versions of these tools have been reviewed. There exists a need to evaluate the basic psychometric properties of the HCR-20$^{V3}$ and VRAG, including their predictive, concurrent and incremental validity, and inter-rater reliability. In Australia, relative to other countries, there is a lack of research evaluating the HCR-20 (for exceptions see Campbell, 2007; Campbell, Shepherd, & Ogloff, in press; Ogloff et al., 2005). This is concerning given that it remains the most commonly used violence risk assessment tool in Australia (Singh et al., 2014). Its continued use must be evidence based. Guy’s (2008) research indicated that the country the HCR-20 is evaluated in does not influence its validity. However, this can only be determined following empirical research. Research must show that such outcomes are indeed observed here. Consequently, the authors of the HCR-20$^{V3}$ have called for it to be evaluated across countries (Douglas et al., 2013). Furthermore, established conclusions on the psychometric properties of tools must be based on an accumulation of research, not singular studies. As de Vogel and colleagues (2014) note, “while the empirical evidence for the HCR-20 is strong and convincing, the psychometric properties and clinical value of the HCR-20$^{V3}$ still needs to be established…future validation studies on the final version of the HCR-20$^{V3}$ are definitely needed. Both retrospective and prospective research should focus on inter-rater reliability, concurrent and predictive validity and clinical applicability for the assessment and management of violence risk in different settings and countries” (p. 110 - 120).
The current study aims to add to the emerging body of research exploring the psychometric properties of the most important SPJ and actuarial violence risk assessment tools, the HCR-20^{V3} and the VRAG-R in Australia. This is necessary if the revised tools are to be presented as evidence based forms of violence risk assessment.
Chapter Four

Research Methodology

4.1 Introduction

The first section of this chapter describes the research setting and relevant legislation. An overview of the research design, aims, hypotheses and procedures is then provided. An explanation of outcome definitions and measures is presented, followed by a description of the administration of the violence risk assessment tools. The chapter concludes with a description of the sample demographics.

4.2 Research Setting

4.2.1 The Victorian Institute of Forensic Mental Health

The Victorian Institute of Forensic Mental Health (VIFMH; also known as Forensicare) is the statewide statutory authority for the provision of forensic mental health services in Victoria, Australia (VIFMH, 2015). Forensicare’s mission is to “provide effective mental health services in a safe and secure environment to people who have both a mental disorder and a history of criminal offending or who present a serious risk of such behavior” (VIFMH, 2005, pg 1).

Established in 1997 under the previous Mental Health Act 1986 (VIC)\(^3\), Forensicare provides clinical services across inpatient, outpatient and specialist treatment contexts at several different sites in Victoria, Australia.

\(^3\) Given the pseudo-prospective nature of the study, it is noted that the Mental Health Act 1986 (VIC) was the act effective during the real-time admission periods. Although the current act, Mental Health Act 2014 (VIC) is still referred to as the relevant sections of the act remain unchanged.
One of these sites, the Thomas Embling Hospital (TEH), is Victoria’s only secure forensic mental health hospital.

Prior to TEH, Victorian forensic psychiatric patients were located at the Rosanna Forensic Psychiatric Centre (RPFC). Established in 1989, the RPFC had an 18-bed acute unit (known as ‘M6’) and a 20-bed rehabilitation unit (known as ‘M5’) (Daffern, Mayer, & Martin, 2004). The RPFC was disestablished in April 2000, upon which patients were transferred to the newly established TEH.

TEH comprises 116 beds distributed across seven units. Patients are allocated to units based on gender, the severity of their mental illness and the level of security required. The units are: Argyle and Atherton (acute male units in a high security setting); Barossa (acute female unit in a high security setting); Bass (sub-acute male unit in a high security setting); Canning (sub-acute male unit in a medium-high security setting); Daintree (rehabilitation mixed-gender unit in a low-medium security setting); and Jardine (rehabilitation mixed-gender unit in a low secure setting) (VIFMH, 2016).

The ‘Acute Care Program’ is designed to provide intensive intervention to patients in the acute phase of mental illness. The program is delivered to patients on Argyle, Atherton, Barossa and Bass. The ‘Continuing Care Program’ is designed for patients who have maintained a stabilised mental state and level of independence over a sustained period of time, but who still require secure care due to chronic symptomatology or the presence of behaviours that represent a risk to the community (VIFMH, 2016).
4.2.2 Legislative Framework

Forensicare operates within a complex legislative framework. While Forensicare’s responsibilities are governed by the *Mental Health Act 2014* (VIC), services are also determined by the *Crimes (Mental Impairment and Unfitness to be Tried Act, 1997)* (VIC); *Sentencing Act 1991* (VIC) and the *Corrections Act 1991* (VIC).

There are several legal pathways through which a patient may be admitted to TEH. Patients can either be ‘security patients’ or ‘involuntary patients’ or ‘forensic patients’. Security patients can fall under the *Mental Health Act 2014* (VIC) or *Sentencing Act 1991* (VIC). Under s. 16(3)(b) of the *Mental Health Act 2014* (VIC), mentally ill prisoners can be certified and transferred to TEH under a Restricted Hospital Order as security patients. Under s. 93(1)(e) of the *Sentencing Act 1991* (VIC), the court may transfer a mentally ill offender to TEH as a security patient by way of sentence on a Hospital Security Order. Security patients are typically transferred to TEH following psychiatric treatment within prison mental health units, such as the Acute Assessment Unit at the Metropolitan Assessment Prison (male correctional facility) and the Marrmak Unit at the Dame Phyllis Frost Centre (female correctional facility).

Involuntary patients can fall under the *Mental Health Act 2014* (VIC) or *Sentencing Act 1991* (VIC). Under s. 12 of the *Mental Health Act 2014* (VIC), prisoners placed in TEH whose sentence has expired may be detained as involuntary patients in TEH if they: 1) Continue to display symptoms of mental illness and require treatment; and/or 2) are assessed as being a risk to the community and ongoing hospitalisation is required for the purposes of continued treatment in a secure environment. Civil psychiatric patients (i.e.,
non-forensic patients) who are mentally ill and cannot be managed in less restrictive mainstream psychiatric facilities can also be transferred to TEH as involuntary patients under s. 12 of the Mental Health Act 2014 (VIC). Furthermore, under s. 93(1)(d) of the Sentencing Act 1991 (VIC), the court may transfer a mentally ill offender to TEH from court as an involuntary patient on a Hospital Order, instead of imposing a sentence. Section 56 of the Corrections Act 1986 (VIC) reiterates these aforementioned pathways for the transfer of prisoners to an approved mental health service (i.e., TEH).

Under the Crimes (Mental Impairment and Unfitness to be Tried Act 1997) (VIC), individuals found Not Guilty by Reason of Mental Impairment (NGRMI) or who are deemed unfit to stand trial may be given a custodial disposition (i.e., Custodial Supervision Order (CSO)) to remain in TEH. These patients are referred to as ‘forensic patients’.

In summary, there are several legal pathways through which a patient can be admitted to TEH as either security, involuntary or forensic patients. Upon discharge, patients are either returned to prison; transferred to a Secure Extended Care Unit (SECU) or other civil psychiatric facility; or released directly into the community.

4.3 Research Overview

4.3.1 Research Design

The current research was based on a pseudo-prospective file review of 100 Australian adult forensic psychiatric patients who were hospitalised at TEH between April 2000 and December 2010⁴. The study was pseudo-

⁴ Three participants included in the study were transferred from RPFC to TEH. The earliest admission date is October 1996.
prospective because tools were scored in real-time in 2015, based on discharge information from 2000-2010. During their hospital stay, participants received violence risk assessments using either the Historical Clinical Risk-Management - 20 Version 2 (HCR-20\textsuperscript{V2}) and/or the Violence Risk Appraisal Guide (VRAG).

Participants were subsequently discharged to either prison or community. As the research question relates to the predictive validity of the violence risk assessment tools in community settings, participants transferred back to prison upon discharge were considered to have ‘delayed entry’ into the community.

For participants discharged directly into the community, support services arranged through discharge planning is part of standard care (e.g., housing, medical and ongoing psychiatric care within the community). However, in the current study, the extent of care was limited to supervision under an Area Mental Health Service (AMHS); placement at a Community Care Unit (CCU); or placement in a supported accommodation facility. Participants discharged to SECU’s were not included in the study. This is due to the long-term length of stay typically associated with SECU placements (Department of Health and Human Services, 2014).

Researchers then retrospectively conducted violence risk assessments using updated versions of the previously utilised violence risk assessment tools: The Historical Clinical Risk-Management - 20 Version 3 (HCR-20\textsuperscript{V3}) and/or the Violence Risk Appraisal Guide - Revised (VRAG-R). Therefore, participants received violence risk assessments using previous and current forms of the HCR-20 (versions 2 and 3) and/or the VRAG (original and revised).
Recidivism data were provided for the period 1\textsuperscript{st} of April 2000 – 31\textsuperscript{st} of January 2013. This enabled a maximum follow-up period of 12 years and 10 months. During this time, movements in and out of correctional systems were recorded, as were periods of hospitalisation. This enabled the calculation of ‘time at risk’ in the community post-discharge from TEH. Predictive validity was assessed in terms of the tools ability to predict post-release violent recidivism over the follow-up period. Estimates of predictive validity also extended to ‘any recidivism’ and ‘non-violent recidivism’ (see descriptions in section 4.9.3).

4.3.2 Research Aims

The current research had three aims: First, to evaluate the predictive validity of the HCR-20\textsuperscript{V3} and the VRAG-R; second, to compare the predictive validity of these tools to their predecessors (i.e., the HCR-20\textsuperscript{V2} and the VRAG respectively); and third, to compare the predictive validity of the HCR-20\textsuperscript{V3} and VRAG-R. Although the focus of the research is predominantly on the HCR-20\textsuperscript{V3}, given its widespread use in Australia, the VRAG-R was also included to: 1) Add to the scant evidence base regarding the predictive validity of the VRAG-R in Australia and internationally; 2) assist in comparing the SPJ and actuarial approaches to violence risk assessment, particularly as the VRAG represents one of the most commonly used actuarial tools (Singh et al., 2014). Furthermore, the VRAG presented an ideal choice, given that VRAG data was available in the comparative datasets (see Section 4.6.4).
4.3.3 Hypotheses

It was hypothesised that:

1. **Association:** There will be a significant and positive association between violent recidivism and scores on the HRC-20\textsuperscript{V3} and VRAG-R. The HCR-20\textsuperscript{V3}'s Total scores and SRR's will correlate significantly with violent recidivism. The VRAG-R's Total scores and Bins will correlate significantly with violent recidivism. The association between frequency of violent offence charges post-release and scores on the HCR-20\textsuperscript{V3} and VRAG-R was also investigated.

2. **Concurrent Validity:** The HCR-20\textsuperscript{V3} and VRAG-R will demonstrate acceptable concurrent validity (i.e., significant correlations of $r > .50$, as per Munro, 2001). Previous and current versions of the tools will correlate significantly (i.e., the HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} will correlate significantly, as will the VRAG and VRAG-R). Furthermore, the HCR-20\textsuperscript{V3} and VRAG-R would also correlate significantly, as will static measures (i.e., the HCR-20\textsuperscript{V3} H-Scale and VRAG-R). Finally, the HCR-20\textsuperscript{V2} scales will correlate significantly with the HCR-20\textsuperscript{V3} scales.

3. **Prediction:** The ability of the HCR-20\textsuperscript{V3} and the VRAG-R to predict future violence will be greater than chance. The observed AUC’s will be similar AUC’s produced by the HCR-20\textsuperscript{V2} and VRAG, and non-significantly different.

4. **HCR-20\textsuperscript{V3} Predictive strength of scales:** The H-Scale will be the strongest predictor of post-release violence over the entire follow-up
period. However, when controlling for time at risk, the C-Scale and R-Scale will emerge as the strongest predictors of post-release violence over the entire follow-up period.

5. *Discrimination*: Participants scoring above the median score on the HCR-20$^{V3}$ and VRAG-R will be significantly more likely than those scoring below the median to 1) recidivate generally; and 2) recidivate violently, during the entire follow-up period.

6. *Survival Time*: There would be statistically significant differences in survival time for general and violent recidivism between HCR-20$^{V3}$ and VRAG-R scores (participants scoring above or below the median Total score) and risk categories (HCR-20$^{V3}$ SRR’s and grouped VRAG-R Bins).

7. *Incremental Validity*: Incremental validity will be assessed within the HCR-20$^{V3}$, and between tools in predicting violent recidivism. Within the HCR-20$^{V3}$, Relevance ratings will demonstrate incremental predictive validity over the Presence ratings. Between tools, it was hypothesised that the PCL-R would not add significant incremental predictive validity to the HCR-20$^{V3}$ (Total scores and SRR’s) nor the VRAG-R Total scores. It was also hypothesised that the HCR-20$^{V3}$ would add incremental validity to the VRAG-R, but the VRAG-R would not add incremental validity to the HCR-20$^{V3}$. 
8. *Inter-rater Reliability:* The HCR-20\textsuperscript{V3} and VRAG-R will demonstrate acceptable inter-rater reliability (ICC > .40 as per Fleiss, 1981).

4.4 Sources of Data

Six sources of data were obtained in order to conduct the study, as follows:

1. Archival data in the form of the participants’ TEH files.
2. Comparative datasets containing HCR-20\textsuperscript{V2} and VRAG scores.
3. Recidivism data from Victoria Police in the form of official records extracted from the Law Enforcement Assistance Program (LEAP).
4. Reincarceration data and forensic psychiatric rehospitalisation data from the Department of Justice Prisoner Information Management System (PIMS).
5. Civil psychiatric rehospitalisation data from the Client Management Interface (CMI) database.
6. Coronial data from the National Coronal Information System (NCIS).

4.5 Ethics

Prior to the commencement of data collection, ethics approval for the research study was received from: The Swinburne University Human Research Ethics Committee (SUHREC); Department of Justice Human Research Ethics Committee (JHREC); Victoria Police Research Coordinating Committee (RCC); and the National Coronal Information System (NCIS; previously known as the National Coroners Information System). Forensicare’s Research Committee also approved the project. The project formed part of a larger existing research project within Forensicare.
Amendment requests were submitted to the relevant ethics committees at different times throughout the research process regarding: The extension of data collection periods and access periods, and the addition of researchers to the research team. These requests were approved by all relevant ethics committees. Ethics approval notices and letters of support are displayed in Appendix F.

4.5.1 Ethical considerations

Ethical clearance was granted to collect the data without the data being de-identified and without participant’s consent. Data were initially identifiable, although was de-identified following data linkage. Consent was not obtained for several reasons. Firstly, due to the retrospective nature of the study, it was anticipated that individuals may not be contactable due to death or relocation. Notably, the passage of time from the first admission to TEH (April 2000) to the submission of ethical applications was approximately 14 years and 9 months. Secondly, due to the nature of the sample, it was also anticipated that the last known address may actually be a correctional facility or housing support service.

It was also considered that forensic psychiatric patients may be more difficult to contact than other groups. This is due to a high rate of itinerancy, unemployment and limited social ties amongst this group. It was anticipated that efforts to contact participants may therefore require disclosure of Forensicare service provision. This was considered to be intrusive, especially given the stigma associated with involvement in forensic mental health services. Finally, logistical issues in contacting participants (the number of participants originally anticipated was $N = 186$) was also prohibitive. Given
these considerations, it was decided that obtaining consent would be impracticable.

The research project was funded by Swinburne University of Technology and the Centre for Forensic Behavioural Science. There were no commercial interests or conflicts of interest.

4.6 Accessing the data

4.6.1 Archival Data

Participant files were transferred from TEH to another Forensicare location, the Community Forensic Mental Health Service (CFMHS), which is co-located with the Centre for Forensic Behavioural Science (CFBS). Most of the files were transported to the CFBS, however for participant’s who had a large number of files, the student researcher completed the scoring of the HCR-20\textsuperscript{V3} and VRAG-R at TEH. In total, 875 files were processed. The average number of files was 8.75 per participant. Scoring the HCR-20\textsuperscript{V3} and VRAG-R required on averaged 2-3 hours per participant. The student researcher scored all violence risk assessment instruments for all participants. Hard-copy versions of the completed HCR-20\textsuperscript{V3}’s and the VRAG-R were stored securely in a locked filing cabinet at the CFBS. Data were entered into a SPSS dataset, which was securely stored on the Forensicare database. Files were then stored in a locked filing cabinet.

4.6.2 Victoria Police Data

Recidivism data were obtained from official Victoria Police records. A 100% match rate was found (i.e., Victoria Police was able to match all individuals on the list of participants provided with the LEAP database). Three
types of data were received: 1) Violent and non-violent offending based on charges; 2) contacts with police (including person checks, vehicle checks and search warrants); and intervention orders (IVO’s). Recidivism data were provided for the period 1st of April 2000 – 31st of January 2013. This enabled a maximum follow-up period of 12 years and 10 months.

4.6.3 Department of Justice Data and Regulation

Two types of data were received from the Department of Justice: 1) Custody episodes (includes prison and police custody) and; 2) Orders. Custody episodes were used to inform participant’s movements in and out of the correctional system during the follow-up period, thus informing time at risk. Data on various orders instated during the follow-up period were provided, including parole orders, Community Based Orders (CBO’s), and Intensive Correction Orders (ICO’s). This data was used for descriptive purposes.

4.6.4 Comparative Data (Sourcing the sample)

In order to satisfy the research aim of comparing the HCR-20V3 and the VRAG-R to their respective earlier versions, databases containing scores for HCR-20V2 and the VRAG were sought from previous research on patients at TEH. Two datasets were identified and requested: Chu (2010) and Campbell (2007).

The Chu (2010) dataset contained 70 participants who had received both a HCR-20V2 and a VRAG assessment whilst inpatients at TEH. This dataset was considered the primary dataset because it contained both HCR-20V2 and VRAG scores.
Given that a maximum of 70 participants could be sourced from the Chu dataset, the Campbell dataset was also included as a means of increasing sample size. The Campbell dataset contained 116 participants who received a HCR-20\textsuperscript{V2} assessment whilst inpatients at TEH. Both the Chu (2010) and Campbell (2007) datasets also contained Psychopathy Checklist – Revised (PCL-R) Total scores.

The combination of the datasets resulted in a pool of $N = 186$ participants for the HCR-20 component of the study and $N = 70$ for the VRAG component of the study. Both datasets were originally created in a research capacity for the purposes of doctoral theses.

4.6.5 Refining the Sample

The sample of $N = 186$ participants was then refined through exclusionary criteria. Firstly, any matching cases across the two datasets were excluded ($N = 13$). Due to the Chu (2010) dataset being the primary dataset, duplicated cases were removed from the Campbell (2007) dataset.

Coronial data were then used to identify any participants who had become deceased during the follow-up period. It was necessary to exclude these participants because the absence of post-release recidivism data could be incorrectly classified as the participant not recidivating, when in fact the participant did not have the opportunity to recidivate because he/she was deceased. Eleven participants were removed from the study on this basis.

As the primary interest of the study was to assess the predictive validity of violence risk assessment tools in the Victorian community, the decision was made to remove participants who were deported on discharge, or discharged to SECU’s. Nineteen participants were removed from the
sample on this basis. Finally, \( N = 19 \) participants were excluded from the study due to an inability to locate all relevant inpatient files and information.

This sample refinement process is summarised briefly below in Table 21.

Table 21

*Reasons for Exclusion from the Sample*

<table>
<thead>
<tr>
<th>Reason</th>
<th>Number of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching cases</td>
<td>13</td>
</tr>
<tr>
<td>Deceased participants</td>
<td>11</td>
</tr>
<tr>
<td>Deported/SECU discharge location</td>
<td>19</td>
</tr>
<tr>
<td>Inability to locate files/missing information</td>
<td>19</td>
</tr>
</tbody>
</table>

*Note.* Final dataset, \( N = 100 \).

Through the course of refining the sample, 33 participants were removed from the initial sample. This resulted in a final sample pool of \( N = 124 \) (\( N = 52 \) on the Chu (2010) dataset and \( N = 72 \) on the Campbell (2007) dataset.

As the Chu (2010) dataset was the primary dataset, each of the \( N = 52 \) participants on the dataset received a HCR-20\textsuperscript{V3} assessment. Participants were then randomly sampled from the Campbell dataset (using the Excel Randomisation Function) in order to increase the sample size.

The \( N = 52 \) participants from the Chu (2010) dataset included male and female participants who also received an original VRAG assessment. The decision was made to exclude female participants from this sample in the current study, due to the following reasons: 1) A limited number of females (\( N = 10 \)) were available on the dataset, thus preventing against the computation of separate statistical analyses for females; 2) females were not included in the construction sample of the original VRAG and the VRAG-R; and 3)
evidence suggesting that gender differences exist in the predictive validity of the VRAG, in which the VRAG is not effective in predicting violence risk amongst females (Hastings, Krishnan, Stuewig, & Tangney, 2008; see Chapter 2 on the VRAG Literature Review for more information).

Due to time constraints on the data collection period, the decision was made to conclude data collection at \( N = 100 \). This was deemed a sufficient sample size for the power required to run the statistical analyses. Furthermore, given that the datasets in isolation had been used in previous doctoral research projects assessing violent recidivism, it was likely that in combination they would provide a sufficient number of participants.

This resulted in a final sample size of \( N = 42 \) for the VRAG component of the study and \( N = 100 \) for the HCR-20 component of the study. In other words, of the final 100 participants in the original comparative dataset, 42 had received both a HCR-20\(^{V2}\) and a VRAG assessment, and 58 had received a HCR-20\(^{V2}\) assessment only.

### 4.6.6 Completing the Datasets

The Chu (2010) dataset had HCR-20\(^{V2}\) H-Scale items scored using all available information at discharge. The C-Scale scores were coded within the first seven days post admission to TEH, and the R-Scale scores were coded during admission. Therefore, the rating of scales at discharge was required. A research assistant coded all the C-Scale scores on discharge (2 months prior to date of discharge), as well as R-Scale scores on discharge. Figure 1 provides a visual illustration of the three datasets used to complete the current research.
**Figure 1  Dataset Composition**

**HCR-20\textsuperscript{V2}**

**Chu (2010) Dataset**

- H-Scale (H1-H10) at Discharge
- H-Scale Total Score

**Campbell (2007) Dataset**

- H-Scale (H1-H10) at Discharge
- H-Scale Total Score

- C-Scale (C1-C5) at Intake
- C-Scale (C1-C5) at Discharge*
- C-Scale Total Score
  (at Intake and at Discharge)

**HCR-20\textsuperscript{V3}**

**Brookstein (2016) Dataset**

- H-Scale (H1-H10) at Discharge
- C-Scale (C1-C5) at Discharge
- R-Scale (R1-R5) at Discharge

- R-Scale (R1-R5) at Discharge*
- R-Scale Total Score

* Denotes sections of the Chu (2010) dataset that required a research assistant to code discharge scores.
4.7 Procedure

4.7.1 Data Linkage

The steps taken to code and link the data in an ethical manner are as follows:

1. De-identified datasets were received from researchers Chu (2010) and Campbell (2007). Numerical identifiers and full names of the participants were received in a separate document.

2. Following the removal of duplicate cases, a sample list with matching identifiers was created.

3. The list of identifiers and names was provided to the Department of Justice, Victoria Police and the NCIS in a secure password-protected document.

4. The Department of Justice provided the PIMS data in the form of a password-protected Excel document. These documents were accompanied with a list of numerical identifiers, which were used to match the data to the original set of identifiers provided by the research team (i.e., two phases of identifier matching was required).

5. Victoria Police provided the recidivism data in the form of a password-protected Excel document. Once again, two phases of identifier matching was required.

6. NCIS data was provided to researchers in the form of a password-protected Excel document as per the Access Agreement between the research team and the NCIS.

7. Once all the data was coded into a de-identified SPSS dataset, the data was analysed using IBM’s Statistical Package for the Social Sciences (SPSS; Version 22).
4.7.2 Avoiding Bias

1. To ensure the researcher remained blind to the HCR-20\textsuperscript{V2} and original VRAG scores, this data was not accessed until after the HCR-20\textsuperscript{V3} and VRAG-R scores had been coded.

2. To ensure that the researcher remained blind to the outcome data, violence risk assessment tools were scored and entered into the SPSS database prior to accessing and entering recidivism data.

3. To ensure that the researcher remained blind to the reliability scoring outcomes, reliability scores were the last type of data to be entered into the dataset.

4.8 Definitions

4.8.1 Violence

Given that the research evaluated the predictive validity of violence risk assessment tools, it was imperative that the definition of violence adopted in this research corresponded with formal definitions described in the instrument manuals.

Violence was defined as “actual, attempted, or threatened infliction of bodily harm on another person” (Douglas et al., 2013; p. 36). This is the formal definition of violence as stated in the HCR-20\textsuperscript{V3} manual. Attempted actions are included in the definition of violence because success or failure of violent actions is the only differentiating factor between attempted and completed violent acts (Douglas et al., 2013). Threats may include overt expressions of intent to harm (e.g., “I am going to assault you”) or other communications and intimidating behaviours that invoke fear of harm.
According to the HCR-20\textsuperscript{V3} manual, bodily harm extends to both physical harm and serious psychological harm. Serious psychological harm includes a fear of physical injury, or other cognitive or emotional consequences. In support of the decision to include serious psychological harm as part of bodily harm, the authors of the HCR-20\textsuperscript{V3} have highlighted that this reflects the empirical reality of psychological harm potentially being more damaging to persons than physical harm (Douglas et al., 2013).

The willfulness of actions is fundamental to their classification as either violent or non-violent. The act must have been deliberate, not accidental or legally sanctioned (e.g., violence in military or law enforcement contexts; or violence that occurs with the victims consent such as sexual masochism). The requirement for deducing willfulness is simply a limited awareness that the act, attempt or threat may result in harm. Notably, legal parameters on culpability and criminal responsibility are not commensurate with determining deliberation. In acts of violence where the accused is found not guilty on account of mental illness, the act is still considered to be violent in nature. Therefore, in the case of forensic patients at TEH who were found Not Guilty by Reason of Mental Impairment (NGRMI), a violent index offence is evidence of a history of interpersonal violence.

Therefore, there are five key elements to the HCR-20\textsuperscript{V3} definition of violence: (a) The person engaged in an act (or omission); (b) with some degree of willfulness; (c) that caused or had the potential to cause; (d) physical or serious psychological harm; (e) to another person or persons (Douglas et al., 2013).

Exclusionary criteria: Within the HCR-20\textsuperscript{V3} definition of violence, property damage and harm to animals are not considered violent, unless they
are conducted in a manner that induces fear of harm in another person/s. Notably, there are some slight differences in the way violence is defined in the HCR-20\(^3\) and the VRAG-R. Similarly, the VRAG-R includes offences such as assault, sexual assault, armed robbery, forcible confinement, threatening with a weapon, and pointing a firearm as violent offences (Rice et al., 2013). Possession of a weapon is not classified as a violent offence, which is consistent with the current HCR-20\(^3\) definition of violence. The main point of difference however is the classification of robbery and arson as violent offences in the HCR-20\(^3\) and the current study, but not in the VRAG-R. Justifications for the classification of these offences as violent in nature is provided below and is generally founded on the physical and psychological harm to person(s) that these acts have the potential to achieve. This inconsistency in the definition of violence is only applicable to two participants (4.76%) of the VRAG-R sample whose first violence offence was either arson or robbery.

### 4.8.2 Violent recidivism

Violent recidivism was defined as any act that 1) would meet the aforementioned definition of violence; that 2) were committed during the follow-up period (post-discharge from TEH to 31/01/2013).

Other recorded violent recidivism characteristics included: 1) the type of first violent offence committed; 2) the frequency of violent offences in the follow-up period; and 3) the severity of violence across the follow-up period. Tallying the frequency of violent offences (i.e., total number of violent acts) across the follow-up period allows for calculation of the rate of violent recidivism. In conjunction with non-violent recidivism, a reoffending versatility
characteristic was also calculated. See Section 4.10 (violence severity) and 4.11 (violence versatility) for further information.

Weapons offences were not considered to be part of violent offending because weapons offences included possession of regulated/unregistered weapon or unsafe carrying of a weapon (i.e., ‘carry dangerous article’), which was not considered to meet the definition of violence as specified above. In the case where a weapon was used in the commission of a violent offence (e.g., ‘assault with weapon’), this offence was coded under the assault category.

Due to the serious psychological harm component of the violence definition, acts that may cause serious psychological harm were included in violent recidivism (e.g., stalking and forcible confinement or kidnapping). Arson and fire setting were also captured under the definition of violence as they may invoke both physical and psychological harm.

Burglary is defined in Section 76 of the Crimes Act (1958)(Vic) as: “A person is guilty of burglary if he enters any building or part of a building as a trespasser with intent (a) to steal anything in the building or part in question; or (b) to commit an offence (i) involving an assault to a person in the building or part in question; or (ii) involving any damage to the building or to property in the building or part in question”. Therefore, although burglary by legal definition potentially includes the intent to assault, there is no guaranteed actual harm or fear of harm. Therefore, burglary is not defined as a violent offence.

Aggravated burglary on the other hand is defined as: “A person is guilty of aggravated burglary if he or she commits a burglary and (a) at the time has with him or her any firearm or imitation firearm, any offensive
weapon or any explosive or imitation explosive; or (b) at the time of entering
the building or the part of the building a person was then present in the
building or part of the building and he or she knew that a person was then so
present or was reckless as to whether or not a person was then so present”
(Section 77, Crimes Act (1958)(Vic)).

Therefore, aggravated burglary potentially includes a fear-inducing
element in the ‘person present’ form. However, as outlined previously,
because carrying a weapon does not constitute a violent offence, a
differentiation between aggravated burglary where a weapon was present
(coded as non-violent offending), and aggravated burglary where a person
was present (coded as violent offending) was made. This was possible as
Victoria Police provided the ‘person present’ specifier.

Both robbery and armed robbery were classified as violent offences
due to their definitions in the Crimes Act (1958)(Vic). As per Section 75,
robbery was included because the definition states that an individual is guilty
of robbery if in the commission of the crime they use “force on any person or
puts or seeks to put any person in fear that he or another person will be then
and there subjected to force”. Armed robbery is therefore automatically
included because it also includes this element of fear.

Table 22 provides the categories and examples of offences that were coded
as violent recidivism.
Table 22

Violent Offence Categories with Examples

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homicide Offences</td>
<td>Murder, manslaughter, culpable driving causing death</td>
</tr>
<tr>
<td>Sexual Offences</td>
<td>Rape, exhibitionism, indecent act, Sexual penetration of a child under 16-years-old</td>
</tr>
<tr>
<td>Assault Offences</td>
<td>Attempted murder, assault, recklessly/intentionally cause injury, aggravated burglary, armed robbery</td>
</tr>
<tr>
<td>Theft Offences with</td>
<td>Robbery, armed robbery, aggravated burglary (person present)</td>
</tr>
<tr>
<td>Violence</td>
<td></td>
</tr>
<tr>
<td>Kidnap Offences</td>
<td>Kidnapping, false imprisonment, Abduction, hold against will, unlawfully detain</td>
</tr>
<tr>
<td>Arson Offences</td>
<td>Arson, fire-setting</td>
</tr>
<tr>
<td>Threat Offences</td>
<td>Use threatening words, threat to kill, extortion with threats</td>
</tr>
<tr>
<td>Stalking Offences</td>
<td>Stalk person/s, harassment</td>
</tr>
</tbody>
</table>

4.8.3 General Recidivism

Although the focus of the research was the predictive validity of the tools in relation to violence, non-violent offences were also considered. General recidivism analyses included both violent and non-violent offences.
No offence hierarchy was created for non-violent recidivism. Table 23 provides the categories and examples of offences that were coded as non-violent recidivism.

Table 23

*Non-violent Offence Categories with Examples*

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Property Offences</td>
<td>Criminal damage</td>
</tr>
<tr>
<td>Drug Offences</td>
<td>Use/possess/traffic/cultivate drugs, drink driving</td>
</tr>
<tr>
<td>Deception Offences</td>
<td>Obtain property by deception, handle /receive stolen goods, deal proceeds of crime, state false name, fraud</td>
</tr>
<tr>
<td>Obstruction of Justice</td>
<td>Resist arrest, hinder police</td>
</tr>
<tr>
<td>Theft Offences</td>
<td>Burglary (excl. person present), theft, go equipped</td>
</tr>
<tr>
<td>Legal Order Offences</td>
<td>Breach order, fail to answer bail, escape legal custody</td>
</tr>
<tr>
<td>Public Order Offences</td>
<td>Public nuisance, cause disturbance, offensive public behaviour, loiter, beg alms, fare evasion, prostitution</td>
</tr>
<tr>
<td>Driving Offences</td>
<td>Exceed speed limit, unlicensed driving, drive in reckless manner</td>
</tr>
<tr>
<td>Other Offences</td>
<td>Attempt to commit indictable offence, Incitement, conspiracy to commit crime</td>
</tr>
</tbody>
</table>
4.8.4 Recidivism Related data

In addition to violent and non-violent reoffending, Victoria Police also provided data on contacts with police (including person checks, vehicle checks and search warrants) during the follow-up period. The frequency of contacts was recorded for the purpose of descriptive data. IVO’s instated during the follow-up period against participants were also used for descriptive purposes.

4.9 Offence Hierarchy: Coding Violence Severity

An offence hierarchy was used to determine the severity of violent offending in the follow-up period (i.e., what was the most severe form of violence perpetrated). A hierarchy was created based on an adapted version of the CFBS Offence Hierarchy (see Ogloff et al., 2012; Ducat, 2013; and Kesic, 2011). Offences were coded as 1 = stalking; 2 = threat; 3 = arson; 4 = forcible confinement; 5 = theft offences with violence; 6 = assault; 7 = sexual; 8 = homicide. The offence hierarchy is graduated in reference to fear of harm, potential harm and definite harm. It starts with fear-inducing acts (stalking and threats); to fear inducing with an element of potential physical harm and physical restraint (arson, forcible confinement and theft offences with violence); to definite physical harm (assault, sexual offending and homicide). Table 22 above (Violent Offence Categories with Examples) demonstrates the hierarchy in order, with examples.

4.10 Coding Offending Versatility

Criminal versatility related to the range of types of offences committed post-discharge, throughout the follow-up period. Versatility was coded using
the 15 criminal offence categories as listed in the PCL-R coding scheme (Item 20: Criminal versatility) (Hare, 2003). Offences that were included in the current study but unrepresented in the PCL-R coding scheme (i.e., stalking) were added as additional offence categories for the purposes of this research. Notably, as with correct PCL-R scoring administration, it is not the number of offence categories satisfied that were coded, but rather the nominated coding scheme where: 0 = participant committed 3 or fewer offences; 1 = participant committed 4 or 5 types of offences; or 2 = participant committed 6 or more types of offences. See Appendix G, Offence Categories for Criminal Versatility.

4.11 Administration of Violence Risk Assessment Tools

Violence risk assessment tools were retrospectively coded at the point of discharge from TEH, for both previous and current versions. The student researcher was the only rater of the tools, with the exception of additional two raters for inter-rater reliability purposes (see Section 4.16). In-patient files were read in entirety. Typical sources of information included a bio-psycho-social assessment, court reports, intensive case reviews, memorandums, daily notes and discharge summaries (compiled by nurses, psychiatrists, psychologists, occupational therapists and social workers).

4.12 User Qualifications

Due to the SPJ nature of the HCR-20\textsuperscript{V3}, professional skill and judgment is required. Authors of the HCR-20\textsuperscript{V3} note that users of the tool should possess the following qualifications: 1) Knowledge of violence; 2) expertise in individual assessment; and 3) expertise in mental disorder. Prior
to the commencement of data collection, the student researcher completed HCR-20\textsuperscript{V3} and PCL-R training provided by Professor Kevin Douglas and Professor James Ogloff. Through education and practicum components of the Doctor of Psychology (Clinical and Forensic) degree, the student researcher was considered to meet all user qualifications.

No formal training for the VRAG-R was received due to: 1) No formal training workshops being available in Australia; and 2) the actuarial nature of the instrument. Despite this, Professor James Ogloff received training in VRAG administration and provided training and guidance as required.

Ongoing supervision and clinical guidance was provided by research supervisors, Professors Michael Daffern and James Ogloff, who have extensive experience in violence risk assessment.

4.13 HCR-20\textsuperscript{V3} Administration

The HCR-20\textsuperscript{V3} authors specify 7 steps in completing the HCR-20\textsuperscript{V3}.

1) Gather and document basic case information: Reviewed in-patient files and collected evidence relating to each of the violence risk factors.

2) Complete Presence ratings

3) Complete Relevance ratings

4) Integrate case information through formulation

5) Identify and describe most likely scenarios for future violence

6) Recommend strategies for managing violence risk

7) Complete SRR.
The following sections of the HCR-20\textsuperscript{V3} were scored: Presence ratings; Relevance ratings; and Summary Risk Ratings (SRR). A description of these ratings are provided in the following sub-section. Other elements of the HCR-20\textsuperscript{V3} such as ‘Development of primary scenarios of future violence’ and ‘Development of Case Management Plans’ were not completed. For the readers benefit however, a brief description is provided. Risk scenarios are the link between risk formulations and risk management plans, in which clinicians construct narratives about forms of violence the subject may, possibly, perpetrate. Risk scenarios are “not a prediction about what will happen; rather, it is a projection about what could happen” (Douglas et al., 2013, p. 57, emphasis in original). In the HCR-20\textsuperscript{V3}, risk scenarios are considered in terms of nature, severity, imminence, frequency/duration and likelihood. Risk scenarios are future-oriented, and encourage the clinician to draw on the formulation, enriching the assessment with real potential future acts of violence.

Risk management is the final stage in the assessment process, and demonstrates the ideally seamless transition from assessment to management. The Risk-Need-Responsivity (RNR) model provides the framework for risk management plan (for a detailed review, see Andrews, 2012), ensuring that interventions are targeted appropriately towards changeable factors, being delivered in a way that is responsive. In the HCR-20\textsuperscript{V3}, risk management plans are constructed in terms of monitoring, treatment, supervision and victim safety planning activities.
4.13.1 Presence Ratings

Presence ratings are coded on a 3-level nominal scheme. Risk factors are coded as either N (No; information indicates that the risk factor is absent); P (Possible/Partial; information indicates that the risk factor is possibly or partially present); or Y (Yes; information indicates that the factor is definitely present). Notably, Possible/Partial ratings are also made when information indicates that the risk factor is present, but this information is inconclusive or contradicts other sources of information (Douglas et al., 2013).

As the HCR-20\textsuperscript{V3} Presence ratings are scored on a nominal system, an ordinal scale was created for this study by transposing the ratings to numerical scores where No = 0; Partial/Possible = 1; and Yes = 2. Therefore, a maximum Total score of 40 was possible, with maximum scale scores for the H, C and R scales being 20, 10 and 10 respectively.

For three items in the HCR-20\textsuperscript{V3}, assessors are also asked to consider whether the Presence rating is ‘definite’ or ‘provisional’. This question pertains to items H6 (Major Mental Disorder); H7 (Personality Disorder) and C3 (Symptoms of Major Mental Disorder). This was completed as part of the presence ratings and coded as 0 = Provisional and 1 = Definite.

4.13.2 Relevance Ratings

Relevance ratings are also coded on a 3-level nominal scheme of Low, Moderate or High. A rating of Low is made when information indicates that the factor is of low relevance to the development of risk management strategies. A factor could be of low relevance because it is absent, because it is present but unlikely to be functional in the future perpetration of violence, or because it is present but unlikely to impair risk management strategies. A
Moderate rating is made when information indicates that the risk factor is possibly or partially relevant to causing violence and impairing risk management strategies, however this relationship is somewhat unclear or limited. Finally, a High rating is made when a risk factor’s role in causing violence and impairing risk management strategies is ‘clear and substantial’.

Ratings were transposed to a numerical scheme where 0 = Low; 1 = Moderate and 2 = High. Therefore, a maximum total relevance rating score of 40 was possible, with maximum scale scores for the H, C and R scales being 20, 10 and 10 respectively.

4.13.3 Summary Risk Ratings

The HCR-20\textsuperscript{V3} specifies three SRR’s: case prioritisation, serious physical harm and imminence of violence. Following scoring of the ratings and development of a formulation, a final risk judgment (SRR) of either Low, Moderate or High was made for the Future violence/Case prioritisation SRR. In coding, these ratings were transposed to values of 1, 2 or 3 respectively.

4.13.4 Risk Formulation

A description of formulation in general is provided in Section 2.3.6. Regarding the HCR-20\textsuperscript{V3} specifically, formulation begins when assessors rate the Relevance ratings. As the HCR-20\textsuperscript{V3} manual (Douglas et al., 2013) notes:

Formulation furthers this process by requiring evaluators to integrate separate risk factors into a conceptually meaningful framework that explains a person’s violence. Ideally, we need to tell a story about an individual that integrates the many pieces of information available to
us. It is necessary to derive an individual theory of risk, to help us make sense of risk, and therefore how best to intervene and manage such risk (pp. 53).

In the administration instructions, Douglas and colleagues (2013, see p. 54) have directed users of the HCR-20$^\text{V3}$ to Hart and Logan’s (2011) work on the essential features of a formulation: inferential (explains reasons of causes of behaviour), action-oriented (formulation informs risk reduction plans), theory-driven, individualised, narrative (qualitative, tells a story about the subject), diachronic, testable and ampliative (generates new information about the subject). While there are many theories of violence, the authors have suggested several theories and conceptual models including the General Personality and Cognitive Social Learning theory (see Andrews, Bonta, & Wormith, 2010); Good Lives Model (Ward, 2002); and Weerasekera’s (1996) “Four P” model. Other methods such as constructing hierarchies of risk factors, creating risk factor clusters or using a decision theory framework (see Section 2.3.4) may be beneficial. Such theoretically driven models assist in generating hypotheses about the causes and perpetration of violence (Douglas et al., 2013).

In the current study, although there is no formal development of a violence risk formulation, it is conceptually impractical to score the SRR and use the HCR-20$^\text{V3}$ as a SPJ tool without having a formulation of the participants risk for violence. A working formulation was created but not coded into the data.
4.13.5 Missing HCR-20\textsuperscript{V3} Scores and Omitting Participants

On the HCR-20\textsuperscript{V3} coding scheme, both Presence and Relevance ratings may be omitted. Items are only omitted if there is no information available concerning the risk factor. Risk factors are not omitted due to the evaluator’s uncertainty about the presence or relevance of the risk factor (in such cases, the item is coded as “P”) (Douglas et al., 2013). Due to the nominal rating system, omitted risk factors were coded as Omitted = 0.

Participants are omitted from the sample based on a threshold of missing data whereby no more than five items could be omitted (no more than two items from the Historical scale, and one from each of the Clinical and Risk Management scales) (Webster et al., 1997; Douglas, Eaves, & Hart, 1997). If this limit was exceeded, the assessment was considered invalid and was not included in analyses. This threshold was selected in keeping with the thresholds set in the HCR-20\textsuperscript{V2}, and consequently the Chu (2010) and Campbell (2007) studies. It was also made in adherence to the HCR-20\textsuperscript{V3} authors calling for evaluators to omit items sparingly (Douglas et al., 2014).

Given that these clinical files had clearly been used for the purpose of research with the Chu (2010) and Campbell (2007) studies, it was anticipated that the material contained in the files would be of sufficient quality for data extraction. This proved true to the researcher’s experience, as no participants required exclusion from the study due to excessive missing data.

4.13.5.1 Prorating HCR-20\textsuperscript{V3} Scores

When a violence risk assessment tool had scores omitted but the number of omitted scores was below the threshold for complete exclusion, the score was prorated. Total scores and scale scores were prorated for both
Presence and Relevance ratings. Scores were prorated according to the following formula, which was adopted for prorating scores in both the Chu (2010) and Campbell (2007) research projects.

\[ S_p = \frac{S_r \times N}{N - n} \]

*Sr* refers to the raw scale score or Total score; *N* refers to the total number of items in the scale (i.e., 10, 5 or 5 items for H, C and R scales respectively) or HCR-20 V3 (i.e., 20 items); *n* refers to the number of omitted items; and *Sp* refers to the prorated score produced. All scores were prorated using an Excel calculator based on the above formula.

### 4.13.6 HCR-20\(^{V3}\) Coding Timeframe

Given the retrospective nature of the study, coding ranges for the H, C and R scales of the HCR-20\(^{V3}\) were selected. All relevant historical information was considered for the H-Scale. The timeframe is essentially the participants ‘lifetime’.

The HCR-20\(^{V3}\) manual notes that evaluators should determine a specific timeframe for coding the C-Scale, and recommend an optimal timeframe of greater than one month but less than six months (Douglas et al., 2013). As such, the C-Scale was coded based on all clinical information for the two months prior to discharge. This decision was made on the basis that: 1) two months of clinical notes was considered to be comprehensive enough to enable the capturing of clinical information; and 2) in general, the average length of stay of acute patients in TEH population is approximately 74 to 84 days (VIFMH, 2012). As a judgment had to be made about the rating range
prior to data collection, this indication was used to support a decision of two months file review. Once data was collected, the median length of stay in the sample was in fact found to be 59 days. Scoring the scales at the point of discharge also allowed the rater to capture current functioning (e.g., mental state) and known forthcoming contextual considerations (e.g., discharge location, familial support etc.) directly prior to release/transfer. Such information would not be known had the scales been rated earlier in the participants hospital stay.

The HCR-20\textsuperscript{V3} manual also notes that evaluators should determine a specific timeframe for coding the R-Scale. It is advised that the optimal timeframe is somewhere in the “near future”, greater than one month but less than 12 months (Douglas et al., 2013; p. 48). A decision was made that the R-Scale would be coded based on risk management information pertaining to the six months following release. This decision was made on the basis that six months represents the half-way point in the recommended timeframe. Clinically, it was also considered to be a reasonable prospective rating period, given that a proportion of the sample was transferred to prison upon discharge, and therefore had delayed entry into the community.

4.14 VRAG-R Administration

The VRAG-R was completed in its entirety (i.e., where possible, all 12 items were scored). In completing the VRAG-R, additional resources such as the Cormier Lang System for Quantifying Criminal History (Quinsey et al., 2006) and Facet 4 of the PCL-R (Hare, 2003) are required.
4.14.1 The Cormier Lang System for Quantifying Criminal History

The Cormier-Lang System is used to determine the seriousness of a participant’s criminal history in both Item 5 (criminal history score for non-violent offences) and Item 8 (criminal history score for violent offence). An adaptation of Akman and Normandeau’s (1967) system, the Cormier-Lang system uses empirically established weights based on offence severity to produce criminal history scores for violent and non-violent offending. Although predicated on the Criminal Code of Canada, it is easily applied to criminal offences in the Australian jurisdiction. The system takes into account both charges and convictions. In the case where a charge and conviction relate to the same incident and sufficient information is known about the offence, the charge is scored (i.e., the highest category). In the case where the exact type of offence is unknown, an “at least” method is adopted (i.e., offence scored in lowest category) (Quinsey et al., 2006). See Appendix H for the Cormier-Lang scoring rubric.

In the current study, criminal histories were captured through official criminal records and additional documentation, such as the Corrections Victoria memorandum (which includes the number of incarceration periods and types of offences), forensic histories taken on admission to TEH and in extant psychological court reports.

4.14.2 Psychopathy Checklist - Revised

Item 12 of the VRAG-R (Anti-sociality) is represented by Facet 4 of the PCL-R, including PCL-R items 10 (Poor Behavioural Controls); item 12 (Early Behaviour Problems); item 18 (Juvenile Delinquency); item 19 (Revocation of Conditional Release) and item 20 (Criminal Versatility). As per standard PCL-
R scoring, each of the five items is rated as 0, 1, or 2 where 0 = item does not apply to the individual; 1 = item applies to a certain extent but not the degree required for a score of 2 and 2 = item applies to the individual. The scores across the 5 items are then summed to create a PCL-R Facet 4 score. The VRAG-R item is then scored accordingly. See Appendix I for the PCL-R Facet 4 scoring rubric. Notably, the PCL-R scores used for hypothesis 7 (incremental validity) were derived from the Chu (2010) and Campbell (2007) comparative datasets, which contained PCL-R Total scores. However, PCL-R scores used to inform VRAG-R item 12 (Anti-sociality) were completed by the student researcher.

4.14.3 Missing VRAG-R scores and Omitting Participants

According to the original VRAG manual, an offender’s score is still considered valid when no more than four items are missing. This threshold for missing data has not altered in the VRAG-R. No participants were omitted from the VRAG-R sample due to missing data. None of the participants in the Chu (2010) dataset had VRAG score omissions.

4.14.4 Prorating VRAG-R Scores

VRAG-R Total scores were prorated according to the official manual (Harris et al., 2015). There are two potential prorating phases in the VRAG-R: Within-item prorating and VRAG-R Total score prorating. All scores were prorated using an Excel calculator based on the procedure outlined below. Prorated scores were then used to determine VRAG-R Bin categorisation. Scores were prorated through the following process:
1) Add up the Total score from scored items (original Total score)

2) If the score is positive\(^5\), determine the highest possible positive score
   the participant could have obtained on those scored items (highest
   possible for scored)

3) Divide the original Total score by highest possible scored, which = \(X\)

4) Determine the highest possible number of positive scores the
   participant could have obtained on missing items (highest possible for
   missing)

5) Multiply highest possible for missing by \(X\), which = \(Y\)

6) Add the original Total score and \(Y\)

Within-item prorating relates to items 3 and 12. Item 3 (History of
Alcohol or Drug Problems) is scored by allotting one point for each of the
following six questions that can be answered ‘yes’: Alcohol problem before
age 18, Illicit drug problem before age 18, Alcohol involved in a prior offence,
Illegal drug involved in a prior offence, Alcohol involved in a current offence,
Illegal drug involved in a current offence. If there is no evidence to inform any
of these six questions, the item 3 must be prorated. The formula for prorating
is detailed below where the prorated score \(A^p\) is computed from \(N^1\) (number
of items scored ‘yes’) and \(N^T\) (total number of items scored):

\[
A^p = \frac{N^1}{N^2} \times 6
\]

\(^5\) If the score is negative, follow the same procedure as outlined above but
add a negative number to the previously obtained negative number.
Item 12 (Anti-sociality) is prorated as per the PCL-R manual. Only one of the PCL-R Facet 4 items can be missing, if more than one is missing, Item 12 is omitted. The PCL-R manual provides a table where the evaluator locates the Total score based on completed items.

4.15 Reliability Scoring

Two separate research assistants assisted with the reliability scoring for the HCR-20V3 and the VRAG-R respectively. For both the HCR-20V3 and VRAG-R, a minimum of 10% of the sample was selected for use in reliability scoring. Of the 100 participants who received a HCR-20V3 assessment, ten participants were selected for inter-rater reliability purposes. Of the 42 participants who had a VRAG-R score, five participants were selected for VRAG-R inter-rater reliability purposes. Research assistants were selected based on their qualifications and training in using violence risk assessment tools. All participants were randomly selected for inter-rater reliability through the Excel ‘randomisation’ function.

The research assistant scoring the HCR-20V3 is a clinical and forensic psychologist who had received formal training in the HCR-20V3. She was also employed at TEH as a psychologist during a specific period that fell within the data collection period (i.e., April 2000 – December 2010). The decision was made to retract any participants selected for reliability if the research assistant either knew the patient directly or knew of the patient. Therefore, for all participants selected and accepted for reliability scoring, it was ensured that the research assistant had no prior knowledge of these patients. The research assistant completed the same sections of the HCR-20V3 as the
student researcher (i.e., Presence rating; Relevance rating and SRR) and scored under the exact same date range conditions.

The research assistant scoring the VRAG-R is a clinical and forensic psychology doctoral student. No formal training was received for scoring the VRAG-R due to no formal training workshops being available in Australia. Once again, Professor James Ogloff provided training and consultation was available as required.

4.16 Calculating Time at Risk

‘Time at risk’ was defined as the number of days the participant spent residing in the community (i.e., not incarcerated or admitted into a mental health facility), where he/she had the opportunity to re-offend, prior to being charged for an offence. Victoria Police data provided dates of criminal charges.

In order to calculate time at risk, time spent incarcerated and/or rehospitalised post-release must be subtracted from time spent in the community post-release. Time spent incarcerated was provided by the Department of Justice in the form of the PIMS data. This is essentially ‘custody episodes’ and included prison and police custody. See Section 4.7.3 for further information.

Forensic rehospitalisations (i.e., readmissions to TEH following discharge) were captured within the PIMS incarceration periods because legally, participants entering TEH under legislation including s. 16(3)(b) of the Mental Health Act (2014) and s. 93 of the Sentencing Act (1991) are still considered ‘prisoners’. Logically then, this data is captured under the Department of Justice as opposed to the Victorian mental health system. This
required manual partitioning of incarceration periods and forensic rehospitalisation periods listed in the PIMS data, using the Client Management Interface (CMI) database (which details all admissions and separations within mental health services in Victoria).

Alternatively, forensic patients subject to a Custodial Supervision Order (CSO) under the Crimes (Mental Impairment and Unfitness to be Tried Act 1997) and involuntary patients under s. 12 of the Mental Health Act (2014) can be detained at TEH following expiration of their sentence. Movements for these patient groups are not captured within the PIMS data because their legal sanctions are not under the Department of Justice but rather, the Victorian mental health system. Civil rehospitalisations were also derived from the CMI system.

Time at risk is essentially then a combination of four post-release considerations: Time spent in the community; time spent incarcerated, time spent in forensic rehospitalisation (TEH); and time spent in psychiatric civil rehospitalisation. Time at Risk was calculated based on the following formula:

\[
TR = TD - DI - DR(f + nf)
\]

TR: Time at risk considering both forensic psychiatric and civil psychiatric rehospitalisations.

TD = Total number of days from discharge to end of the follow-up period.

DI = Total number of days spent incarcerated post-discharge from TEH to the end of the follow-up period.

DR(\(f + nf\)) = Total number of days spent in forensic psychiatric and non-forensic (i.e., civil psychiatric) rehospitalisation.
Time at risk variables were prepared for three different offence types: 1) Any offence in the follow-up period; 2) first violent offence in the follow-up period; and 3) first non-violent offence in the follow-up period. This is necessary because the date of first offence changes depending on the offence of interest.

Recidivism data were provided for the period 1\textsuperscript{st} of April 2000 – 31\textsuperscript{st} of January 2013. This enabled a maximum follow-up period of 12 years and 10 months. Based on the earliest discharge date within the sample (3\textsuperscript{rd} of August 2000) and conclusion of the follow-up period (31\textsuperscript{st} January, 2013), the maximum possible time at risk was 4,564 days. The average time at risk for any offending was 1442.93 days ($Md = 783$, $SD = 1514.83$, $Min = 1$, $Max = 4502$). The average time at risk for violent offending was 1935.19 days ($Md = 1397.50$, $SD = 1622.03$, $Min = 1$, $Max = 4502$).

4.17 Sample Demographics

4.17.1 Age, Gender, Ethnicity and Martial Status

The final sample comprised one hundred forensic psychiatric patients ($N = 73$ Males; $N = 27$ Females). The age at admission ranged from 18.34 years to 62.03 years ($M = 32.80$; $SD = 9.95$). The age at discharge ranged from 18.51 years to 63.18 years ($M = 33.51$; $SD = 10.20$).

Most participants were Australian born ($N = 78$), five of whom were Indigenous Australians. The remainder was born overseas including from Asia (7%); Europe (8%); New Zealand (5%) and the United Kingdom (2%).

The cultural composition of the sample was considered in terms of individuals who were from an English Speaking Background (ESB);
Aboriginal and Torres Strait Islander (ATSI); or Culturally and Linguistically Diverse (CALD). CALD has replaced the previously used term of ‘Non-English speaking background’ (NESB), and is more inclusive (Queensland Government, 2010). Participants were identified as being CALD if they were not Indigenous Australians and had a cultural heritage different to that of an Anglo-Australian culture (Queensland Government, 2010; Sawrikar & Katz, 2009). Based on this definition, the sample comprised 74% ESB; 5% ATSI and 21% CALD.

    The marital status of the participants on discharge were as follows: Never married or in a de-facto relationship (53%); presently married or in a de-facto relationship (11%); divorced or separated (30%) and widowed (6%). Notably, of those widowed, three participants were admitted to TEH for the index offence of murdering their partners.

4.17.2 Legal Status, Unit Location and Length of Stay

    Legal status on admission was as follows: Security patients (88%), involuntary patients (8%) and forensic patients (4%). Legal status is subject to change during the course of a participant’s admission. For example, a security patient whose sentence expires during the admission period but who still remains mentally disordered may be detained at TEH as an involuntary patient. Legal status on discharge was as follows: Security patients (72%), involuntary patients (20%) and forensic patients (8%).

    On admission, participants were placed on the following units: Argyle (36%); Atherton (35%); and Barossa (26%). Three participants were admitted to RPFC M6 (2%) and M5 (1%) units and transferred to TEH in April 2000.
The average length of inpatient stay in the sample was 257.99 days ($SD = 615.82; Min = 5; Max = 3037$). However, the median was 59 days.

Based on legal status at discharge, the average length of stay was 1297.25 days for forensic patients, 463.40 days for involuntary patients and 86.21 days for security patients. Although forensic patients represented only 8% of the sample on discharge, their average length of stay was much longer than other groups.

4.17.3 Education and Work History

Data on educational history was attained for most of the sample. No educational history was available for one participant and only partial educational history could be located for another participant. Two participants received no formal education. Across the entire sample, 97% completed primary school and 13% completed secondary school. Twenty-nine percent of the sample engaged in some level of formal vocational training (e.g., certificates or apprenticeships); and 6% had commenced tertiary education (not necessarily completed). The highest level of education attained was as follows: Primary school (64%); secondary school (5%); vocational training (23%) and tertiary education completed (5%). Twelve percent had never been employed; whereas 88% had some form of employment history (i.e., record of at least one employment episode).

4.17.4 Previous Hospitalisation

Most participants had previous admissions to forensic and/or non-forensic mental health units. Fifty-five percent of the sample had prior forensic psychiatric admissions (e.g., to TEH or psychiatric units within prisons such
as the Acute Assessment Unit at the Melbourne Assessment Prison). Of those with a previous forensic psychiatric admission, the average number of previous admissions was 3.60 (Min = 1; Max = 22; SD = 3.33). Seventy-seven percent of the sample had previous civil psychiatric admissions. Of those who had a previous psychiatric admission, the average age of first psychiatric admission was 24.62 years (Min = 11; Max = 53).

4.17.5 Suicide and Self-harm

Most of the sample had a history of self-harm (51%) and/or a history of serious suicide risk (60%) recorded in the in-patient file. Just under half of the sample (41%) had previously attempted suicide. Presence of a history of suicide and self-harm was only recorded if they were explicitly stated in in-patients files.

4.17.6 Substance Use

A participant was considered to have a history of substance use if it was recorded explicitly in the participant’s file. Seven percent of the sample had no recorded substance use across the lifespan. Thirteen percent of the sample used one substance only, whilst 80% were poly-substance users (i.e., used more than one substance). The percentage of the sample who had a history of use with the following substances were: Cannabis (88%); alcohol (78%); amphetamines (62%); heroin (56%); cocaine (12%) and other substances (e.g., benzodiazepines, magic mushrooms, Lysergic acid diethylamide (LSD), inhalants) (36%).

More recent substance abuse was also recorded in addition to historical abuse (however recent use is included in the above statistics).
Substance use in the 12 months preceding participant’s discharge date from TEH was also recorded. Percentage of substance use was as follows: Cannabis (65%); alcohol (59%); heroin (35%); amphetamines (35%) cocaine (5%) and others (e.g., benzodiazepines, inhalants). The abuse of prescription medications was also recorded for 18% of the sample.

4.17.7 Psychiatric Diagnoses (Admission)

Psychiatric diagnoses made prior to admission to TEH were coded into the dataset by the student researcher based on the current DSM-V (American Psychiatric Association, 2013). This data was sourced from in-patient files and the Department of Health’s Client Management Interface (CMI) system. Eighty-nine percent of the sample received at least one previous psychiatric diagnosis. Forty-three percent of the total sample had previously recorded co-morbid psychiatric diagnoses, 46% had only one previously recorded psychiatric diagnosis, and 11% had no previously recorded psychiatric diagnosis.

The distribution of previously received diagnoses on admission across the total sample was: Schizophrenia Spectrum and other Psychotic Disorders (68%); Depressive Disorders (29%); Bipolar and related Affective Disorders (13%); Substance Related Disorders (19%); Trauma and Stress-Related Disorders (6%); Anxiety (3%); Obsessive-Compulsive Disorder (OCD) (1%); Paraphillic disorders (1%); and other disorders (i.e., Dissociative, Somatisation and Eating Disorders) (4%).

On admission, 64% of the sample had a previously recorded personality disorder diagnosis. Of these 64 participants, 41% had a single previous personality disorder diagnosis, whereas 23% had two or more
personality disorder diagnoses. The distribution of personality disorders was as follows: Antisocial (62.50%); Borderline (21.88%); Paranoid (6.25%); Obsessive-compulsive (4.69%); Schizoid (6.25%); Histrionic (7.81%); Dependent (12.50%); Narcissistic (7.81%); Avoidant (3.13%); Schizotypal (1%) and Personality Disorder Not Otherwise Specified (NOS) (10.94%).

A cross tabulation of history of personality disorder and history of major mental disorder revealed that the majority of participants who had a major mental disorder also had a personality disorder (60 participants). Twenty-nine participants had only a major mental disorder recorded, and four participants had only a personality disorder recorded. Seven participants had no previous major mental illness diagnosis or personality disorder diagnosis. Other previously received diagnoses were: Intellectual disability (3%); acquired brain injury (2%); alcohol related cognitive defects (1%); eating disorder (2%); Huntington’s disease (1%).

4.17.8 Psychiatric Diagnoses (at Discharge)

Upon discharge from TEH, formal diagnoses are recorded by the consultant psychiatrist in the discharge summary. These are reported in terms of major mental illnesses, personality disorders and other relevant diagnoses (e.g., chronic illness and medical diagnoses). Diagnoses may be based on interviews, observation and structured assessments made by the multi-disciplinary team. Ninety-five percent of the sample was diagnosed with mental illness upon discharge from TEH (excluding personality disorder diagnoses). Of the total sample, 64% received a single diagnosis, whereas 27% received co-morbid psychiatric diagnoses. Across primary and secondary diagnoses, the distribution was as follows: Schizophrenia
Spectrum and other Psychotic Disorders (72%); Depressive Disorders (8%); Bipolar and Related Disorders (6%); Substance Related Disorders (17%); Trauma and Stress-related Disorders (5%); OCD (1%); and Eating Disorder (1%).

Thirty percent of the total sample was diagnosed with a personality disorder upon discharge. Of these 30 participants, 22 were diagnosed with a single personality, whilst seven were diagnosed with co-morbid personality disorders. One participant was diagnosed with three personality disorders upon discharge. The prevalence of personality disorder diagnoses across the entire sample was as follows: Anti-social Personality Disorder (13%); Borderline Personality Disorder (11%); Paranoid Personality Disorder (4%); Narcissistic Personality Disorder (2%); and Personality Disorder Not Otherwise Specified (NOS) (4%). Obsessive-Compulsive, Schizotypal, Dependent and Avoidant personality disorders were each represented in 1% of the total sample.

Other relevant diagnoses recorded on discharge were: Intellectual Disability or cognitive impairment (6%); Ganser Syndrome (1%); Parkinsons (1%); Aspergers (1%); Huntingtons (1%); and other medical diagnoses (3%).

Based on a historical medical record and medical investigations undertaken during the index admission, 54% of the sample was considered to have a chronic illness, including: Epilepsy (7%); Hepatitis A, B or C (41%); Asthma (5%); Diabetes (5%); Morbid obesity (8%); Cancer (1%) and others (e.g., High Blood Pressure, Chronic Pain, Cerebral Palsy, and ear or liver complications) (15%).
4.17.9 Institutional Behaviour and Use of Seclusion

Inpatient aggressive behaviour was recorded in terms of physical aggression, verbal aggression and property damage. This data was derived from official incident report forms, regular in-patient risk assessments, HoNOS (Health of the Nation Outcome Scales) charts, and daily nursing notes. 36.36% of the sample had at least one instance of physical aggression and 62.63% had at least one instance of verbal aggression. Evidence of property damage was found for 31.31% of the sample.

The use of seclusion rooms was also recorded. Under s. 3 of the Mental Health Act 2014 (VIC), seclusion is defined as the ‘sole confinement of a person to a room or any other enclosed space from which it is not within the control of the person confined to leave’. The Mental Health Act (2014, VIC) defines seclusion as a restrictive intervention which may be used when “A person receiving mental health services in a designated mental health service may be kept in seclusion if seclusion is necessary to prevent imminent and serious harm to the person or to another person” (p.102). At TEH, purpose built seclusion rooms are utilised for such purposes.

Fifty participants were secluded on at least one occasion during their hospital stay. However, it should be noted that seclusion is not always used in the case of inpatient aggression (e.g., it can also be used as a temporary ‘transition’ space between prison and the mainstream TEH unit, or as a means of separating patients, or for the purposes of monitoring in risk to self).

4.17.10 Forensic History

Sixty-six percent of the sample had previous violent offences (not including index offence/s). Namely, murder (2%), sexual offences (non-
penetration) (8%), sexual offences (penetration) (5%), assault (59%), robbery (13%), armed robbery (16%), aggravated burglary (4%), threats (25%), arson (4%), kidnap and confinement offences (4%) and stalking (2%).

Eighty-one percent of the sample had a previous non-violent offence. Types of non-violent offending included: Theft offences (60%); drug offences (42%); property offences (38%); burglary (30%); weapons offences (26%); obstruction of justice (13%); breach of order (48%) and other miscellaneous offences (e.g., driving offences) (50%).

4.17.11 Index Offence(s)

Index offence(s) were defined as offence(s) that directly contributed to the participant’s admission to TEH or prison prior to TEH transfer. Twenty participants’ index offences were exclusively violent; whereas thirty-three participant’s index offences were exclusively non-violent. For 47 participants, the index offending included both violent and non-violent offending.

For 67% of the sample, the index offence(s) included at least one violent offence. The distribution of violent index offences was as follows: Assault (39%); threats (15%); robbery (11%); armed robbery (11%); murder (10%); sexual offences (5%); stalking (2%); kidnap and forcible confinement (2%); aggravated burglary (person present) (2%); and arson (1%).

The severity of violence in the index offence(s) (i.e., the most severe type of violent offence across all index offences) was also calculated based on the CFBS offence hierarchy. For participants who had violent offending as part of their index offences, assault was the most severe violent offence for 55.2% of the sample, followed by theft with violence (16.4%), homicide
(14.9%), sexual offences (7.5%), threat to persons (3%), stalking (1.5%) and arson (1.5%).

Eighty percent of the sample had non-violent offending in the index offences. Non-violent index offences were distributed as follows: Theft offences (43%); breach order (31%) property offences (26%); weapons offences (23%); drug offences (18%); burglary (13%); obstruction of justice (12%); and others (e.g., driving offences) (42%).

4.17.12 Offending in the Follow-up Period

Sixty-three percent of the sample committed an offence post-release. In offences post-release, 50% of the sample was charged with at least one violent offence; whereas 57% of the sample were charged with at least one non-violent offence. Six percent committed only violent offences post-release; 13% committed only non-violent offences post-release; and 44% committed both violent and non-violent offending in the follow-up period. Of the 50 participants who committed violent offences, the most common type of first violent offence was assault (35%), followed by theft with violence (8%), sexual violence (4%), arson (2%) and threats (1%). The average number of violent offences committed across the follow-up period was 2.69 (Min = 1; Max = 23; SD = 4.72).

The most severe violent offence committed during the entire follow-up period (based on the CFBS violent offence hierarchy) was as follows: threat (1%); arson (1%); theft with violence (3%); assault (34%); sexual violence (10%) and homicide (1%).

Of the 57 participants who committed non-violent offences, the most common type of first non-violent offences were theft offences (15%); legal
order offences (8%); public order offences (7%); drug offences (6%); property
damage offences (6%); aggravated burglary (weapon present only) (5%);
weapons offences (4%); obstruction of justice (3%) and deception offences
(3%). The average number of non-violent offences committed across the
follow-up period was 9.67 (Min = 1; Max = 139; SD = 20.64).

Considering the 63 participants who recidivated, the criminal versatility
displayed in the follow up period (as per the PCL-R Item 20) was as follows: 3
or fewer offences (26%); 4 or 5 types of offences (14%) and 6 or more types
of offences (24%).

Information regarding other recidivism related data (i.e., IVO’s, contacts with police and placement on orders post release) was also
collected. Thirty-three participants had IVO’s lodged against them during the
follow-up period. Participants were always defendants of IVO’s, never
applicants. Sixty-two percent of the sample had contact/s with police during
the follow-up period. Sixty-seven participants were placed on legal orders
post-release (including parole orders, CBO’s, ICO’s and others).

4.17.13 Discharge Pathways and Legal Orders

Most of the sample (89%) was still considered symptomatic (i.e.,
exhibiting full or partial symptoms of major mental illness as per the HCR-
20\textsuperscript{V3} C3 item) upon discharge. Ninety-five percent of the sample was
discharged from TEH whilst still being prescribed psychiatric medications.

Participants were discharged form TEH from the following units: Argyle
(30%); Atherton (28%); Barossa (25%); Bass (7%); Canning (4%); Daintree
(4%) and Jardine (2%). Upon discharge, participants were either released
directly into the community (63%) or transferred back to prison (37%).
For those participants released to the community, relevant orders may be in place. Forensic patients on a CSO under the *Crimes (Mental Impairment and Unfitness to be Tried Act 1997)* (VIC) are generally granted ‘extended leave’ into the community upon discharge from TEH. Legally, this takes effect in the form of a Non-Custodial Supervision Order (NCSO). Eleven percent of the sample were granted extended leave from TEH. Nine percent were discharged onto parole. Eleven percent were discharged onto Community Based Orders (CBO’s) and 28% were discharged onto Community Treatment Order (CTO’s). Sixteen participants had an active IVO against them at the point of discharge.

### 4.18 Representativeness of the Sample

In order to establish whether the sample was representative of the TEH population, the sample was compared with the TEH 2014-2015 population on 4 factors: Proportion of male and female patients; legal status dispositions; average length of stay and psychiatric diagnoses. Unfortunately, information on the average age of patients to TEH at admission and discharge could not be sourced. The TEH 2014-2015 population was selected as the most recent population snapshot available, to inform applicability of findings to recent hospital operation.

Between 2014 and 2015, there were 76 male admissions and 33 female admissions to TEH (VIFMH, 2015). Therefore, on any given day during the financial year, approximately 70% of the TEH patient population was male. This is similar to the gender divide in the current sample, in which 73% of the sample was male.
During the period when participants were hospitalised in TEH, July 2002 – June 2003, there were 142 admissions, of which 82% were security patients, 12% involuntary patients and 6% forensic patients (VIFMH, 2003). This is similar to the distribution within the current sample, where security patients comprised 88% of the sample, involuntary patients 8% and forensic patients 4%.

The average length of stay in the sample differed considerably from the average in the general TEH population. Based on legal status on discharge, the average length of stay in the TEH patient population between 2013 – 2014 was 86 days for security patients; 2,112 days for forensic patients; and 191 days for involuntary/other patient types (VIFMH, 2014). This differs considerably from the current sample, where the average length of stay was approximately 248 days for security patients; 291 days for forensic patients and 356 days for involuntary patients. This discrepancy may be a reflection of many factors, such as severity of mental illness in the sample, or changes in patient care, treatment and intervention since the establishment of TEH in 2000. This will be discussed further in Chapter 7, Discussion.

Regarding the use of seclusion, 50 participants had at least one episode of seclusion during their admission. During the period of 2013 – 2014 at TEH, use of seclusion varied by unit where the number of patients secluded was as follows: Argyle N = 25; Atherton N = 27; Barossa N = 12 and Canning N = 5 (VIFMH, 2015). Within the current sample, based on unit of admission, the number of patients secluded was similar, where: Argyle N = 20; Atherton N = 15; Barossa N = 14 and M6 = 1. Notably however, these outcomes represent seclusion episodes over the entire admission period, as opposed to a discrete one-year period. Therefore, the number of patients
secluded on male acute units being slightly less than the 2013 – 2014 statistics is not surprising given that the average length of stay at TEH in the sample was 257.99 days (i.e., less than a year).

The most recently published annual report providing a snapshot of the TEH patient population regarding psychiatric diagnoses was in June 2009. At this stage, the primary diagnoses were: Schizophrenia and related disorders (93%); affective disorders (depressive disorders and bipolar related disorders) (2.5%); adjustment disorders (trauma and stress-related disorders) (2.5%); personality disorder (1%) and others (1%) (VIFMH, 2009). In the current sample, the distribution was as follows: Schizophrenia and related disorders (72%); affective disorders (depressive disorders and bipolar related disorders) (14%); adjustment disorders (trauma and stress-related disorders) 5%; substance related disorders (17%); and others (2%). In comparison, the lower incidence of schizophrenias and higher incidence of affective disorders may be due to differences in classification of affective disorders with psychotic features. In the current sample, this presentation was coded according to the DSM-V system under affective disorders.

4.19 Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE)

Reporting of the research was conducted in accordance with the Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE; Singh et al., 2014). The RAGEE is a recently developed 50-item checklist that provides reporting guidelines for studies evaluating the predictive validity of violence risk assessment tools (Singh et al., 2014). See Appendix J for items of the RAGEE checklist that the current research has fulfilled.
Chapter Five

Statistical Methods

5.1 Introduction

The purpose of the current chapter is to describe the data preparation and statistical methods employed. Data cleaning and assumption testing processes are addressed, followed by a description of the statistical analyses. A detailed explanation of the Receiver Operating Characteristic (ROC) and the Area Under Curve (AUC) is provided due to predictive validity being a key focus of the study. All data analysis was conducted using IBM’s Statistical Package for the Social Sciences (SPSS; Version 22). Notably, a Swinburne University statistician provided consultation as required.

5.2 Hypotheses and Statistical Analyses

The eight hypotheses and statistical analyses employed to investigate them are listed below. It was hypothesised that:

1. Association: There will be a significant and positive association between violent recidivism and scores on the HRC-20\(^{V3}\) and VRAG-R. The HCR-20\(^{V3}\)’s Total scores and SRR’s will correlate significantly with violent recidivism. The VRAG-R’s Total scores and Bins will correlate significantly with violent recidivism. The association between frequency of violent offence charges post-release and scores on the HCR-20\(^{V3}\) and VRAG-R was also investigated. Kendall’s Tau was used to evaluate these hypotheses. A Chi-squared Test of Independence was also employed to test associations amongst nominal variables.
2. **Concurrent Validity:** The HCR-20\textsuperscript{V3} and VRAG-R will demonstrate acceptable concurrent validity (i.e., significant correlations of $r > .50$, as per Munro, 2001). Previous and current versions of the tools will correlate significantly (i.e., the HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} will correlate significantly, as will the VRAG and VRAG-R). Furthermore, the HCR-20\textsuperscript{V3} and VRAG-R would also correlate significantly, as will static measures (i.e., the HCR-20\textsuperscript{V3} H-Scale and VRAG-R). Finally, the HCR-20\textsuperscript{V2} scales will correlate significantly with the HCR-20\textsuperscript{V3} scales. Kendall’s Tau and Pearson’s correlation were used to test this hypothesis.

3. **Prediction:** The ability of the HCR-20\textsuperscript{V3} and the VRAG-R to predict future violence will be greater than chance. The observed AUC’s will be similar AUC’s produced by the HCR-20\textsuperscript{V2} and VRAG, and non-significantly different. The ROC was used to test this hypothesis.

4. **HCR-20\textsuperscript{V3} Predictive strength of scales:** The H-Scale will be the strongest predictor of post-release violence over the entire follow-up period. However, when controlling for time at risk, the C-Scale and R-Scale will emerge as the strongest predictors of post-release violence over the entire follow-up period. Binomial logistic regression was used to test this hypothesis.

5. **Discrimination:** Participants scoring above the median score on the HCR-20\textsuperscript{V3} and VRAG-R will be significantly more likely than those scoring below the median to 1) recidivate generally; and 2) recidivate
violently, during the entire follow-up period. A Chi-squared Test of Independence was used to test this hypothesis.

6. *Survival Time:* There would be statistically significant differences in survival time for general and violent recidivism between HCR-20\textsuperscript{V3} and VRAG-R scores (participants scoring above or below the median Total score) and risk categories (HCR-20\textsuperscript{V3} SRR’s and grouped VRAG-R Bins). Kaplan-Meier Survival Analysis were used to test this hypothesis.

7. *Incremental Validity:* Incremental validity was assessed within the HCR-20\textsuperscript{V3}, and between tools in predicting violent recidivism. Within the HCR-20\textsuperscript{V3}, Relevance ratings will demonstrate incremental predictive validity over the Presence ratings. Between tools, it was hypothesised that the PCL-R would not add incremental predictive validity to the HCR-20\textsuperscript{V3} (Total scores and SRR’s) nor the VRAG-R Total scores. It was also hypothesised that the HCR-20\textsuperscript{V3} would add incremental validity to the VRAG-R, but the VRAG-R would not add incremental validity to the HCR-20\textsuperscript{V3}. Sequential (hierarchical) logistic regression was used to test this hypothesis.

8. *Inter-rater Reliability:* The HCR-20\textsuperscript{V3} and VRAG-R will demonstrate acceptable inter-rater reliability (ICC > .40 as per Fleiss, 1981). The Intraclass Correlation Coefficient (ICC) was used to test this hypothesis.
5.3 Data Cleaning

The data cleaning process has already been described in Chapter 4, Section 4.7.5, and is summarised briefly here. Establishment of the final dataset required the merging of the two datasets from Campbell (2007) and Chu (2010). The sample was then refined based on exclusionary criteria, resulting in 124 remaining participants. Data collection ceased at 100 participants due to time constraints. The dataset was made complete through the addition of the demographic data, HCR-20\textsuperscript{V3} and VRAG-R scores, and HCR-20\textsuperscript{V2} updated discharge scores for C-Scale and R-Scale (see Chapter 4, Section 4.7.6).

5.4 Missing Data

Very little descriptive data (e.g., demographics, psychiatric data, inpatient behaviour, forensic history etc.) was missing. Omitted scores on the HCR-20\textsuperscript{V3} and VRAG-R were coded as missing. In the case of the HCR-20\textsuperscript{V3}, ‘Omit’ is an actual scoring category, which can be selected on the basis that there is not enough information to score the item. Therefore, these ratings are conceptually understood to be ‘missing data’ for the purposes of data analysis.

A missing value analysis using Little’s Missing Completely At Random (MCAR; Little, 1988) test was conducted for all Presence and Relevance ratings on the HCR-20\textsuperscript{V3}, as well as individual VRAG-R items. For the HCR-20\textsuperscript{V3}, all items with missing data on the H-Scale and C-Scale were found to be missing at random. There were no missing items recorded on the R-Scale. All items on the VRAG-R had data missing at random.
Regarding scores on the violence risk assessment tools, participants with omitted scores received a prorated score. Methods for prorating scores are outlined in Chapter Four, Sections 4.13.5.1 (HCR-20) and 4.14.4 (VRAG). As prorating of scores already accounts for omitted items (up to a certain threshold of missing items), ‘missing data’ was not computed.

5.5 Statistical Assumptions

The data was initially screened through descriptive statistics including measures of central tendency (mean, mode and median) and measures of variability (range, variance and standard deviation). Assumption checks were made for normality, homoscedasticity and independence for most analyses. Analyses requiring assumption checks are outlined in the relevant analysis below. Due to the ordinal nature of variables relating to violence risk assessment scores, normality checks were interpreted cautiously given metric considerations for non-continuous data.

Normality was assessed through visual methods (histograms and probability-plots (P-P Plots)) and statistical methods (kurtosis, skew and the Kolmogorov-Smirnov test (K-S test; Chakravarti, Laha, & Roy, 1967)). The assumption of homoscedasticity was examined using Levene’s test, parametric or non-parametric depending on the normality of the data (Levene, 1960). Outliers were identified through Box-plots.

Most variables violated assumptions of normality, therefore non-parametric statistical methods were employed (e.g., Kendall’s Tau correlation instead of Pearson’s correlation). Transformations were not computed due to the difficulty this lends to interpretation (i.e., consequential changes to hypothesis testing) (Grayson, 2004). Instead, robust statistical analyses were
used where possible to enable the accuracy of the statistical model despite assumption failures (Field, 2009).

In some analyses, bootstrapping was used to foster more robust models. Bootstrapping (Efron & Tibshirani, 1993) works by estimating properties of the sampling distributions through treating the sample as a population, from which smaller ‘bootstrap samples’ can be drawn (Field, 2009). Since bootstrapping is included as an automated option in some statistical software programs like SPSS, bootstrapping has gained prominence in psychological research (Fan, 2003).

5.6 Statistical Analyses

Unless otherwise stated, conventional statistical parameters of a \( p < .05 \) significance levels and 95% Confidence Intervals (95% CI’s) were adopted across all the statistical analyses.

5.6.1 Predictive Validity

Validity generally refers to whether an instrument measures what it purports to measure. Predictive validity is a validity index indicating the degree to which the instrument predicts a criterion measure, which in the case of the current study is violence (Cohen et al., 2010). To assess the predictive validity of an instrument, test scores can be obtained at one point in time and criterion measures obtained at a future point in time. Alternatively, retrospective designs and pseudo-prospective designs (such as in the current research method) allow for criterion measures to exist prior to test score calculation, provided that researchers remain blind to these outcomes.
Judgments of predictive validity are sometimes based on statistical evidence in the form of indexes derived from odd's ratios and 2 x 2 contingency tables (e.g., false positive and false negative cross-tabulations) and correlations between test scores and outcomes (Cohen et al., 2010; Rice et al., 1995). However, such techniques are limited by prediction errors and not including base rates of violence (i.e., frequency of violence in a given population). Mossman (1994; p. 783) called for researchers to “describe accuracy accurately”, and noted that predictive accuracy was oftentimes quantified as a function of correct predictions, without reference to the effects of base rates and clinician biases for or against Type I or Type II prediction errors. Correlational techniques are limited in populations where the base rate for the target behavior is low because of the diminishing size of the correlation as the criterion variable departs from base rates of 50% (Douglas et al., 1999; Mullen et al., 2009). Rice and Harris (1995) note that in risk assessment research, the base rate of violence is typically lower than 50%. Therefore, correlational techniques are ineffective for measuring the predictive validity of violence risk assessment tools (Rice et al., 1995). Mossman (1994) recommended the use of the Receiver Operating Characteristic (ROC) to overcome correlational limitations by establishing accuracy of predictions independent of base rates of violence and prediction errors (Mullen et al., 2009; Rice et al., 1995).

5.6.1.1 The Receiver Operating Characteristic (ROC)

Originally developed for use with radar and signal detection systems, the ROC has been adopted in many other fields such as radiology, where its unique properties are required for predictive purposes and diagnostic
certainty (Hanley et al., 1982). The ROC has now become the analysis of choice in violence risk assessment research (Hanley et al., 1982; Davis, 2010). Within the violence risk assessment context, the ROC is used to represent the probability that, within a random sample, a truly violent person would receive a higher score on a risk assessment instrument than a truly non-violent person (Nicholls et al., 2004). This is applicable to research where data consists of dependent dichotomous outcomes (i.e., violent or non-violent) (Douglas et al., 1999). Such is the case in violent risk assessment.

The ROC-AUC parameter was selected in the current study because it can establish the predictive accuracy of a risk assessment tool, compare accuracy between instruments, and indicate trade-offs between specificity and sensitivity at different decision thresholds (i.e., chosen cut-off scores/points) (Dolan et al., 2000; Douglas et al., 1999).

5.6.1.2 The ROC curve

The graph upon which a ROC curve will be placed has true positives along the y-axis, and false positives along the x-axis (Mossman & Somoza, 1991). True positives refer to the probability that a person who was predicted to be violent was actually violent. This is also known as sensitivity, which is the accuracy with which the outcome is predicted (i.e., proportion of violent people who were correctly predicted to be violent) (Mullen et al., 2009). False positives refer to the probability that a person who was predicted to be violent was actually non-violent. This is known as specificity (i.e., proportion of non-violent people who were correctly predicted to be non-violent). Therefore, sensitivity corresponds to individuals predicted to be violent, whilst specificity corresponds to individuals predicted to be non-violent.
The false positive to true positive ratio (FP:TP) indicates how many incorrect predictions occurred for every correct prediction. In turn, true negatives refer to the probability that a person who was predicted to be non-violent was actually non-violent; and false negatives refer to the probability that a person who was predicted to be non-violent was actually violent.

Thus, the ROC graph is essentially an illustration of the relationship between sensitivity and specificity for diagnostic tests. By plotting the true positive rate against the false positive rate, the ROC provides the ‘hit rate’ along various thresholds of the predictor variable. According to Metz (1978), the ROC curve “describes the inherent detection characteristics of the test... (and) the receiver of the test information can operate at any point on the curve by using an appropriate decision threshold” (p. 288, emphasis in original). Therefore, at any given level of specificity, the receiver (i.e., researcher or clinician) will know the sensitivity (Douglas et al., 1999). In summary, ROC curves take the form of a graph where the true positive rate (TPR; sensitivity of the predictor) is plotted as a function of the false positive rate (FPR; 1-specificity) (Mossman et al., 1991).

As a result, there is a trade-off between sensitivity and specificity at varying cut-off levels whereby, as sensitivity increases, specificity decreases, and vice versa. Ideally, accuracy should be described in a way that reflects this trade-off independent of the receiver’s cut-off or decision threshold. Therefore, ROC curves can be used to determine the optimal cut-off scores that maximize sensitivity and specificity. More relevant to this thesis however is the AUC, which provides the overall discriminatory ability of the violence risk assessment tool.
5.6.1.3 The Area Under Curve (AUC)

As the name suggests, the AUC is the portion of the graph that falls under the ROC curve. The AUC is a statistical index for interpreting the overall accuracy of a predictor, representing the proportion of true positives over false positives (Mullen et al., 2009; Douglas et al., 1999). The AUC essentially represents the probability that a randomly selected true positive and true negative are correctly ordered by the test (Campbell, 2007). In other words, the AUC value indicates the probability that a randomly chosen person who scores positive on the dependent measure (is subsequently violent), will score higher on the predictive measure than a randomly chosen nonviolent person (Mossman et al., 1991).

The diagonal line indicates chance prediction, where true positives are equal to false positives (i.e., for every patient accurately predicted to be violent, another patient was inaccurately predicted to be violent). This diagonal line is referred to as the ‘line of no information’ where the test has no predictive value. A curve above this line indicates that prediction is occurring at rates above chance. Therefore, as the curve deviates from the diagonal line, prediction improves. In violence risk assessment, the ROC curve will generally fall somewhere between the ‘ideal test’ point and the line of no information. The AUC value indicates the location of the ROC curve, relative to the line of no information (Hanley et al., 1982).

The AUC ranges from 0 – 1, where 0 = perfect negative prediction; 0.5 = chance prediction; and 1 = perfect positive prediction (Douglas et al., 1999). The AUC is interpreted such that an AUC of .80 would indicate that 80% of those predicted to be violent would be violent. The ROC curve approximates perfect accuracy as it moves towards the left and top boundaries of the
graph, as \( 1 = \) prefect predictive accuracy (Hanley et al., 1982). Therefore instruments that distinguish violent from non-violent individuals with nearly perfect accuracy will have an AUC approaching 1 (Dolan et al., 2000).

There is no formal categorisation for the interpretation of AUCs, and there is much variation within the literature. AUCs in the range of 0.75 to .80 have been considered to be moderate to large effect sizes (Mossman, 1994; Rice, 1997; Dolan et al., 2000). Others have noted that an AUC of 0.71 is equivalent to a Cohens \( d \) of .80 (large effect size), AUC of 0.64 is equivalent to a Cohens \( d \) of .50 (moderate effect size), while AUC’s of 0.56 is equivalent to a Cohens \( d \) of .20 (Cohen, 1988; Douglas et al., 1999; Rice & Harris, 2005). Again, others have considered AUC’s greater than 0.75, or a Cohens \( d \) greater than 0.50 to be a large effect size (Dolan et al., 2000). Elsewhere, AUC’s around 0.54 are regarded as small effects; 0.63 are moderate effects and those greater than 0.71 are considered large effects (Rice et al., 2005). For the current research, the following AUC thresholds were adopted based on an overview of AUC categories in the literature: \(< 0.65 = \) Small effect; \( 0.65 – 0.70 = \) Moderate effect and \( \geq 0.70 = \) Large effects. AUC’s were reported with 95% Confidence Intervals. Figure 2 illustrates a sample ROC curve and the AUC it produces, with notable elements of the graph described.
5.6.1.4 Other Relevant Discriminatory Indices

The AUC is beneficial as a global characterisation of the discrimination capabilities of diagnostic tools, and succinctly conveys performance information (Mossman, 2013). Although the AUC is oftentimes used to summarise the accuracy of tools in a single number, in isolation it cannot capture more detailed meaningful information about a tools performance, such as the trade-off between sensitivity and specificity, and false-positive and false-negative errors (i.e., PPV and NPV) (Martinez-Camblor, Carleos, & Corral, 2013; Mossman, 2013; Shea, 2015). Indeed, one test may be superior to another in terms of identifying positive and negative individuals, despite having the same AUC (Campbell, 1994). The adequacy of diagnostic tools
requires consideration of cut-offs and the benefits and consequences of errors. Therefore, it is crucial to report other relevant indices of discrimination to truly gauge the performance of tools.

Other relevant discriminatory indices include Positive Predictive Power (PPP) and Negative Predictive Power (NPP). PPP represents the accuracy of predictions that stated that individuals would be violent, whilst NPP represents the accuracy of predictions that stated that individuals would not be violent. Therefore, PPV represents the probability that participants with a positive test truly have the outcome, whereas NPV represents the probability that participants with a negative test truly don't have the outcome. In order to calculate PPV and NPV, the tools must be treated as though they produced a dichotomous outcome (i.e., a positive or negative test result). Clearly, neither tools produce this: the HCR-20\textsuperscript{V3} produces Low, Moderate or High SRR’s; and the VRAG-R produces nine Bin categories. Therefore dichotomous outcomes were created by grouping SRR’s as follows: Low vs. (Moderate and High), and (Low and Moderate) vs. High. Dichotomous outcomes were also created ascending through the various Bins of the VRAG-R as cut offs: 1 vs. (2 – 9), (1 – 2) vs. (3 - 9). A cut-off essentially represents the minimum value that will constitute a test-positive result. That being said, there is no utility in reporting indices for the minimum possible score (i.e., Low category on the HCR-20\textsuperscript{V3} and Bin number 1 on the VRAG-R), because this would essentially classify all cases as test positive (i.e., PPV would be equal to the base rate in the sample). Therefore, minimum possible scores are not reported in discriminatory indices.

PPV and NPV are not reported for the HCR-20\textsuperscript{V3} Total scores analysis, as the SPJ tool is not designed to be utilised in this manner, rather, the
clinical decision making is designed to result in SRR’s (Guy, 2008). Reporting such results encourages improper use of the tools (i.e., using Total scores as cut-offs for predicting violent recidivism), and defeats the purpose and essence of SPJ tools. Therefore, these indices based on coordinates along the curve have only been provided for the HCR-20\textsuperscript{V3} SRR, as this is the final predictive entity produced by the tool. In a similar vein, these indices of discrimination were not reported for the VRAG-R Total scores because the Bin number is the final predictive entity produced by the tool, and demonstrates use of the tool in the intended way.

5.6.1.5 Comparing AUC’s

Hypothesis 3 relates to the comparison of the HCR-20\textsuperscript{V3} and the VRAG-R to their predecessors (i.e., the HCR-20\textsuperscript{V2} and original VRAG respectively), using the same participants for comparisons.

Comparative situations differ from those where a single ROC is produced. Due to the statistical complexity of comparing ROC curves, there are two main issues to consider: How to compare the curves (i.e., which statistic will be employed); and how to determine or approximate the distribution under the null (Martinez-Camblor et al., 2013). Previous research has suggested that the AUC is robust to departures from normality, and biases in estimates of AUC’s for parametric and nonparametric procedures are very small (i.e., the two methods yield similar estimates of AUC’s) (Hajian-Tilaki, Hanley, Joseph, & Collet, 1997). Therefore, concerns about distorted or imprecise AUC estimates should not be an issue in selecting parametric or non-parametric approaches, regardless of the type of distribution.
Previous studies have used the AccuROC program (Vida, 1997) which uses the chi-squared statistic to calculate and compare ROC curves and their AUC’s (e.g., Douglas et al., 1999). Alternatively, others have used bootstrap measures to perform this analysis (e.g., Gardner et al., 1996). Another popular method as developed by DeLong and colleagues (1998) involves a completely non-parametric model based on properties of the Mann-Whitney statistic, and is preferred for continuous data. In the current study, a different method as described by Hanley and McNeil (1983) was utilised as it was previously employed in Chu’s (2010) research, amongst others (e.g., Snowden et al., 2007; Snowden et al., 2010). This statistically sophisticated technique uses nonparametric statistical properties and Gaussian modelling assumptions for estimating the variance of two correlated areas (DeLong et al., 1998). This nonparametric approach is superior to the parametric approach because it makes no assumptions regarding the distribution of parameters (Campbell, Douglas, & Bailey, 1989). Although previous research suggests that the DeLong and colleagues (1988) method and the Hanley and McNeil (1983) method yield very similar AUC estimates, the Hanley and McNeil (1983) method was also selected due to the ordinal nature of the data in some analyses.

Hanley and McNeil (1983) addressed the issue of comparing AUC’s that have been derived from the same participants. They note the importance of taking into account the correlation between the two areas that stems from the paired nature of the data. That is, “when two or more empirical curves are constructed based on tests performed on the same individuals, statistical analysis on differences between curves must take into account the correlated nature of the data” (DeLong et al., 1988, p. 837).
When two or more ROC curves are generated using the same participants at one time period, the comparison of the two algorithms requires more formal statistical criteria to decipher whether differences in accuracy are random or real. Hanley and McNeil (1983) state that in order to determine whether the difference between the two areas (generated by the same participants) is significant, $z$ tests for dependent groups must be calculated where: $A_1$ and $SE_1$ refer to the AUC and estimated standard error of the ROC area associated with modality 1; and $A_2$ and $SE_2$ refer to these quantities for modality 2. Modality refers to the risk assessment tools being compared. The $r$ quantity represents the correlation between the two areas ($A_1$ and $A_2$) by using the same sample of patients. The formula for calculating the critical ratio $z$ is as follows:

$$z = \frac{A_1 - A_2}{\sqrt{(SE_1^2 + SE_2^2 - 2rSE_1SE_2)}}$$

The null hypothesis in this case is that underlying areas are the same from both modalities. The $-2rSE_1SE_2$ term is important because failure to remove fluctuations that the paired design has already eliminated from the sampling variability will lead to a large denominator (and therefore a small critical ration $z$), which reduces the chance of detecting differences between the two areas.

$r$ is calculated using two intermediate correlation coefficients calculated using either the Pearson product-movement correlation method or Kendall’s Tau. Kendall’s Tau is obviously more suited to non-parametric data and in tools scored on ordinal scales (Hanley et al., 1983). The first correlation is the
correlation coefficient for the risk ratings given to non-recidivists from the two modalities ($R_A$). The second is the correlation coefficient for the ratings given to recidivists from the two modalities ($R_N$). The formula for $r$ is provided below, however a table developed by Hanley and McNeil (1983) provides for simple calculation of $r$ through input of the AUC's (i.e., $A_1$ and $A_2$) and the correlation coefficients $R_A$ and $R_N$ (see Appendix K for Hanley and McNeil’s (1983) Table on correlation coefficients).

$$r = \frac{\text{Cov}(A_1,A_2)}{\text{SE}_1\text{SE}_2}$$

Due to the association between the two areas, it would be inappropriate to calculate the Standard Errors (SE’s) of difference without taking this correlation into account. This is because the two indices are likely to “fluctuate in tandem when derived from a single sample” (Hanley et al., 1983; p. 840).

In the current analyses, the student researcher calculated the AUC’s (i.e., $A_1$ and $A_2$); their respective standard errors (i.e., $\text{SE}_1$ and $\text{SE}_2$) and the correlation coefficients (i.e., $R_N$ and $R_A$) using SPSS Version 22. These were then used to obtain the correlation coefficient ($r$) through Hanely and McNeil’s (1983) table. These values were then imputed into the critical ratio $z$ formula using a Microsoft Excel spreadsheet. Once the critical ratio $z$ was calculated, tables of normal distribution are used where cutoff values $z \geq +1.96$ or $z \leq 1.96$ represent evidence that the “true” areas under the ROC curves were different (p. 840). Although these equations were calculated by hand, a
statistical software package used in biomedical research, MedCalc for Windows, version 16.2.1 (MedCalc Software, 2016), was used to verify the results. This software has been used in previous studies to compare (e.g., Grann et al., 2000), and is similar to other computer software like AccuROC (Vida, 1997), used in studies like Douglas, Ogloff, Nicholls and Grant (1999). However, the MedCalc software was selected because it allows for pairwise comparisons of dependent ROC curves, with specifications that calculations are computed according to the Hanley and McNeil (1983) methodology.

In summary, the Hanley and McNeil (1983) method for comparing AUC’s was adopted in the current research because it provided for a more sensitive comparison by accounting for the smaller sampling variability induced by studying each participant twice. Using an unpaired Z-test that assumes independence may fail to detect significant differences between areas, but this method ensures that comparisons can be made more sensitive by taking the smaller sampling variability into account. Guidelines around sample size requirements for conducting such analyses have also been provided and are discussed within the relevant hypothesis in Chapter Six, Results.

5.6.2 Correlational Analyses - Kendall’s Tau

Although correlational techniques are dependent on base rates of violence, they were reported in the current study because of future consideration for their common use in meta-analytic studies (Strub et al., 2014). Kendall’s tau (symbolized by τ) is a non-parametric correlation that was used for correlational analyses. Kendall’s tau was selected instead of Spearman’s correlation coefficient as research suggests that, in comparison
to Spearman’s Correlation Coefficient, it is a more superior estimate of the actual correlation in the population and leads to more accurate generalisations (Field, 2009; Howell, 1997).

Kendall’s Tau was used to address hypothesis 1: Association between violence risk assessment tools and violent recidivism; and concurrent validity between previous and current versions of the respective tools. Concurrent validity pertains to the relationship between instruments (i.e., whether an instrument correlates with another previously-validated instrument that measures the same construct). Therefore, Kendall’s Tau was used to evaluate the concurrent validity between the: HCR-20V3 and the HCR-20V2, the original VRAG and the VRAG-R, and the HCR-20V3 and VRAG-R.

Interpretation of the strength of the relationship was as follows: 0 – 0.25 = little or no relationship, 0.26 – 0.49 = fair relationship, 0.50 – 0.69 = moderate to good relationship, 0.70 – 0.89 = good relationship, 0.90 – 1.0 = very good relationship (Munro, 2001).

5.6.3 Chi-Square Test of Independence

A Chi-Squared Test of Independence (also known as the Chi-Square Test of Association) was used to explore the relationship between categorical variables. It works by comparing the frequency of cases in categories of one variable across categories of another variable, and indicates what values should be expected if there is no association between the two variables (Pallant, 2009). The chi-square test requires two assumptions: Independence of data and expected frequencies greater than five. Independence of data was met as each participant only contributed to one cell of the contingency table. Expected frequencies greater than five indicate that the sampling
distribution approximates a chi-square distribution. While expected frequencies should ideally be greater than five, in larger contingency tables it is acceptable to have up to 20% of expected frequencies below five, however there is a loss of statistical power. When the assumption is met, the Pearson Chi-square statistic is reported, which indicates if the variables are independent or related. Any violation of expected frequencies indicates that the sample size is too small and that the sampling distribution will not approximate a chi-squared distribution. This violation is noted in the specific analyses and was dealt with using Fisher’s exact test and/or data pooling.

Effect sizes are reported using Cramer’s V (as opposed to the phi coefficient) because contingency tables in the analyses are always larger than 2 x 2. Judging the size of the effect was done in accordance with Pallant (2009) guidelines where .01 = Small effect; .30 = Medium effect and .50 = Large effect. Odd’s Ratios are reported for smaller contingency tables (i.e., HCR-20^V3) but not for larger contingency tables (i.e., VRAG-R) as this would not be a useful nor efficient way of interpreting effect size (Field, 2009). The Chi-square Test of Independence was used to assess hypothesis 6 (discrimination) and a component of Hypothesis 1 (association).

5.6.4 Binomial Logistic Regression

Binomial (binary) logistic regression was used to answer hypothesis 5 (HCR-20^V3 predictive strength of scales) and hypothesis 9 (incremental validity). Logistic regression allows for the prediction of a discrete outcome (such as group membership) from a set of variables and provides the probability of a particular outcome for each group (Tabachnick & Fidell, 2007).
Binomial logistic regression differs from multinomial logistic regression when the outcome has only two categories (e.g., violent vs. non-violent; or recidivist vs. non-recidivist) (Field, 2009). Unlike the multiple regression equation where the outcome is predicted from predictor variables, logistic regression predicts the probability of the outcome occurring, given known values of the predictors (Field, 2009). The analysis attempts to maintain strong prediction whilst eliminating predictors – thus simplifying the prediction equation. Parameter values are estimated using maximum-likelihood estimation, to maximise probability (Field, 2009).

It is a flexible technique that requires no assumptions regarding the distribution of predictor variables (i.e., predictors do not have to be normally distributed, linear or have equal variances within groups) (Tabachnick et al., 2007). Interactions between predictor variables can also be accounted for through two-way interactions or higher-order interactions in the case of many predictor variables.

Logistic regression does however require that IV’s are continuous or categorical (ordinal IV’s need to be treated as either continuous or nominal variables) and a dichotomous DV. Sample size requirements vary from 15 – 50 cases per IV (Laerd Statistics, 2015). Therefore, a sample of 100 participants was considered sufficient.

There are three different types of logistic regression: Standard, sequential (hierarchical) or stepwise. In contrast to standard where all predictors enter the equation simultaneously and stepwise where statistical criteria is used in a hypothesis-generating context, sequential regression is hypothesis-testing driven and involves the researcher specifying the order of entry of predictors into the model (Tabachnick et al., 2007). Kerlinger (1986,
p. 454) advocates that theory should be used to guide the order of entry as there exists no “substitute for the depth of knowledge of the research problem”. This opinion has been mirrored by others noting that “the data analyst knows more than the computer” (Henderson & Velleman, 1981, p. 391). Therefore, sequential logistic regression was selected. Based on the volume of research available on the HCR-20, the student researcher could engage in informed decision-making about predictor variable entry.

There are two types of inferential tests in logistic regression: tests of models and tests of individual predictors. As an initial step, a comparison of the constant-only model with a model containing the constant and all predictors (i.e., Constant-Only vs. Full Model) is used to assess whether predictor as a group contribute to prediction of the outcome. The log-likelihood test is used to assess the fit of the model by indicating how much unexplained information remains after the model has been fitted (Field, 2009). Therefore, large log-likelihood statistic values indicate poorly fitting models (Field, 2009). Evaluation of the contribution of individual predictors to the model can be assessed through the Wald test, likelihood ratio test and Lagrange multiplier test. The Wald test was used due to it being a default, simple process (Tabachnick et al., 2007). The Wald Statistic has a Chi-square distribution and informs whether predictors make significant contributions to the prediction of the outcome (Field, 2009). The model chi-square statistic measures the difference between the constant-only model and the model with predictors included.

Logistic regression was also used to assess the incremental validity associated with the Presence and Relevance features of the HCR-20$^V_3$, as well as the incremental validity between the HCR-20$^V_3$, VRAG-R and PCL-R.
It was hypothesised that the Relevance ratings would result in a significant increase in the model fit, demonstrating incremental validity over the Presence ratings. The analysis was not conducted for SRR's because theoretically, the SRR has already accounted for Relevance ratings. It was also hypothesised that the HCR-20\textsuperscript{V3} and PCL-R would add incremental validity to the PCL-R, but that the converse would not be true. Hierarchical regression was not used to test incremental validity related hypotheses because the dependent variable (violent recidivism) was coded as a dichotomous outcome, not continuous (Laerd Statistics, 2015).

5.6.5 Odds Ratios

Odds Ratio's (OR's) are crucial to the interpretation of logistic regression. OR’s represent the probability of an event occurring divided by the probability of that event not occurring (Field, 2009). Statistically, the OR’s are interpreted as a change in odds resulting from a unit change in the predictor. OR’s can be used when the DV is dichotomous. Hence, logistic regression is essentially multiple regression but where the outcome variable is categorical and the predictor variables are continuous or categorical (Field, 2009).

In the current study, OR’s were used because they provide practical probability estimates by evaluating whether scores above or below the median distinguish between outcome groups (e.g., recidivists and non-recidivists). OR’s were derived from a non-parametric Chi-Squared Test of Independence ($\chi^2$), with the median score as the point of differentiation.

Median splits were calculated using the SPSS Rank procedure. The median split was computed using the 50% percentile rank point. This divided
continuous scores into categorical data (i.e., two groups, those falling below and above the median). Ties (when participant scores were equal to the median) were dealt with by using the mean of the two values, which is the default option.

An OR value greater than 1 indicates that as the predictor increases, so does the odds of the outcome occurring, and vice versa for values less than 1 (Field, 2009). For example, an OR of 2.5 can be interpreted as one group being 2.5 times more likely than another group to possess a particular criterion (e.g., post-release violence). The inverse is applicable to those scoring below the median (i.e., less likely). Fleiss and colleagues note that OR’s between 2.5 and 3 are considered large and clinically important associations (Fleiss, Williams, & Dubro, 1986).

5.6.6 Survival Analysis

Survival analysis was used to incorporate timing of violence into analyses. Survival analysis deals with the amount of time it takes for a specific event to occur (i.e., a violent offence). By definition then, participants who did not commit a violent act during the follow-up period are considered to have ‘survived’, whilst those who do commit a violence offence are deemed as having ‘failed’. The outcome variable is time at risk until the event occurs; and the grouping variable is scores on violence risk assessment tools (Kleinbaum & Klein, 2012). For participants who survive, this quantity is equal to time at risk until the end of follow-up.

The sample size was considered sufficient for survival analyses, as Eliason (1993) recommends sample sizes of 60 if 5 or fewer covariates are estimated. Missing data presents an unusual complication in survival
analyses. Missing data can occur because cases survive until the end of the follow-up period, and therefore time to failure is not known. These cases with unknown survival time are referred to as censored cases (Tabachnick et al., 2007).

There are two different types of survival analysis available, depending on the nature of the data and research question at hand: The product-limit (Kaplan-Meier) method (Kaplan & Meier, 1958); and the Cox proportional-hazards model (also known as Cox Regression; Cox, 1972). The Kaplan-Meier method compares survival curves between groups to investigate whether there is a statistically significant difference between them. Alternatively, the Cox Regression assesses the relationship between survival time and predictors, while controlling for other covariates (Tabachnick et al., 2007). The Kaplan-Meier is a non-parametric estimator. It calculates survival statistics each time an event is observed and produces mean and median survival time, standard errors and a 95% confidence interval. Because mean survival times are effected by censored cases however, they are calculated based on areas under survival curves as opposed to averages (Hosmer, Lemeshow, & May, 2008). Thus, when using Kaplan-Meier, the median is more important as a measure of central tendency (Laerd Statistics, 2015). There are also several statistics for evaluating group differences (Log-rank, Tarone and Wilcoxon). These tests differ depending on case weighting and the point at which groups start to diverge.

In the current study, the Kaplan-Meier method were used to provide an estimation and interpretation of survivor functions and comparison of functions (Kleinbaum et al., 2012). Survival analyses were run based on both Total scores and risk categories. Median splits were used to group Total
scores into categorical data (participants scoring above or below the median Total score). This was necessary to prevent the Kaplan-Meier survival function plotting a different survival curve for each Total score. Whilst the HCR-20\textsuperscript{V3} already had risk categories in the form of SRR’s, risk categories were created for the VRAG-R by grouping the nine VRAG-R Bins into three groups to create categories of Low (bottom three Bins), Moderate (middle three Bins) and High (top three Bins).

5.6.7 Intraclass Correlation Coefficient

Reliability refers to the consistency of a measure and reflects the proportion of “real” information about a construct captured by the instrument (Field, 2009; Landers, 2015). Although inter-rater correlations can be computed using Pearson’s \( r \) in the case of two raters, this method fails to take the variance between raters into account (Barrett, 2001). Other measures of agreement such as Cohen’s Kappa (\( k \)) aren’t able to capture “near misses” in disagreement, or are not available through statistical packages such as SPSS, as used in the current study (Shea, 2015).

The Intraclass Correlation Coefficient (ICC) was used to investigate the inter-rater reliability of the HCR-20\textsuperscript{V3} and the VRAG-R in the current study because it captures the strength of association between scores as well as the level of agreement between raters. ICC is defined as “the correlation between one measurement (either a single rating or a mean of several ratings) on a target and another measurement obtained on that target” (Shrout & Fleiss, 1979, p. 422). The ICC is a useful measure of inter-rater reliability because it
is highly flexible and takes differences between ratings and correlations between raters into account (Landers, 2015).

There are several different ICC’s that are distinguished along several decision points: One-way or two-way designs; consistency of order of ratings or agreement on the variable being rated; raters as fixed or random variables; and reliability of single ratings or averaged ratings (Howell, 2010). As a result, six different formulas for calculating ICC are available. Selection of formulas depends on the purpose of the research, study design and the type of measurements used.

There are three different models of ICC: 1) One-way random; 2) two-way random; or 3) two-way mixed. In one-way random, each participant is assessed by a different set of randomly selected raters. In two-way random, each participant is assessed by each rater, and raters have been randomly selected. In two-way mixed, each participant is assessed by each rater, and the raters are the only raters of interest. The type of ICC refers to systematic differences between trials or raters and must also be selected: Consistency or Absolute Agreement. Furthermore, one restriction applies to the use of the ICC: Each case rated must have the same number of ratings (Landers, 2015). This requirement was satisfied in the current study.

The ICC is displayed as ICC (#, k). The # symbol relates to the model chosen, whereas the second symbol (k) denotes the form. Form is based on whether the reliability is calculated from a single measurement, or the average across two or more measurements taken by different raters.

The ICC two-way random (i.e., ICC (2,k)) was selected because 1) each reliability participant was assessed by each rater, and 2) raters are seen
as stemming from a population of all possible raters, thus enabling
generalisation to other raters. Absolute Agreement was also selected
because systematic differences were deemed relevant. This decision is
based on guidelines for computing ICC outlined by Landers (2015) and
Shrout and Fleiss (1979), which is the primary text regarding ICC.

Reliability analyses were run separately for: The HCR-20\textsuperscript{V3} and VRAG-R Prorated Total scores; the HCR-20\textsuperscript{V3} scale scores; HCR-20\textsuperscript{V3} Relevance ratings; and HCR-20\textsuperscript{V3} SRR's. Both the single-measure (ICC\textsubscript{1}) and average measure (ICC\textsubscript{2}) are reported in the results. This provides data on the
reliability of the risk assessment tools when a single professional is using it
(ICC\textsubscript{1}) or when several clinicians are rating and relying on averaged ratings
(ICC\textsubscript{2}). Both are reported, however ICC\textsubscript{1} is considered to be the actual
reliability rating, given that assessments are typically made by a single
professional in the clinical context. Notably, ICC\textsubscript{2} ratings are generally higher
due to multiple assessors in the decision-making process.

ICC ranges from 0.0 to 1.0, and values are high when there is little
variation in scores given by raters. An ICC of .7 indicates that 70% of
variability in scores captured by the instrument(s) represents the construct
being measured, whilst 30% represents random variation. Fleiss (1981) has
provided direction for judging reliability by ICC values where < 0.40 is
considered ‘poor’, 0.40 – 0.59 is ‘moderate’, 0.60 – 0.74 is ‘good’ and ≥ 0.75
is ‘excellent’. This is reasonably similar to other classifications described in
the literature (e.g., Landis et al., 1977).
Chapter Six

Results

6.0 Introduction

The results are presented in order of the stated hypotheses. Due to having 100 participants, percentages are not reported in analyses where $N = 100$.

6.1 Hypothesis 1: Association

It was hypothesised that there would be an association between violent recidivism and scores on the HCR-20$^{V3}$ and VRAG-R. The HCR-20$^{V3}$’s Total scores and SRR’s will correlate significantly with violent recidivism. The VRAG-R’s Total scores and Bins will correlate significantly with violent recidivism. Kendall’s Tau was used to test this hypothesis (one tailed because the hypothesis was directional). A Chi-squared Test of Independence was also employed to test associations amongst nominal variables.

Association between violence risk assessment tools and violence

Table 24 displays the correlations between the tools (HCR-20$^{V3}$ SRR, HCR-20$^{V3}$ Total score, VRAG-R Bin, VRAG-R Total score) and violent recidivism. As can be seen, all correlations were positive and significant, with a fair to moderate strength of association.
Table 24

*Kendall’s Tau Correlations between Violence Risk Assessment Tool Scores and Violent Recidivism*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCR-20\textsuperscript{V3} Prorated Total Score</td>
<td>100</td>
<td>.29**</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V3} SRR</td>
<td>100</td>
<td>.49**</td>
</tr>
<tr>
<td>VRAG-R Prorated Total Score</td>
<td>42</td>
<td>.31**</td>
</tr>
<tr>
<td>VRAG-R Bins</td>
<td>42</td>
<td>.32**</td>
</tr>
</tbody>
</table>

*Note* * Significant at the $p < .05$ level; ** Significant at the $p < .01$ level.

Chi squared Tests of Association were used to assess the association between nominal variables (e.g., HCR-20\textsuperscript{V3} SRR’s and VRAG-R Bins) and violent recidivism. For the HCR-20\textsuperscript{V3}, a 3 x 2 design was used (Low, Moderate, High x Violent offence/ No Violent offence). Table 25 provides the contingency table of HCR-20\textsuperscript{V3} SRR scores and post-release violence. It can be seen that as the SRR levels increase from Low to High, the number of participants charged with a violent offence post-release increases. Notably, none of the participants rated Low committed a violent offence in the follow-up period.
Table 25

Cross-tabulation of HCR-20V3 SRR Scores and Post-Release Violence

<table>
<thead>
<tr>
<th>SRR</th>
<th>Charged with Violent Offence Post-Release</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No (100%)</td>
</tr>
<tr>
<td>Low</td>
<td>16 (100%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>22 (57.89%)</td>
</tr>
<tr>
<td>High</td>
<td>12 (26.09%)</td>
</tr>
</tbody>
</table>

The assumption of expected frequencies was met as all expected counts were greater than 5 (0% of cells had expected counts less than 5). Therefore, the Chi-square test for association was reported as opposed to Fisher’s exact test. There was a statistically significant association between HCR-20V3 SRR and violent offence charges during the follow-up period, $\chi^2(2) = 27.469$, $p = .000$. Cramer’s $V$ indicated that the strength of the association was large, $\phi_c = .524$, $p = .000$. In comparing groups, for participants who received a High SRR, the odds of being charged with a violent offence post-release was 3.9 times higher than if they had received a Moderate SRR. Odd’s Ratios are not computed for Low SRR because no participants rated Low were charged with violent offences during the follow-up period.

**VRAG-R**

For the VRAG-R, a 9 x 2 design was used. As an initial step, a cross-tabulation of VRAG-R Bins and violent recidivism was computed (see Table 26). Through eye-ball-ing the data, it can be seen that as the Bin number increases, the number of participants who committed violent offences...
generally increases. Notably, none of the participants in Bin numbers 1-4 committed a violent offence in the follow-up period.

Table 26

*Cross-tabulation of VRAG-R Bins and Number of Participants Engaging in Post-Release Violence*

<table>
<thead>
<tr>
<th>Bin Number</th>
<th>Charged with Violent Offence Post-Release</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1 – 3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4 – 6</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>7 – 9</td>
<td>9</td>
<td>12</td>
</tr>
</tbody>
</table>

The assumption of expected frequencies was not met as all expected counts were below 5 (100% of cells had expected counts less than 5).

Although there is some leeway in larger tables, the minimum requirement is that no more than 20% of expected counts can be less than 5 (Field, 2009).

Therefore, Fisher’s exact test (FET) was reported as opposed to the Chi-square test of association. There was a statistically non-significant association between VRAG-R Bins and violent offence charges during the
follow-up period (\( p = .118 \), two tailed FET). Cramer's V indicated that the strength of the association was large but non-significant, \( \varphi_c = .501 \), \( p = .139 \).

Figure 3 displays a clustered bar chart. Violent offence charges are present from Bins 5-9, peaking clearly at Bin 9 (scores < -24 represent Bin 1; scores > 27 represent Bin 9).

Figure 3 Clustered Bar Chart displaying Violent Recidivism across VRAG-R Bins

The meaningfulness of this analysis is limited by the number of cases falling into each category. This is concerning because Fisher's exact test is ill-suited to large contingency tables, however it was adopted here because it is
the common remedy for violations of expected frequencies (Field, 2009; Laerd Statistics, 2015).

To overcome this, the VRAG-R Bins were pooled to increase frequencies in categories. Three categories were created: Low (bottom 3 Bins); Moderate (middle 3 Bins) and High (top 3 Bins). The analysis was re-run in a 3 x 2 design. The assumption of expected frequencies was not met as 33.3% of the cells had expected counts below 5. Fisher’s exact test is reported. There was a statistically non-significant association between VRAG-R Bins and violent offence charges during the follow-up period ($p = .161$, two-tailed FET). Cramer’s V indicated that the strength of the association was fair-to-moderate but non-significant, $\phi_c = .284$, $p = .161$.

**Association between Tools and Frequency of Violent Charges Post-Release**

The association between frequency of violent offence charges post-release and scores on the HCR-20$^{V3}$ and VRAG-R was also explored. Kendall’s Tau was used to evaluate these hypotheses (two-tailed because the hypothesis was non-directional). Significant positive correlations between frequency of violent charges in the follow-up period and the HCR-20$^{V3}$ Prorated Total score ($\tau = .265$, $p < .01$) and the VRAG-R Prorated Total score ($\tau = .290$, $p < .05$) were found.

**6.2 Hypothesis 2: Concurrent Validity**

It was hypothesised that the HCR-20$^{V3}$ and VRAG-R will demonstrate acceptable concurrent validity (i.e., significant correlations of $r > .50$, as per Munro, 2001). Previous and current versions of the tools will correlate significantly (i.e., the HCR-20$^{V2}$ and HCR-20$^{V3}$ will correlate significantly, as
will the VRAG and VRAG-R). It was also hypothesised that the HCR-20\textsuperscript{V3} Total score and H-Scale score will correlate significantly with the VRAG-R.

The Kendall’s Tau statistic and Pearson’s correlation was used to evaluate this hypothesis. Only 99 participants were available for the HCR-20 analyses because one participant exceeded the threshold for missing data on the HCR-20\textsuperscript{V2} dataset. Table 27 displays the correlations between violence risk assessment tools. As can be seen, all correlations were positive and significant at the \( p < .01 \) level, with a moderate strength of association.

Table 27

*Kendall’s Tau Correlations between Total Scores on Violence Risk Assessment Tools (HCR-20\textsuperscript{V2}, HCR-20\textsuperscript{V3}, VRAG and VRAG-R)*

<table>
<thead>
<tr>
<th></th>
<th>HCR-20\textsuperscript{V2}</th>
<th>HCR-20\textsuperscript{V3}</th>
<th>VRAG</th>
<th>VRAG-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCR-20\textsuperscript{V2}</td>
<td>1</td>
<td>.53**</td>
<td>.42**</td>
<td>.41**</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V3}</td>
<td>1</td>
<td>.47**</td>
<td>.45**</td>
<td></td>
</tr>
<tr>
<td>VRAG</td>
<td>1</td>
<td>.57**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VRAG-R</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

*Note. N = 99 for HCR-20 analyses; N = 41 for VRAG analyses. * Significant at the \( p < .05 \) level; ** Significant at the \( p < .01 \) level (one-tailed).*

The Kendall’s Tau statistic was adopted as a precautionary measure, due to variations in normality of the data at the univariate level. However, given the moderate strength of correlations, correlations were re-run using the Pearson’s correlation. As Table 28 displays, all correlations were stronger than identified in Kendall’s Tau correlations, and significant at the \( p < .01 \) level.
Table 28

Pearson’s Correlations between Total Scores on Violence Risk Assessment Tools (HCR-20\textsuperscript{V2}, HCR-20\textsuperscript{V3}, VRAG and VRAG-R)

<table>
<thead>
<tr>
<th></th>
<th>HCR-20\textsuperscript{V2}</th>
<th>HCR-20\textsuperscript{V3}</th>
<th>VRAG</th>
<th>VRAG-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCR-20\textsuperscript{V2}</td>
<td>1</td>
<td>.74**</td>
<td>.62**</td>
<td>.67**</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V3}</td>
<td>1</td>
<td>.61**</td>
<td>.66**</td>
<td></td>
</tr>
<tr>
<td>VRAG</td>
<td></td>
<td></td>
<td>.79**</td>
<td></td>
</tr>
<tr>
<td>VRAG-R</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Note. \(N = 99\) for HCR-20 analyses; \(N = 41\) for VRAG analyses. * Significant at the \(p < .05\) level; ** Significant at the \(p < .01\) level (one-tailed).

Association between HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} Scales

Correlations between the H, C and R-scales of the HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} were computed and are displayed in Table 29. Only 99 participants were available for the C-scale analyses because one participant on the HCR-20\textsuperscript{V2} dataset exceeded the threshold for missing scores on C-Scale.

Table 29

Pearson’s Correlations between HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} Scales.

<table>
<thead>
<tr>
<th></th>
<th>H-Scale</th>
<th>C-Scale</th>
<th>R-Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCR-20\textsuperscript{V3}:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-Scale</td>
<td>.66**</td>
<td>.31**</td>
<td>.27**</td>
</tr>
<tr>
<td>C-Scale</td>
<td>.53**</td>
<td>.63**</td>
<td>.48**</td>
</tr>
<tr>
<td>R-Scale</td>
<td>.44**</td>
<td>.53**</td>
<td>.50**</td>
</tr>
</tbody>
</table>

Note. \(N = 100\) for HCR-20 H-Scale and R-Scale analyses; \(N = 99\) for HCR-20 R-Scale analyses. * Significant at the \(p < .05\) level; ** Significant at the \(p < .01\) level (one-tailed).
**Association between Static Measures on Tools**

It was hypothesised that the static component of the HCR-20\(^{V3}\) (i.e., the Historical scale) and the VRAG-R as a static measure would be significantly correlated. A Kendall’s Tau correlation (one-tailed due to directionality) indicated that they were significantly correlated, \(\tau = .561, p < .01\).

**6.3 Hypothesis 3: Prediction**

It was hypothesised that the ability of the HCR-20\(^{V3}\) and the VRAG-R to predict future violence would be greater than chance; and that AUC’s would be similar and not significantly different from AUC’s produced by the HCR-20\(^{V2}\) and VRAG. The ROC was used to test this hypothesis. Excel formula template was used to calculate Positive Predictive Value (PPV) and Negative Predictive Value (NPV).

**6.3.1 HCR-20\(^{V3}\)**

HCR-20\(^{V3}\) Prorated Total scores \(N = 100\) discriminated between violent and non-violent participants \([AUC = 0.70, SE = .05, 95\% CI of .59 - 0.80, p = 001]\). Results were significant with a moderate effect size. HCR-20\(^{V3}\) SRR \(N = 100\) discriminated between violent and non-violent participants \([AUC = 0.77, SE = .05, 95\% CI of 0.68 – 0.86, p = .000]\). Results were significant with a large effect size. Figure 4 displays the ROC curves for both the HCR-20\(^{V3}\) Prorated Total scores and HCR-20\(^{V3}\) SRR’s. The indices of discrimination are also displayed for the HCR-20\(^{V3}\) SRR in Table 30 below across Moderate and High categories.
Figure 4  ROC Curves for the HCR-20$^v3$ Total Scores and SRR’s with Violent Recidivism as Outcome.

Diagonal segments are produced by ties.
Table 30

*Indices of Discrimination for HCR-20\textsuperscript{V3} SRR as Cut-Off*

<table>
<thead>
<tr>
<th>Cut-Off</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Moderate</td>
<td>1.00</td>
<td>0.32</td>
<td>0.60</td>
<td>1.00</td>
</tr>
<tr>
<td>High</td>
<td>0.68</td>
<td>0.76</td>
<td>0.74</td>
<td>0.70</td>
</tr>
</tbody>
</table>

*Note. N = 100. PPV = Positive Predictive Value; NPV = Negative Predictive Value. Minimum possible scores are not reported because all cases are test-positive at this level.*

ROC curves were also produced for the HCR-20\textsuperscript{V3} Relevance ratings Total scores (*N* = 100; Relevance ratings calculated at item level as opposed
to sub-item level). Relevance ratings discriminated between violent and non-violent participants [AUC = 0.71, SE = .05, 95% CI of 0.60 - 0.81, p = .000]. Results were significant with a large effect size. The ROC curve is displayed in Figure 5.

AUC’s for the HCR-20V3 Presence and Relevance ratings were compared using Hanley and McNeil’s (1983) method. The average area between curves was 70.03. The Kendall’s Tau correlations between the paired ratings were $r_N = 0.61$ (non-violent participants) and $r_A = 0.67$ (violent participants), with average correlation being 0.64. These values are then inputted into the Table provided by Hanley and McNeil (1983; see Appendix K), which displays the correlation coefficient between two ROC areas as a function of average ratings and areas. With an average area of 70.03 and an average correlation of 0.64, the correlation between areas is approximately $r = 0.61$. The critical $z$-ratio was then calculated to test the null hypothesis that difference between areas is random. Using the equation outlined in section 5.6.1.5, $z = -0.23$. This quantity was then referred to tables of normal distribution using conservative cut off value of $z \geq 1.96$. Given that the resulting $z$ ratio is below this, this was taken as evidence that the “true” ROC areas were not different.
6.3.2 VRAG-R

The VRAG-R Prorated Total score \((N = 42)\) discriminated between violent and non-violent participants \([\text{AUC} = 0.71, \ SE = .08, \ 95\% \ CI \ of \ 0.56 – 0.87, \ p = .019]\). Results were significant with a large effect size. Indices of discrimination were identical for the VRAG-R Bins, \([\text{AUC} = 0.71, \ SE = .08, \ 95\% \ CI \ of \ 0.56 – 0.87, \ p = .019]\). Indices of discrimination are provided for the VRAG-R Bins in Table 31. The ROC curves are displayed in Figure 6.
Table 31

*Indices of Discrimination for VRAG-R Bin Numbers as Cut-off*

<table>
<thead>
<tr>
<th>Bin Number</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>.04</td>
<td>.46</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>.08</td>
<td>.47</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>.34</td>
<td>.55</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>.84</td>
<td>.47</td>
<td>.57</td>
<td>.78</td>
</tr>
<tr>
<td>7</td>
<td>.63</td>
<td>.60</td>
<td>.57</td>
<td>.66</td>
</tr>
<tr>
<td>8</td>
<td>.47</td>
<td>.69</td>
<td>.56</td>
<td>.61</td>
</tr>
<tr>
<td>9</td>
<td>.31</td>
<td>.91</td>
<td>.75</td>
<td>.61</td>
</tr>
</tbody>
</table>

*Note.* $N = 42$. PPV = Positive Predictive Value; NPV = Negative Predictive Value. Minimum possible scores are not reported because all cases are test-positive at this level. Cut-off for Bin Number 3 is not computed because no participants were assigned to Bin Number 3. Indices of discrimination for Bin Number 3 are identical to Bin Number 2.
Table 32 below summarises the performance of the HCR-20\textsuperscript{V3} and VRAG-R in discriminating violent from non-violent participants.

Table 32

<table>
<thead>
<tr>
<th>Measure</th>
<th>AUC</th>
<th>Effect Size</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCR-20\textsuperscript{V3} Total Score</td>
<td>0.70</td>
<td>Medium</td>
<td>0.59 – 0.80</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V3} SRR</td>
<td>0.77</td>
<td>Large</td>
<td>0.68 – 0.86</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V3} Relevance Ratings</td>
<td>0.71</td>
<td>Large</td>
<td>0.60 – 0.81</td>
</tr>
<tr>
<td>VRAG-R Total Score</td>
<td>0.71</td>
<td>Large</td>
<td>0.56 – 0.87</td>
</tr>
</tbody>
</table>

Note. $N = 100$. VRAG-R $N = 42$. 
It was also hypothesised that the observed levels of accuracy produced by the HCR-20\textsuperscript{V3} and VRAG would be similar to levels of accuracy demonstrated by the HCR-20\textsuperscript{V2} and VRAG respectively. The ROC analysis was used to determine this hypothesis.

The procedure for comparing ROC curves to determine whether differences are real or random is outlined in Chapter Five Section 5.6.1.5. The Kendall’s Tau correlation coefficient was adopted in the current analyses due to the non-normal distribution of the data. A two-tailed analyses was run as there was no reason or previous research to indicate a priori that modalities two (i.e., HCR-20\textsuperscript{V3}; VRAG-R) would be superior to modalities one (i.e., HCR-20\textsuperscript{V2}; VRAG).

6.3.3 HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3}

The HCR-20\textsuperscript{V3} Prorated Total scores were used as the basis of comparison because no SRR is available in the HCR-20\textsuperscript{V2}. Only 99 participants were available for the analyses because one participant in the HCR-20\textsuperscript{V2} dataset met the exclusionary criteria (i.e., the participant had nine missing Presence scores). Therefore, this participant was also excluded in the HCR-20\textsuperscript{V3} variable. For this reason, HCR-20\textsuperscript{V3} AUC’s reported here differ somewhat from AUC’s reported in Table 32. Separate ROC curves were run for HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3}. Results are summarized in Table 33 below, followed by Figure 7, displaying the combined ROC curves.
Table 33
*Indices of Discrimination for the HCR-20\textsuperscript{V2} and the HCR-20\textsuperscript{V3} with Violent Recidivism as Outcome*

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>SE</th>
<th>p-Value</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V2}</td>
<td>.77</td>
<td>.046</td>
<td>.000</td>
<td>.682</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V3}</td>
<td>.69</td>
<td>.053</td>
<td>.001</td>
<td>.588</td>
</tr>
</tbody>
</table>

*Note. N = 99.*

Figure 7  ROC Curves for the HCR-20\textsuperscript{V2} and the HCR-20\textsuperscript{V3} with Violent Recidivism as Outcome

The calculations for comparing AUC’s are now produced from the data in Table 33 above. The HCR-20\textsuperscript{V2} AUC was 77.27 (SE = 4.6) and the HCR-
AUC was 69.08 ($SE = 5.3$). The average area was therefore 73.18. The Kendall Tau correlations between the paired ratings were $r_N = 0.46$ (non-violent participants) and $r_A = 0.56$ (violent participants). Therefore the average correlation between ratings was 0.51. These values are then inputted into the Table provided by Hanley and McNeil (1983; see Appendix K). With an average area of 73.18 and an average correlation of 0.51, the correlation between areas is approximately $r = 0.47$. The critical $z$-ratio was then calculated to test the null hypothesis that difference between areas is random. Using the equation outlined in section 5.6.1.5, $z = 1.60$. This quantity was then referred to tables of normal distribution using conservative cut off value of $z \geq 1.96$. Given that the resulting $z$ ratio is below this, this was taken as evidence that the “true” ROC areas were not different.

This process was reviewed using MedCal statistical software, which provides pairwise comparisons and more detailed information than manually computed calculations. The results indicated that the difference between areas was 0.08 ($SE = 0.04$). The $z$ statistic was 1.89 ($p = 0.06$), indicating that the two compared areas are not significantly different.

### 6.3.4 VRAG and VRAG-R

Analyses were run using both the Total Prorated Scores and Bin categories. Separate ROC curves were run for the VRAG and the VRAG-R. Results are summarized in Table 34 below, followed by the combined ROC curves, displayed in Figure 8. Forty-one participants were available in the
VRAG dataset for analyses due to one participant meeting exclusionary
criteria on the original VRAG dataset.

Table 34

*Summary of the Predictive Performance of the VRAG and the VRAG-R with
Violent Recidivism as Outcome*

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>SE</th>
<th>p-Value</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>VRAG Total Score</td>
<td>.73</td>
<td>.079</td>
<td>.011</td>
<td>.578, .886</td>
</tr>
<tr>
<td>VRAG Bins</td>
<td>.74</td>
<td>.078</td>
<td>.010</td>
<td>.583, .889</td>
</tr>
<tr>
<td>VRAG-R Total Score</td>
<td>.72</td>
<td>.079</td>
<td>.016</td>
<td>.565, .875</td>
</tr>
<tr>
<td>VRAG-R Bins</td>
<td>.72</td>
<td>.079</td>
<td>.017</td>
<td>.563, .874</td>
</tr>
</tbody>
</table>

*Note.* N = 41.
Figure 8  ROC Curves for the VRAG and VRAG-R (Total Scores and Bins) with Violent Recidivism as Outcome

Separate ROC curves for the VRAG and VRAG-R Total score comparisons and Bin number comparisons are also presented for visual comparison in Figure 9 and Figure 10.
Figure 9  VRAG and VRAG-R Total Score Comparisons

Diagonal segments are produced by ties.
The calculations for comparing AUC’s are now produced from the data in Table 34 above. The VRAG Total score AUC was 73.21 (SE = 7.9) and the VRAG-R Total score AUC was 72.01 (SE = 7.9). The average area was therefore 72.61. The Kendall Tau correlations between the paired ratings were $r_N = 0.56$ (non-violent participants) and $r_A = 0.52$ (violent participants). Therefore the average correlation between ratings was 0.54. These values were inputted into the Table provided by Hanley and McNeil (1983), which displays the correlation coefficient between two ROC areas as a function of average ratings and areas. With an average area of 72.61 and an average correlation of 0.54, the correlation between areas is approximately $r = 0.50$. The critical $z$-ratio was then calculated, whereby $z = 0.15$. This quantity was
then referred to tables of normal distribution using conservative cut off value of $z \geq 1.96$. Given that the resulting $z$ ratio is below this, this was taken as evidence that the “true” ROC areas were not different.

This process was reviewed using MedCal statistical software. Once again, AUC’s were basically identical: VRAG (AUC = 0.73, SE = 0.08, 95% CI = 0.57 – 0.86); VRAG-R (AUC = 0.72, SE = 0.08, 95% CI = 0.56 – 0.85). The results indicated that the difference between areas was 0.01 (SE = 0.06). The $z$ statistic was 0.20 ($p = 0.85$), indicating that the two compared areas are not significantly different.

The process was repeated for the VRAG, where computations were based on VRAG Bins (AUC = 73.56; SE = 7.8) and VRAG-R Bins (AUC = 71.89; SE = 7.9), the average area therefore being 72.73. The Kendall Tau correlations between the paired ratings were $r_N = 0.69$ (non-violent participants) and $r_A = 0.64$ (violent participants). Therefore the average correlation between ratings was 0.66. These values were then inputted into the Table provided by Hanley and McNeil (1983), which displays the correlation coefficient between two ROC areas as a function of average ratings and areas. With an average area of 72.73 and an average correlation of 0.66, the correlation between areas is approximately $r = 0.62$. The critical $z$-ratio was then calculated, $z = 0.24$. This quantity was then referred to tables of normal distribution using the conservative cut off value of $z \geq 1.96$. Given that the resulting $z$ ratio is below this, this was taken as evidence that the “true” ROC areas were not different.
MedCal computations revealed that AUC’s were similar: VRAG Bin (AUC = 0.74, \(SE = 0.08, 95\% \text{ CI} = 0.56 – 0.86\)); VRAG-R Bin (AUC = 0.72, \(SE = 0.08, 95\% \text{ CI} = 0.56 – 0.85\)). The results indicated that the difference between areas was 0.02 (\(SE = 0.06\)). The \(z\) statistic was 0.30 (\(p = 0.77\)), indicating that the two compared areas were not significantly different.

6.4 Hypothesis 4: HCR-20\textsuperscript{V3} Predictive Strength of Scales

It was hypothesised that the H-Scale would be the strongest predictor of post-release violence over the entire follow-up; however, when controlling for time at risk, the C-Scale and R-Scale would emerge as the strongest predictors of post-release violence over the entire follow-up period. Binomial logistic regression was used to test this hypothesis.

Prior to conducting the analysis, assumptions of linearity and multicollinearity were checked to ensure the validity of the analysis. The assumption of linearity was assessed using the Box-Tidwell procedure, which involved adding interaction terms for each continuous IV and its natural log transformation (Hosmer & Lemeshow, 2000). Statistical significance was set at \(p < .017\) due to a Bonferroni correction (i.e., \(.05 \div 3 = .017\)). Results indicated that all interactions were non-significant, meaning that the assumption of linearity of the logit was met across all three predictor variables (H-Scale, C-Scale and R-Scale). The assumption of multicollinearity was assessed by running a linear regression analysis and using values of Tolerance and VIF as indicators of the presence of collinearity. Menard (1995) recommended that tolerance values less than 0.1 indicate collinearity problems. Myers (1990) recommended that VIF values greater than 10 also
indicate collinearity problems. Result indicated that tolerance values were greater than 0.1 and Variance Inflation Factor (VIF) values were less than 10 across all variables. This leads to the conclusion that there were no serious multicollinearity concerns.

6.4.1 AUC’s for HCR-20^V3 Scales

AUC’s were also run for the HCR-20^V3 scales as part of assessing the predictive strength of scales. Results are displayed in Table 35 below, with ROC curves displayed in Figure 11.

Table 35

<table>
<thead>
<tr>
<th>Scale</th>
<th>AUC</th>
<th>SE</th>
<th>Sig.</th>
<th>95% CI: Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-Scale</td>
<td>.65</td>
<td>.055</td>
<td>.012</td>
<td>.539</td>
<td>.754</td>
</tr>
<tr>
<td>C-Scale</td>
<td>.69</td>
<td>.053</td>
<td>.001</td>
<td>.590</td>
<td>.797</td>
</tr>
<tr>
<td>R-Scale</td>
<td>.62</td>
<td>.056</td>
<td>.039</td>
<td>.509</td>
<td>.730</td>
</tr>
</tbody>
</table>

Note. 95% CI = 95% Confidence Interval. Under the non-parametric assumption.
6.4.2 Violent Recidivism

A sequential binomial logistic regression was performed to ascertain the effects of H-Scale, C-Scale and R-Scale Total scores on the likelihood that participants had a violent offence charge post-release. The H-Scale was added in the first block, and C-Scale and R-Scale added in the second block. Theoretically, this allows for a baseline of a static risk scale, and assessment of additional variance captured by the dynamic risk scales.

Initially, a test of the full model with all predictors entered against a constant only model was statistically significant, \( \chi^2 (1, N = 100) = 7.402, p < .01 \). This indicated that the addition of one or more of the predictor variables to the model would significantly increase its predictive power, and that the
predictors as a set reliably distinguished between those who recidivated violently and those who did not.

When the predictor variables were added to the model, the model was statistically significant, $\chi^2 (3, N = 100) = 14.605, p < .005$. The model explained 18.1% (Nagelkerke $R^2$) of the variance in post-release violent offence charge and correctly classified 64% of cases. Sensitivity was 68%, specificity was 60%, PPV was 62.9% and NPV was 65.2%. Of the three predictor variables, only the C-Scale score was statistically significant (as show in Table 36). Therefore, the C-Scale was the strongest predictor of post-release violent offence charges. An increase to any of the three scales was associated with an increase of having a violent offence charge post-release. The increase in odds was 1.13; 1.35 and 1.01 for the H-Scale, C-Scale and R-Scale respectively.

### Table 36

**Sequential Logistic Regression Predicting Likelihood of a Violent Offence Charge Post-Release based on H-Scale, C-Scale and R-Scale Prorated Scores**

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>p</th>
<th>Odds Ratio</th>
<th>95% CI for OR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>H-Scale</td>
<td>.121</td>
<td>.091</td>
<td>1.763</td>
<td>1</td>
<td>.184</td>
<td>1.129</td>
<td>.944</td>
</tr>
<tr>
<td>C-Scale</td>
<td>.300</td>
<td>.134</td>
<td>5.014</td>
<td>1</td>
<td>.025*</td>
<td>1.350</td>
<td>1.038</td>
</tr>
<tr>
<td>R-Scale</td>
<td>.010</td>
<td>.114</td>
<td>.008</td>
<td>1</td>
<td>.928</td>
<td>1.010</td>
<td>.808</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.825</td>
<td>1.360</td>
<td>7.910</td>
<td>1</td>
<td>.005</td>
<td>.022</td>
<td></td>
</tr>
</tbody>
</table>

*Note: CI = Confidence Interval.* * indicates significance at the $p < .05$ level.
An examination of residuals assisted in determining whether the model was a good one. Examination of residuals helped to 1) isolate points where the data fitted poorly and (2) isolate points that exerted undue influence on the model (Field, 2009). No outliers were found in the analysis. There were no unusually high values of Cook’s distance, which indicates that there were no individual cases having an influential effect on the model. No standardized residuals approached 3, which would normally warrant inspection.

6.4.3 Controlling for Time at Risk

Time at Risk was added to the model to assess whether the predictive strength of HCR-20\textsuperscript{v3} scales changed given opportunity to offend. It was hypothesised that, when controlling for time at risk, the C-Scale and R-Scale would emerge as the strongest predictors of post-release violence over the entire follow-up period. It was expected that the Time at Risk variable itself would emerge as a significant predictor in the analyses, given that re-offending is reasonably related to and dependent on time at risk. Binomial logistic regression was used to test this hypothesis.

When the time at risk variable was included, results indicated that tolerance values were greater than 0.1 and VIF values were less than 10 across all variables. Thus, the conclusion that there were no serious multicollinearity concerns remained. The assumption of linearity when time at risk was included (new Bonferroni correction set at $p < .012$) was not met for the time at risk variable ($p = .000$). This meant that time at risk was not linearly related to the logit of the dependent variable. Despite this violation to the linearity assumption, the sequential logistic regression was re-run without intervention at first, and then with steps taken to account for the violation.
Time at Risk in Block 1; H-Scale in Block 2 and C-Scale with R-Scale in Block 3. The model was still statistically significant overall $\chi^2 (4, N = 100) = 59.571$, $p < .005$. The Hosmer and Lemeshow Test indicated that the model was not a poor fit in predicting categorical outcomes ($p = .275$) (statistical significance indicates poor fitting models). Time at risk and the C-Scale emerged as the only significant predictors. Whilst this violation remains a limitation of the study, the results are reported below.

Table 37

*Sequential Logistic Regression Predicting Likelihood of a Violent Offence Charge Post-Release based on Time at Risk, H-Scale, C-Scale and R-Scale*

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>$p$</th>
<th>Odds Ratio</th>
<th>95% CI for OR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time at Risk</td>
<td>-.001</td>
<td>.000</td>
<td>28.380</td>
<td>1</td>
<td>.000*</td>
<td>.999</td>
<td>.998</td>
</tr>
<tr>
<td>H-Scale</td>
<td>.031</td>
<td>.123</td>
<td>.062</td>
<td>1</td>
<td>.803</td>
<td>1.031</td>
<td>.810</td>
</tr>
<tr>
<td>C-Scale</td>
<td>.358</td>
<td>.183</td>
<td>3.835</td>
<td>1</td>
<td>.050*</td>
<td>1.431</td>
<td>1.000</td>
</tr>
<tr>
<td>R-Scale</td>
<td>-.033</td>
<td>.147</td>
<td>.051</td>
<td>1</td>
<td>.821</td>
<td>.967</td>
<td>.725</td>
</tr>
<tr>
<td>Constant</td>
<td>-.355</td>
<td>1.752</td>
<td>.014</td>
<td>1</td>
<td>.839</td>
<td>.701</td>
<td></td>
</tr>
</tbody>
</table>

*Note: CI = Confidence Interval. * indicates significance at the $p < .05$ level.*

In overcoming the violation of the linearity assumption, transformation to the data was considered but not computed due to the difficulty this lends to interpretation (i.e., consequential changes to hypothesis testing) (Grayson, 2004). Therefore it was decided that the best course of action would either be to remove problematic data or bootstrap the data.
A review of the Casewise List indicated that there were 4 cases (#13, 16, 21 and 93) that did not fit the model well. All these cases had Studentized Residuals approaching 3, which indicates potential outliers. Therefore, the assumptions were re-run with these cases excluded from the analysis. The assumption of linearity continued to be violated across all three interaction forms. Given that the assumption was violated either way, efforts to re-run the analysis with the four cases excluded were not taken.

Alternatively, a bootstrap procedure was applied to the data (simple sampling with 1000 samples and a bias-corrected and accelerated confidence interval type) to allow for a more robust model. Results indicated that, once again, time at risk ($p = .001$) and C-Scale ($p = .047$) were the only significant predictors of outcome. The model was still statistically significant overall $\chi^2 (4, N = 100) = 59.571$, $p < .001$, with 82% of cases being correctly classified.

6.4.4 Addition of HCR-20$^\text{V3}$ Total Score Variable

The analysis was then extended to include the HCR-20$^\text{V3}$ Total score as a predictor. When the Total score was added, there were no serious multicollinearity concerns as tolerance values were greater than 0.1 and VIF values were less than 10 across all variables. However, the assumption of linearity remained violated for Time at Risk.

The sequential logistic regression was re-run with Time at Risk in Block 1; HCR-20$^\text{V3}$ scales in Block 2; and the HCR-20$^\text{V3}$ Total score in Block 3. At baseline, the residual chi-square statistic indicated that the addition of one or more of the variables would significantly improve its predictive power ($\chi^2 = 46.230; p < .01$). The model fit improved significantly through the
addition of Time at Risk ($\chi^2 = 53.128; p < .01$). Correct classification of cases improved from 50% in the baseline model to 82%. However, the Hosmer and Lemeshow test indicated that the model fit was poor ($p < .05$).

Through introducing the HCR-20$^{V3}$ scales, the model improved significantly ($\chi^2 = 59.571; p < .01$). Correct classification of cases remained at 82%. The Hosmer and Lemeshow test indicated that the model had a better fit ($p = .275$). As in the previous analysis, Time at Risk ($p = .000$) and C-Scale ($p = .050$) were the only significant predictors. Through introducing the HCR-20$^{V3}$ Total score, the model improved significantly ($\chi^2 = 62.575; p < .01$) and remained a good fit (Hosmer and Lemeshow $p = .461$). However, only Time at Risk emerged as a significant variable ($p < .01$).

Given that the Total score represents information already captured by the three scales, its failure to emerge as a significantly predictor is not surprising. As displayed in Table 38, when a separate logistic regression is run for the Total score and the scales, the Total score is a more significant predictor. The model including Time at Risk and scales indicates the model improves significantly from baseline ($\chi^2 = 59.571; p < .01$) and has a good fit (Hosmer and Lemeshow $p = .275$). The alternative model including Time at Risk and Total scores improves significantly from baseline ($\chi^2 = 57.139; p < .01$) and has a good fit, Hosmer and Lemeshow $p = .081$).
Table 38

**Sequential Logistic Regression Predicting Likelihood of a Violent Offence Charge Post-Release based on Time at Risk, HCR-20\(^{V3}\) Scales and HCR-20\(^{V3}\) Total Scores.**

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>p</th>
<th>Odds Ratio</th>
<th>95% CI for OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time at Risk</td>
<td>-.001</td>
<td>.000</td>
<td>28.380</td>
<td>1</td>
<td>.000**</td>
<td>.999</td>
<td>.998</td>
</tr>
<tr>
<td>H-Scale</td>
<td>.031</td>
<td>.123</td>
<td>.062</td>
<td>1</td>
<td>.803</td>
<td>1.031</td>
<td>.810</td>
</tr>
<tr>
<td>C-Scale</td>
<td>.358</td>
<td>.183</td>
<td>3.835</td>
<td>1</td>
<td>.050*</td>
<td>1.431</td>
<td>1.000</td>
</tr>
<tr>
<td>R-Scale</td>
<td>-.033</td>
<td>.147</td>
<td>.051</td>
<td>1</td>
<td>.821</td>
<td>.967</td>
<td>.725</td>
</tr>
<tr>
<td>Time at Risk</td>
<td>-.001</td>
<td>.000</td>
<td>28.380</td>
<td>1</td>
<td>.000**</td>
<td>.999</td>
<td>.998</td>
</tr>
<tr>
<td>Total Score</td>
<td>.098</td>
<td>.050</td>
<td>3.797</td>
<td>1</td>
<td>.051</td>
<td>1.103</td>
<td>.999</td>
</tr>
</tbody>
</table>

*Note: CI = Confidence Interval. * indicates significance at the \(p \leq .05\) level. ** indicates significance at the \(p < .01\) level.*

**6.4.5 Interactions between HCR-20\(^{V3}\) Scales**

The hypothesis was then extended to take interactions between and among the HCR-20\(^{V3}\) scales into account to identify the strongest predictors of post-release violence. The binomial logistic regression was re-run, with H-Scale in Block 1, C-Scale and R-Scale in Block 2, and interactions between scales in Block 3. Initially, a test of the full model with all predictors entered against a constant only model was statistically significant, \(\chi^2 (1, N = 100) = 7.402, p < .01\). This indicated that the addition of one or more of the predictor variables to the model would significantly affect its predictive power. When the predictor variables were added to the model, the model was statistically
significant, $\chi^2 (6, N = 100) = 20.900, p < .005$. The model explained 25.1% (Nagelkerke $R^2$) of the variance in the likelihood of a post-release violent offence charge and correctly classified 63% of cases. None of the variables significantly predicted the outcome, however the C-Scale ($p = 0.77$) and the interaction between H-Scale and R-Scale ($p = 0.62$) were the strongest predictors.

When time at risk was added to the analysis in Block 1, two significant predictors emerged: Time at risk ($p = .000$) and the interaction between H-Scale and R-Scale ($p = .050$). Table 39 below provides a summary of this analysis with violent recidivism as dependent variable, and the following as IV’s: Time at Risk, H-Scale, C-Scale, R-Scale, H-Scale x C-Scale interaction, H-Scale x R-Scale interaction, and C-Scale x R-Scale interaction.
Table 39

*Sequential Logistic Regression Predicting Likelihood of a Violent Offence Charge Post-Release based on Time at Risk, HCR-20\(^{\text{V3}}\) Prorated Scale Scores, and Interactions amongst HCR-20\(^{\text{V3}}\) Scales.*

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>p</th>
<th>Odds Ratio</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time at Risk</td>
<td>-.001</td>
<td>.000</td>
<td>26.489</td>
<td>1</td>
<td>.000*</td>
<td>.999</td>
<td>.998</td>
<td>.999</td>
</tr>
<tr>
<td>H-Scale</td>
<td>.817</td>
<td>.560</td>
<td>2.132</td>
<td>1</td>
<td>.144</td>
<td>2.264</td>
<td>.756</td>
<td>6.777</td>
</tr>
<tr>
<td>C-Scale</td>
<td>-.893</td>
<td>1.248</td>
<td>.512</td>
<td>1</td>
<td>.474</td>
<td>.409</td>
<td>.035</td>
<td>4.727</td>
</tr>
<tr>
<td>R-Scale</td>
<td>1.593</td>
<td>1.019</td>
<td>2.445</td>
<td>1</td>
<td>.118</td>
<td>4.920</td>
<td>.668</td>
<td>36.245</td>
</tr>
<tr>
<td>H-Scale*C-Scale</td>
<td>.036</td>
<td>.081</td>
<td>.196</td>
<td>1</td>
<td>.658</td>
<td>1.037</td>
<td>.884</td>
<td>1.216</td>
</tr>
<tr>
<td>H-Scale*R-Scale</td>
<td>-.143</td>
<td>.073</td>
<td>3.825</td>
<td>1</td>
<td>.050*</td>
<td>.867</td>
<td>.752</td>
<td>1.000</td>
</tr>
<tr>
<td>C-Scale*R-Scale</td>
<td>.100</td>
<td>.075</td>
<td>1.767</td>
<td>1</td>
<td>.184</td>
<td>1.105</td>
<td>.954</td>
<td>1.280</td>
</tr>
<tr>
<td>Constant</td>
<td>-8.380</td>
<td>6.715</td>
<td>1.557</td>
<td>1</td>
<td>.212</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note:* OR = Odds Ratio. CI = Confidence Interval. * indicates significance at the .05 level.

### 6.5 Hypothesis 5: Discrimination

It was hypothesised that participants scoring above the median on the HCR-20\(^{\text{V3}}\) and VRAG-R will be significantly more likely than those scoring below the median to 1) recidivate generally; and 2) recidivate violently, during the follow-up period. A Chi-squared Test of Independence was used to determine this hypothesis.
6.5.1 HCR-20\textsuperscript{V3}

**General Recidivism**

A Chi-squared Test of Independence was used to assess the association between those scoring above and below the median score on the HCR-20\textsuperscript{V3} and general recidivism. A 2 x 2 design was used. Firstly a median split was conducted in SPSS ($N = 100; \text{Md} = 30$).

The assumption of expected frequencies was met as all expected counts were greater than 5 (0% of cells had expected counts less than 5). Therefore, the Chi-square test for association was reported as opposed to Fisher’s exact test. There was a statistically significant association between scoring above or below the HCR-20 Median and criminal offence charges during the follow-up period, $\chi^2(1) = 9.974, p = .002$. Cramer’s V indicated that the strength of the association was moderate and significant, $\varphi_c = .316, p = .002$.

The odds of being charged with a criminal offence during the follow-up period was 3.89 times higher if the participant scored above the median as opposed to below.

**Violent Recidivism**

A Chi-squared Test of Independence was used to assess the association between those scoring above and below the median score on the HCR-20\textsuperscript{V3} and violent recidivism. A 2 x 2 design was used. Firstly a median split was conducted in SPSS ($N = 100; \text{Md} = 30$).

The assumption of expected frequencies was met as all expected counts were greater than 5 (0% of cells had expected counts less than 5). Therefore, the Chi-square test for association was reported as opposed to Fisher’s exact test. There was a statistically significant association between scoring above
or below the HCR-20 Median and violent offence charges during the follow-up period, $\chi^2(1) = 6.784$, $p = .009$. Cramer’s V indicated that the strength of the association was small-to-moderate and significant, $\varphi_c = .260$, $p = .016$.

The odds of being charged with a violent offence during the follow-up period was 2.89 times higher if the participant scored above the HCR-20$^{\text{V3}}$ median as opposed to below.

6.5.2 VRAG-R

**General Recidivism**

A Chi-squared Test of Independence was used to explore whether there was an association between scoring above and below the median score on the VRAG-R and general recidivism. A 2 x 2 design was used. Firstly a median split was conducted in SPSS using VRAG-R Prorated Total score ($N = 42; Md = 11.50$).

The assumption of expected frequencies was met as all expected counts were greater than 5 (0% of cells had expected counts less than 5). Therefore, the Chi-square test for association was reported as opposed to Fisher’s exact test. There was a statistically non-significant association between scoring above or below the VRAG-R Median and criminal offence charges during the follow-up period, $\chi^2(1) = 3.500$, $p = .059$. Cramer’s V indicated that the strength of the association was small-to-moderate and non-significant, $\varphi_c = .289$, $p = .118$. The odds of being charged with a criminal offence during the follow-up period was 3.3 times higher if the participant scored above the VRAG-R median as opposed to below.
Violent Recidivism

The assumption of expected frequencies was met as all expected counts were greater than 5 (0% of cells had expected counts less than 5). Therefore, the Chi-square test for association was reported as opposed to Fisher’s exact test. There was a statistically non-significant association between scoring above or below the VRAG-R Median and violent offence charges during the follow-up period, $\chi^2(1) = 2.403, p = .107$. Cramer’s V indicated that the strength of the association was small-to-moderate and non-significant, $\phi_c = .239, p = .215$. The odds of being charged with a violent offence during the follow-up period was 2.6 times higher if the participant scored above the VRAG-R median as opposed to below.

6.6 Hypothesis 6: Survival Time

It was hypothesised that there would be statistically significant differences in survival time for general and violent recidivism between HCR-20^V3 and VRAG-R scores (participants scoring above or below the median Total score) and risk categories (HCR-20^V3 SRR’s and grouped VRAG-R Bins). Kaplan-Meier Survival Analysis was used to test this hypothesis.

6.6.1 HCR-20^V3

General Recidivism – HCR-20^V3 Total Scores

Kaplan-Meier survival analysis was used to determine whether there was a statistically significant difference in recidivism patterns between participants who scored above or below the median score on the HCR-20^V3.

A median split was conducted using the SPSS Rank procedure ($N = 100; Md = 30$). Although a median of 30 seems high, 46% of participants
received a SRR of High, and 54 participants had a Total score greater or
equal to 30. The median split was computed using the percentile rank (50%)
point. Ties were dealt with using the mean of the two values, which is the
default option. This procedure divided continuous scores into categorical data
whereby total prorated scores were divided into two groups to represent
those scoring above \( N = 53 \) or below \( N = 47 \) the median.

Figure 12 displays Kaplan-Meier survival curves for HCR-20\textsuperscript{V3} Total
score median splits and time to first offence. The percentage of censored
cases across groups was dissimilar: below median (53.2\%) and above
median (22.6\%). Groups were compared using the log rank test. Groups
differed significantly in survival distributions, \( \chi^2(1) = 16.085 \), \( p < .001 \).

Figure 12  Kaplan-Meier Survival Curves for HCR-20\textsuperscript{V3} Total Prorated Score
Median Splits and Time to First Offence
General Recidivism – HCR-20\textsuperscript{V3} SRR’s

The distribution of participants across HCR-20\textsuperscript{V3} SRR’s was as follows: Low ($N = 16$), Moderate ($N = 38$) or High ($N = 46$) risk. The average survival time for general recidivism was 1442.93 days (Min = 1; Max = 4502). The mean survival time across SRR’s was: Low ($M = 2204.13$); Moderate ($M = 1767.16$) and High ($M = 910.33$) days.

Figure 13 displays Kaplan-Meier survival curves for HCR-20\textsuperscript{V3} SRR’s and time to first offence. Each line represents a SRR group. Censored cases are indicated by vertical lines, where participants ‘survived’ throughout their time at risk. Notably however, the percentage of censored cases present across groups was dissimilar: Low (87.5%), Moderate (47.4%) and High (10.9%).

Groups ($N = 100$) were compared using the log rank test. Groups differed significantly in survival distributions, $\chi^2(2) = 25.509$, $p < 0.001$. The overall median survival time for recidivists was 987 days, 95% CI [406.210, 1567.790]. The median survival time for participants who received a HCR-20\textsuperscript{V3} SRR of High was 431 days, 95% CI [183.965, 678.035]; and 1473 days, 95% CI [0.000, 3189.008] for the Moderate group. Furthermore, SPSS could not compute median survival times for the Low group as 87.5% of cases were censored, and median survival time reflects the time at which 50% of participants reached the event.

Log-rank pairwise comparisons were run to determine which SRR groups had different survival distributions. A Bonferroni correction was made with statistical significance accepted at the $p = .0167$ level. Low differed significantly from high ($\chi^2 = 18.950$, $p < .01$) but did not differ from moderate
(\chi^2 = 5.484, p = .019). Moderate differed significantly from high (\chi^2 = 11.386, p < .001).

Figure 13 Kaplan-Meier Survival Curves for HCR-20\textsuperscript{V3} SRR’s and Time to First Offence.

### Violent Recidivism – HCR-20\textsuperscript{V3} Total Scores

Kaplan-Meier survival analysis was used to determine whether there was a statistically significant difference in violent recidivism patterns between participants who scored above or below the median on the HCR-20\textsuperscript{V3}. As before, a median split was conducted and participant’s total prorated scores were divided into two groups to represent those scoring above (N = 53) or below (N = 47) the median score.
Figure 14 displays Kaplan-Meier survival curves for HCR-20\textsuperscript{V3} Total prorated score median splits and time to first violent offence. The percentage of censored cases present across groups was dissimilar: Below Median (63.8\%) and Above Median (37.7\%). Groups ($N = 100$) were compared using the log rank test. Groups differed significantly in survival distributions, $\chi^2(1) = 11.355$, $p = < .001$.

Figure 14 Kaplan-Meier Survival Curves for HCR-20\textsuperscript{V3} Total Prorated Score Median Splits and Time to First Violent Offence.

Violent Recidivism – HCR-20\textsuperscript{V3} SRR’s

Once again, HCR-20\textsuperscript{V3} SRR distribution was as follows: Low ($N = 16$), Moderate ($N = 38$) or High ($N = 46$) risk. The average survival time for violent
recidivists was 1935.19 days (Min = 1; Max = 4502). The mean survival time across SRR’s was: Low (M = 2597); Moderate (M = 2199.63) and High (M = 1486.54) days.

Figure 15 displays Kaplan-Meier survival curves for HCR-20$^v3$ SRR’s and time to first violent offence. Each line represents a SRR group. The percentage of censored cases present across groups was dissimilar: Low (100%), Moderate (57.9%) and High (26.1%). Due to the Low category having 100% censored cases (i.e., no participants rated Low committed a violent offence), there were no failure events, and SPSS could not compute relevant statistics. Furthermore, SPSS could not compute median survival times for the Moderate group as 57.9% of cases were censored, and median survival time reflects the time at which 50% of participants reached the event.

Groups ($N = 100$) were compared using the log rank test. Groups differed significantly in survival distributions, $\chi^2(2) = 21.694$, $p < .001$. The overall median survival time for recidivists was 1473 days, 95% CI [114.285, 2831.715]. The median survival time for participants who received a HCR-20$^v3$ SRR of High was 696 days, 95% CI [370.313,1021.687].

Log-rank pairwise comparisons were run to determine which SRR groups had different survival distributions. The log-rank test was used because it places equal weight to all observations (Tabachnick et al., 2007). A Bonferroni correction was made with statistical significance accepted at the $p = .0167$ level. Low differed significantly from moderate ($\chi^2 = 6.942$, $p = .008$) and high ($\chi^2 = 15.881$, $p = .000$). Moderate differed significantly from high ($\chi^2 = 8.480$, $p = .004$).
Figure 15 Kaplan-Meier Survival Curves for HCR-20 V3 SRR’s and Time to First Violent Offence.

6.6.2 VRAG-R

General Recidivism – VRAG-R Total Scores

Kaplan-Meier survival analysis was used to determine whether there was a statistically significant difference in violent recidivism patterns between participants who scored above or below the median on the VRAG-R. The median value for VRAG-R Total prorated score was \( Md = 11.50 \). Participant’s Total prorated scores were divided into two groups to represent those scoring above \( (N = 21) \) or below \( (N = 21) \) the median.

Figure 16 displays Kaplan-Meier survival curves for VRAG-R Total prorated score median splits and time to first violent offence. The percentage
of censored cases present across groups was somewhat dissimilar: below median (57.1%) and above median (28.6%). Groups (N = 42) were compared using the log rank test. The survival distributions of groups differed significantly, $\chi^2(1) = 5.198$, $p = .023$.

Figure 16 Kaplan-Meier Survival Curves for VRAG-R Total Prorated Score Median Splits and Time to First Offence.

![Survival Functions](image)

**General Recidivism – VRAG-R Risk Bins**

The nine VRAG-R Bins were grouped into three groups to create categories of Low (bottom three Bins), Moderate (middle three Bins) and High (top three Bins). The distribution of participants were as follows: Low ($N = 2$),
Moderate \((N = 19)\) or High \((N = 21)\) risk. The mean survival time across groups was: Low \((M = 448)\); Moderate \((M = 1649.16)\) and High \((M = 927)\) days.

Figure 17 displays Kaplan-Meier survival curves for the grouped VRAG-R Bins and time to first offence. Each line represents a risk group. The percentage of censored cases present across groups was dissimilar: Low (100\%), Moderate (52.60\%) and High (28.60\%). Due to the Low category having 100\% censored cases (i.e., no participants rated Low committed a violent offence), there were no failure events, and SPSS could not compute relevant statistics. Furthermore, SPSS could not compute median survival times for the Moderate group as 52.60\% of cases were censored, and median survival time reflects the time at which 50\% of participants reached the event.

Groups \((N = 42)\) were compared using the log rank test. Groups did not differ significantly in survival distributions, \(\chi^2(2) = 5.454\), \(p = .06\), although approached significance. Log-rank pairwise comparisons were run to determine which risk groups had different survival distributions. The log-rank test was used because it places equal weight to all observations \((\text{Tabachnick et al., 2007})\). A Bonferroni correction was made with statistical significance accepted at the \(p < .0167\) level. Low did not differ significantly from Moderate \((\chi^2 = 0.15, p = .50)\) or High \((\chi^2 = 0.90, p = .34)\), but Moderate differed significantly from High \((\chi^2 = 4.68, p = .03)\).
Violent Recidivism – VRAG-R Total Scores

Kaplan-Meier survival analysis was used to determine whether there was a statistically significant difference in violent recidivism patterns between participants who scored above or below the median on the VRAG-R. The median value for VRAG-R Total prorated score was \( Md = 11.50 \). Participant’s Total prorated scores were divided into two groups to represent those scoring above \( (N = 21) \) or below \( (N = 21) \) the median.

Figure 18 displays Kaplan-Meier survival curves for VRAG-R Total prorated score median splits and time to first violent offence. The percentage of censored cases present across groups was somewhat dissimilar: below
median (66.7%) and above median (42.9%). Groups \((N = 42)\) were compared using the log rank test. The survival distributions of groups did not differ significantly, \(\chi^2(1) = 3.473, p = .062\).

Figure 18 Kaplan-Meier Survival Curves for VRAG-R Total Prorated Score Median Splits and Time to First Violent Offence.

Violent Recidivism – VRAG-R Risk Bins

Once again, the nine VRAG-R Bins were grouped into three groups to create categories of Low (bottom three Bins), Moderate (middle three Bins) and High (top three Bins). The distribution of participants were as follows: Low \((N = 2)\), Moderate \((N = 19)\) or High \((N = 21)\) risk. The mean survival time
across groups was: Low (M = 448); Moderate (M = 1939.58) and High (M = 1291.48) days.

Figure 19 displays Kaplan-Meier survival curves for the grouped VRAG-R Bins and time to first violent offence. Each line represents a risk group. The percentage of censored cases present across groups was dissimilar: Low (100%), Moderate (63.20%) and High (42.9%). Due to the Low category having 100% censored cases (i.e., no participants rated Low committed a violent offence), there were no “events” SPSS could not compute relevant statistics. Furthermore, SPSS could not compute median survival times for the Moderate group as 63.20% of cases were censored, and median survival time reflects the time at which 50% of participants reached the event.

Groups (N = 42) were compared using the log rank test. Groups did not differ significantly in survival distributions, $\chi^2(2) = 3.67$, $p = .16$. Log-rank pairwise comparisons were run to determine which risk groups had different survival distributions. The log-rank test was used because it places equal weight to all observations (Tabachnick et al., 2007). A Bonferroni correction was made with statistical significance accepted at the $p < .0167$ level. Low did not differ significantly from Moderate ($\chi^2 = 0.30$, $p = .59$) or High ($\chi^2 = 0.67$, $p = .41$), neither did Moderate differ significantly from High ($\chi^2 = 3.12$, $p = .08$).
6.7 Hypothesis 7: Incremental Validity

6.7.1 Incremental Validity Within the HCR-20$^{V3}$

A sequential (hierarchical) logistic regression was conducted to assess the incremental validity associated with the Presence and Relevance features of the HCR-20$^{V3}$. It was hypothesised that the Relevance ratings would result in a significant increase in the model fit, demonstrating incremental validity over the HCR-20$^{V3}$ Presence ratings.

Prior to conducting the analyses, assumptions of linearity and multicollinearity were checked to ensure the validity of the analyses. The assumption of linearity was assessed using the Box-Tidwell procedure, with statistical significance being set at $p < .025$ using a Bonferroni correction due
to multiple comparisons. Results indicated that all predictors were non-significant (Presence ratings $p = .374$; Relevance ratings $p = .168$), meaning that they were linearly related to the logit of the dependent variable, and that assumption of linearity of the logit was met.

The assumption of multicollinearity between outcome and predictors was assessed using a linear regression analysis. Result indicated that tolerance values were all greater than 0.1 and VIF values were less than 10 across all variables (Tolerance = .358; VIF = 2.792). This leads to the conclusion that there were no serious multicollinearity concerns.

The Presence ratings were entered into the first block. This produced a significant model, $\chi^2(1) = 12.509, p < .001$. The Hosmer and Lemeshow goodness of fit test indicated that the model was a good fit ($p = .483$; where significance suggests poor model fitting). The model explained 15.7% (Nagelkerke $R^2$) of variance in violent recidivism and correctly classified 59% of cases (sensitivity was 70%, specificity was 48%). The Wald test indicated that the HCR-20$V^3$ Presence ratings were significant predictors of violence ($p = .002$). For a one unit change in HCR-20$V^3$ Prorated Total score, the odds of violent recidivism increased by 1.140 times.

When Relevance ratings were added at Block 2, the model continued to be significant $\chi^2(2) = 14.372, p = .001$ and remained a good fit (Hosmer and Lemeshow test $p = .483$). The model explained 17.8% (Nagelkerke $R^2$) of variance in violent recidivism and correctly classified 67% of cases (sensitivity was 72%, specificity was 62%). Despite these improvements in prediction, the Wald test indicated that Relevance ratings made non-significant improvements in prediction (Wald = 1.819; $p = .177$). It was concluded that the Relevance features of the HCR-20$V^3$ did not demonstrate incremental
validity over the Presence ratings. A summary of the results is provided in Table 40.

Table 40

**Sequential Logistic Regression Predicting Likelihood of Violent Recidivism based on HCR-20\textsuperscript{V3} Presence and Relevance Ratings**

<table>
<thead>
<tr>
<th>Block 1</th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>p</th>
<th>OR</th>
<th>95% CI for OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence</td>
<td>.131</td>
<td>.041</td>
<td>10.066</td>
<td>1</td>
<td>.002</td>
<td>1.140</td>
<td>1.052 1.237</td>
</tr>
</tbody>
</table>

**Block 2**

| Presence    | .059| .066 | 0.802| 1  | .370 | 1.061| 0.932 1.208   |
| Relevance   | .072| .053 | 1.819| 1  | .177 | 1.074| 0.968 1.192   |

*Note.* OR = Odds Ratio; Presence = HCR-20\textsuperscript{V3} Total Prorated score (Presence ratings); Relevance = HCR-20\textsuperscript{V3} Total Relevance ratings. A casewise list was not computed because no outliers were identified in the analysis.

Notably, the decision to add Presence ratings at Block 1 and Relevance ratings at Block 2 was based on practicalities in completing the HCR-20\textsuperscript{V3} (i.e., Presence ratings are always completed prior to Relevance ratings, and it is impractical to assess how relevant a risk factor is to an individual, without first knowing whether or not the factor is present). However, there is no previous empirical research to inform this order of entry. As can be seen in Table 40, Presence and Relevance ratings were both non-significant when considered simultaneously in Block 2. This is likely because the two variables are capturing similar information, becoming redundant when considered simultaneously (indeed, Presence and Relevance ratings were
moderately correlated (\(\tau = .66, p < .01\)). A separate analysis with Relevance ratings being entered into the regression model in isolation also produced a significant model, \(\chi^2(1) = 13.554, p < .001\) in which the HCR-20\(^V3\) Total Relevance ratings were a significant predictor (Wald = 11.479; \(p = .001\)).

Closer inspection using cross-tabulations for the 20 violence risk factors across Presence ratings (Low, Moderate, High) and Relevance ratings (Low, Moderate, High) demonstrated that significant correlations existed for all risk factors (Kendall's Tau-b\(^6\) correlations ranging between \(\tau_b = .39 - .93; p \leq .01\)). Analysis of frequencies within cross-tabulation cells indicated that the highest frequencies were generally in matched ratings (i.e., participants rated 'yes' on Presence were also rated high on Relevance; and participants rated 'no' on Presence were rated low on Relevance. This was the case for the vast majority of risk factors, excluding H3 Relationships (most participants rated yes received a moderate Relevance rating), H4 Employment (most participants rated yes received a moderate Relevance rating), and H8 Traumatic experiences (most participants rated yes received a low Relevance rating). The percentage of participants rated 'Yes' for Presence ratings who also received a High Relevance rating reached as high as 85% (H1 Violence). Participants rated 'no' for Presence always had the highest frequency of participants in the Low Relevance rating category. In fact, for all risk factors, participants with a 'no' Presence rating always had a Low Relevance rating.

\(^{6}\)Kendall's Tau can be reported in two variations: Tau-b and Tau-c. Tau-b is used for square tables (tables where the rows and columns are equal), while Tau-c is used for rectangular tables, which don't have major diagonals.
As an extra precaution, a Stepwise (statistical) logistic regression was computed to evaluate the importance of predictors based solely on statistical criteria (Tabachnick et al., 2007). In the Forward deletion method, stepwise searches for the predictor that best predicts the outcome variable, basing selection on the strength of the correlation between IV’s and the DV. In the Backward deletion method, all IV’s are entered into the model and are deleted one-by-one if they do not significantly contribute to the regression (based on t-test significance).

In the Forward regression, Relevance ratings emerged as the most significant predictor (Wald = 11.479; \( p = .001 \)). However, it is noted that the correlations between IV’s and violent recidivism is almost identical (for Presence ratings, \( r = .29, p < .01 \); for Relevance ratings, \( r = .30, p < .01 \)). In Backward regression, Relevance ratings again emerged as the most significant predictor (Wald = 11.479; \( p = .001 \)). The Presence ratings were removed by the statistical software at Step 2 because no statistically significant contribution was made. However, these results should be interpreted with caution because stepwise regression is controversial method. Interpretation can be hazardous because minor differences (e.g., slight differences in the strength of the correlations - even to second or third decimal place) can profoundly effect the apparent importance of an IV (Tabachnick et al., 2007).

In conclusion, when analysed in separate logistic regressions, both Presence and Relevance ratings emerged as significant predictors of violent recidivism. However, when included together in a stepwise regression, Relevance ratings emerged as the strongest predictor. The advantage of the Relevance ratings should be interpreted with caution, because the Presence
and Relevance ratings were significantly correlated on all risk factors, and produced very similar AUC’s. Analysis at the item-level shows that a significant relationship exists between the ordinal variables across levels, and that a pattern in scoring exists. This suggests that the Presence and Relevance ratings are capturing much of the same information, which leads to redundancy when both are analysed simultaneously. Indeed, ROC curves for the two variables produced identical AUC’s.

6.7.2 Incremental Validity between the HCR-20\textsuperscript{V3}, VRAG-R and PCL-R.

A sequential (hierarchical) logistic regression was conducted to assess the incremental validity associated with the PCL-R, HCR-20\textsuperscript{V3} and VRAG-R associated with violent recidivism. It was hypothesised that the PCL-R would not add significant incremental predictive validity to the HCR-20\textsuperscript{V3} (Total scores and SRR’s) nor the VRAG-R Total scores.

Prior to conducting the analyses, assumptions of linearity and multicollinearity were checked to ensure the validity of the analyses. The assumption of linearity was assessed using the Box-Tidwell procedure, with statistical significance being set at $p < .01$ using a Bonferroni correction due to multiple comparisons (5 terms per model). For the VRAG-R, natural log transformations can only be performed on Total scores greater and excluding zero, as natural log transformations of zero or below result in undefined outcomes. Therefore a constant of +35 was added to the VRAG-R Total scores. Results indicated that all predictors were non-significant, HCR-20\textsuperscript{V3} and PCL-R comparison, (HCR-20\textsuperscript{V3} $p = .270$; PCL-R $p = .189$), VRAG-R and PCL-R comparison (VRAG-R $p = .495$; PCL-R $p = .583$), HCR-20\textsuperscript{V3} and VRAG-R comparison (HCR-20\textsuperscript{V3} $p = .794$; VRAG-R $p = .294$), meaning that
predictors were linearly related to the logit of the dependent variable, and that assumption of linearity of the logit was met.

The assumption of multicollinearity between outcome and predictors was assessed by running a linear regression analysis. Result indicated that tolerance values were all greater than 0.1 and VIF values were less than 10 across all variables (Tolerance = .438 - .521; VIF = 1.918 – 2.284). This leads to the conclusion that there were no serious multicollinearity concerns.

6.7.2.1 HCR-20\textsuperscript{V3} Total Scores and PCL-R comparison

The Casewise list indicated that there were no outliers present in the HCR\textsuperscript{V3} and PCL-R comparison. The HCR-20\textsuperscript{V3} was entered at Block 1, and the PCL-R at Block 2. At stage 1, the HCR-20\textsuperscript{V3} produced a significant model, \( \chi^2(1) = 11.436, p < .01 \). The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the HCR-20\textsuperscript{V3} Total scores was 15.1\% (Nagelkerke \( R^2 \)). At this stage, 60\% of cases were correctly classified.

When the PCL-R ratings were added to the model at Block 2, the model remained significant, \( \chi^2(1) = 22.403, p < .001 \), and the contribution of the PCL-R Total scores was significant, \( \chi^2(1) = 10.967, p = < .01 \). The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the HCR-20\textsuperscript{V3} Total scores and PCL-R in combination was 28\% (Nagelkerke \( R^2 \)). At this stage, 69.5\% of cases were correctly classified. In the final model, only the PCL-R emerged as a significant predictor (Wald = 8.977, \( p = .003 \)). The HCR-20\textsuperscript{V3} was not significant (Wald = 1.036, \( p = .309 \)). Therefore, the contribution of the PCL-R Total scores was significant, and added
incremental validity to the HCR-20 V3 Total scores in the prediction of violent recidivism.

6.7.2.2 HCR-20 V3 SRR’s and PCL-R comparison

The analysis was re-run using the HCR-20 V3 SRR’s. SRR’s were coded as nominal variables, because ordinal variables cannot be used in logistic regression, and therefore ordinal variables must be treated as either nominal or continuous variables (Leard Statistics, 2015). The assumption of linearity is not required for categorical independent variables (Leard Statistics, 2015). Result indicated that tolerance values were all greater than 0.1 and VIF values were less than 10 across all variables (Tolerance = 1.0; VIF = 1.0). The Casewise list indicated that there were was one outlier present in the HCR20 V3 SRR and PCL-R comparison. This studentized residual had a value of -2.468 standard deviations, and was kept in the analysis.

The HCR-20 V3 SRR was entered at Block 1, followed by the PCL-R Total score at Block 2. At stage 1, the HCR-20 V3 SRR produced a significant model, $\chi^2(2) = 32.107, p < .001$. The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the HCR-20 V3 SRR was 38.3% (Nagelkerke $R^2$). At this stage, 71.6% of cases were correctly classified.

When the PCL-R Total scores were added to the model at Block 2, the model remained significant, $\chi^2(3) = 38.884, p < .001$, and the contribution of the PCL-R Total scores was significant, $\chi^2(1) = 6.714, p = < .05$. The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the HCR-20 V3 SRR and PCL-R in combination was 44.8% (Nagelkerke $R^2$). At this
stage, 71.6% of cases were correctly classified. In the final model, only the PCL-R emerged as a significant predictor (Wald = 5.849, \( p = .016 \)). The HCR-20\(^V3\) SRR’s were all non-significant (\( p > .05 \)). Therefore, the contribution of the PCL-R Total scores was significant, and added incremental validity to the HCR-20\(^V3\) SRR’s in the prediction of violent recidivism.

### 6.7.2.3 VRAG-R Total Scores and PCL-R comparison

The Casewise list indicated that there were no outliers present in the VRAG-R and PCL-R comparison. The VRAG-R Total score was entered at Block 1, and the PCL-R was entered at Block 2. At stage 1, the VRAG-R Total scores produced a significant model, \( \chi^2(1) = 6.513, \ p < .05 \). The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the VRAG-R Total scores was 19.2% (Nagelkerke R\(^2\)). At this stage, 59.5% of cases were correctly classified.

When the PCL-R was added to the model at Block 2, the model remained significant, \( \chi^2(2) = 9.715, \ p < .01 \), but the contribution of the PCL-R was non-significant, \( \chi^2(1) = 3.201, \ p = 0.074 \). The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the VRAG-R and PCL-R in combination was 27.6% (Nagelkerke R\(^2\)). At this stage, 61.9% of cases were correctly classified. In the final model, neither the PCL-R (Wald = 2.788, \( p = .095 \)) nor the VRAG-R (Wald = 0.845, \( p = .358 \)) emerged as a significant predictor. Therefore, the contribution of the PCL-R Total scores was non-significant, and did not add incremental validity to the VRAG-R Total scores in the prediction of violent recidivism.
6.7.2.4 HCR-20\textsuperscript{V3} and VRAG-R comparison

A sequential (hierarchical) logistic regression was also conducted to assess the incremental validity associated with the HCR-20\textsuperscript{V3} and VRAG-R, with violent recidivism as outcome. It was hypothesised that the HCR-20\textsuperscript{V3} would add incremental validity to the VRAG-R, but that the VRAG-R would not add incremental validity to the HCR-20\textsuperscript{V3}.

The VRAG-R Total score was entered at Block 1, and the HCR-20\textsuperscript{V3} Total score at Block 2. At stage 1, the VRAG-R Total scores produced a significant model, $\chi^2(1) = 6.513$, $p < .05$. The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the VRAG-R Total scores was 19.2\% (Nagelkerke R\textsuperscript{2}). At this stage, 59.5\% of cases were correctly classified.

When the HCR-20\textsuperscript{V3} Total scores were added to the model at Block 2, the model remained significant, $\chi^2(2) = 16.293$, $p < .001$, and the contribution of the HCR-20\textsuperscript{V3} Total scores was significant, $\chi^2(1) = 9.780$, $p < .01$. The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the VRAG-R and HCR-20\textsuperscript{V3} Total scores in combination was 43\% (Nagelkerke R\textsuperscript{2}). At this stage, 69\% of cases were correctly classified. In the final model, only the HCR-20\textsuperscript{V3} Total scores emerged as a significant predictor (Wald = 5.873, $p < .05$). The VRAG-R was not significant (Wald = 0.023, $p = .880$). Therefore, the contribution of the HCR-20\textsuperscript{V3} Total scores was significant, and added incremental validity to the VRAG-R Total scores in the prediction of violent recidivism.
The analysis was run again with the order of entry reversed. The HCR-20\textsuperscript{V3} Total score was entered at Block 1, and the VRAG-R Total score at Block 2. At stage 1, the HCR-20\textsuperscript{V3} Total scores produced a significant model, $\chi^2(1) = 16.271, p < .001$. The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the VRAG-R Total scores was 43\% (Nagelkerke $R^2$). At this stage, 69\% of cases were correctly classified.

When the VRAG-R Total scores were added to the model at Block 2, the model remained significant, $\chi^2(2) = 16.293, p < .001$, but the contribution of the VRAG-R Total scores was non-significant, $\chi^2(1) = 0.023, p = .880$. The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the VRAG-R and HCR-20\textsuperscript{V3} Total scores remained at 43\% (Nagelkerke $R^2$), as did the percentage of cases were correctly classified (69\%). In the final model, only the HCR-20\textsuperscript{V3} Total scores emerged as a significant predictor (Wald = 5.873, $p < .05$). The VRAG-R was not significant (Wald = 0.023, $p = .880$). Therefore, the contribution of the VRAG-R Total scores was non-significant, and failed to add incremental validity to the HCR-20\textsuperscript{V3} Total scores in the prediction of violent recidivism.

When the analysis was run again using HCR-20\textsuperscript{V3} SRR’s as opposed to Total scores, similar results emerged. The VRAG-R Total score was entered at Block 1, and the HCR-20\textsuperscript{V3} SRR’s at Block 2. At stage 1, the VRAG-R Total scores produced a significant model, $\chi^2(1) = 6.513, p < .05$. The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by
the VRAG-R Total scores was 19.2% (Nagelkerke R^2). At this stage, 59.5% of cases were correctly classified.

When the HCR-20^{V3} SRR’s were added to the model at Block 2, the model remained significant, \( \chi^2(2) = 34.326, p < .001 \), and the contribution of the HCR-20^{V3} SRR’s was significant, \( \chi^2(1) = 27.812, p < .001 \). The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the VRAG-R Total scores and HCR-20^{V3} SRR’s in combination was 74.7% (Nagelkerke R^2). At this stage, 90.5% of cases were correctly classified. In the final model, only the HCR-20^{V3} SRR’s emerged as a significant predictor (Wald = 12.719, \( p < .001 \)). The VRAG-R was not significant (Wald = 1.008, \( p = .315 \)). Therefore, the contribution of the HCR-20^{V3} SRR’s was significant, and added incremental validity to the VRAG-R Total scores in the prediction of violent recidivism.

When the analysis was computed with reverse order of entry (i.e., HCR-20^{V3} SRR’s in Block 1 and VRAG-R Total scores in Block 2), the HCR-20^{V3} SRR’s produced a significant model, \( \chi^2(1) = 33.260, p < .001 \). The Hosmer and Lemeshow test was also non-significant, indicating that the model was not a poor fit. The percentage of variance in the DV explained by the HCR-20^{V3} SRR’s was 73.2% (Nagelkerke R^2). At this stage, 90.5% of cases were correctly classified.

When the VRAG-R Total scores were added to the model at Block 2, the model remained significant, \( \chi^2(2) = 34.326, p < .001 \), but the contribution of the VRAG-R Total scores was non-significant, \( \chi^2(1) = 1.066, p = .302 \). The Hosmer and Lemeshow test was also non-significant, indicating that the
model was not a poor fit. The percentage of variance in the DV explained by the HCR-20$^{V3}$ SRR’s and VRAG-R Total scores in combination remained at 74.7% (Nagelkerke $R^2$), as did the percentage of cases were correctly classified (90.5%). In the final model, only the HCR-20$^{V3}$ SRR’s emerged as a significant predictor (Wald = 12.719, p < .001). The VRAG-R was not significant (Wald = 1.008, p = .315). Therefore, the contribution of the VRAG-R Total scores was non-significant, and did not add incremental validity to the HCR-20$^{V3}$ SRR’s in the prediction of violent recidivism.

6.8 Hypothesis 8: Inter-Rater Reliability

It was hypothesised that the HCR-20$^{V3}$ and the VRAG-R would demonstrate acceptable inter-rater reliability (ICC > .40 as per Fleiss, 1981). Table 41 displays both the single-measure (ICC$_1$) and average-measure (ICC$_2$) coefficients for the HCR-20$^{V3}$ (Total score, scale score, SRR and Relevance ratings) and VRAG-R Total score. Based on Fleiss’ (1981) categorisations, the inter-rater reliability for the HCR-20$^{V3}$ Total score was excellent (ICC = 0.82), whereas SRR was good (ICC = 0.68). The H-Scale was excellent (ICC = 0.83), C-Scale good (ICC = 0.68), and R-Scale moderate (ICC = 0.52). Relevance ratings was moderate (ICC = 0.56) but produced the lowest ICC$_1$ index. The VRAG-R Total score had excellent inter-rater reliability (ICC = .80).
Table 41

*Inter-rater Reliability of the HCR-20\textsuperscript{V3} and VRAG-R.*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Inter-Rater Reliability</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICC\textsubscript{1}</td>
<td>ICC\textsubscript{2}</td>
</tr>
<tr>
<td>HCR-20\textsuperscript{V3} Total Score</td>
<td>.819</td>
<td>.901</td>
</tr>
<tr>
<td>H-Scale</td>
<td>.826</td>
<td>.905</td>
</tr>
<tr>
<td>C-Scale</td>
<td>.683</td>
<td>.811</td>
</tr>
<tr>
<td>R-Scale</td>
<td>.516</td>
<td>.681</td>
</tr>
<tr>
<td>SRR</td>
<td>.675</td>
<td>.806</td>
</tr>
<tr>
<td>Relevance Ratings</td>
<td>.561</td>
<td>.718</td>
</tr>
<tr>
<td>VRAG-R Total Score</td>
<td>.800</td>
<td>.889</td>
</tr>
</tbody>
</table>

*Note.* HCR-20\textsuperscript{V3} = Historical Clinical Risk Management – 20 Version 3.
VRAG-R = Violence Risk Appraisal Guide. HCR-20\textsuperscript{V3} N = 10. VRAG-R N = 5.

ICC\textsubscript{1} = *single measure intraclass correlation coefficient.* ICC\textsubscript{2} = *average-measure intraclass correlation coefficient.* 95% CI = 95% Confidence Interval for single-measures. *Total scores analysis based on prorated scores for both HCR-20\textsuperscript{V3} and VRAG-R.*
7.1 Overview

The HCR-20\textsuperscript{V3} and VRAG-R are the most commonly used violence risk assessment tools within the actuarial and SPJ approaches, respectively. To our knowledge, this study represents the first and only evaluation of the HCR-20\textsuperscript{V3} and VRAG-R in Australia. Using a pseudo-prospective design, 100 Australian forensic psychiatric patients were followed-up for a maximum of 12 years and 10 months post-discharge from the Thomas Embling Hospital.

The mean time at risk for general offending was 1442.93 days ($Md = 783$, $SD = 1514.83$), and 1935.19 days ($Md = 1397.50$, $SD = 1622.03$) for violent offending. Outcome data for both general and violent recidivism were sourced from Victoria Police records. The base rate of offending during follow-up was 63% for general recidivism and 50% for violent recidivism.

The aim of the study was to: 1) evaluate the predictive validity of the HCR-20\textsuperscript{V3} and the VRAG-R; 2) compare the predictive validity of these tools to their predecessors (i.e., the HCR-20\textsuperscript{V2} and the VRAG); and 3) to compare the predictive validity of the HCR-20\textsuperscript{V3} and VRAG-R. Additionally, the concurrent validity and inter-rater reliability of the HCR-20\textsuperscript{V3} and VRAG-R was explored, as was incremental validity relevant to the PCL-R.

Eight distinct hypotheses were put forward. The findings pertaining to each hypothesis are discussed below, followed by a consideration of the research limitations and clinical implications. The chapter concludes with recommendations for future research.
7.2 Scoring Description

Of the 100 participants rated on the HCR-20\textsuperscript{V3}, 16 were rated Low risk, 38 were rated Moderate risk, and 46 were rated High risk. Results revealed that there was an incremental increase in rates of general offending (12.50%, 52.63%, and 89.13%) and violent offending (0%, 42.22%, and 73.91%) across participants found to be at Low, Moderate or High levels of risk respectively.

A similar pattern was observed for the VRAG-R when Bins were pooled into categories of Low, Moderate and High. Of the 42 participants, two were rated Low risk, 19 were rated Moderate risk, and 21 were rated High risk. Rates of general offending (0%, 47.37%, and 71.43%) and violent offending (0%, 36.84%, and 57.14%) increased incrementally across the risk categories.

Interestingly, the percentage of participant’s rated as having a high risk of violence on both tools closely matched the base rates of violent recidivism (50%). On the HCR-20\textsuperscript{V3}, 46% of participants were rated as High risk for violence, and on the VRAG-R, 57.14% of participants were rated as falling within the top three risk Bins.

7.3 Main Findings
7.3.1 Associations with Violence

It was hypothesised that there would be an association between violent recidivism and scores on the HRC-20\textsuperscript{V3} and VRAG-R. Specifically, it was predicted that the HCR-20\textsuperscript{V3} Total scores and SRR’s would correlate significantly with violent recidivism, as would the VRAG-R’s Total scores and
Bins. The association between scores on violence risk assessment tools and frequency of violent offence charges post-release was also explored.

This hypothesis was partially supported. Results indicated that violent recidivism was significantly correlated with the HCR-20\textsuperscript{V3} Total scores, HCR-20\textsuperscript{V3} SRR’s, VRAG-R Total scores and VRAG-R Bins. All correlations were in the positive direction, with a fair to moderately strong relationship.

For nominal outcomes (i.e., HCR-20\textsuperscript{V3} SRR’s and VRAG-R Bins), results showed a significant association between the HCR-20\textsuperscript{V3} SRR’s and violent recidivism, with a large effect. For those rated High, the odds of being charged with a violent offence post-release was 3.9 times higher than if they had been rated as having a Moderate risk of violence.

The association between the VRAG-R Bins and violent recidivism was large but not statistically significant. Pooling of the Bins to increase frequencies did not improve outcomes, and the association between VRAG-R Bins and violent recidivism remained non-significant. Even with pooling of the data, the meaningfulness of this analysis was limited by the number of cases falling into VRAG-R Bins (i.e., small sample size spread across many categories), resulting in low power. Therefore, this result should be interpreted with caution.

The association between scores on violence risk assessment tools and the frequency of violent offence charges was also explored. The mean number of violent offences committed across the follow-up period was 2.69. Regarding the relationship between scores and frequency of violent charges post-release, significant positive correlations were found for both the HCR-20\textsuperscript{V3} Total score, and the VRAG-R Total score. The strength of association
was fair for both tools, indicating that the tools hold value not only for the presence of offending, but also the frequency of offences.

7.3.2 Concurrent Validity

It was hypothesised that the previous and current versions of the HCR-20 and VRAG would demonstrate acceptable concurrent validity (i.e., significant correlations, \( r > .50 \)), as would comparisons of the HCR-20\(^{V3}\) and VRAG-R to each other. It was also hypothesised that static measures (i.e., the HCR-20\(^{V3}\) H-Scale and VRAG-R) would correlate significantly, as would scales of the HCR-20\(^{V3}\) and HCR-20\(^{V2}\).

A matrix of correlations between the HCR-20\(^{V2}\), HCR-20\(^{V3}\), VRAG and VRAG-R were all significant and in the positive direction. The strength of the correlation between the HCR-20\(^{V2}\) and HCR-20\(^{V3}\) Total scores was moderate to good, as was the strength of the correlation between the VRAG and VRAG-R Total scores. The correlation between the VRAG and VRAG-R was the strongest correlation observed. This may be an indication of the static nature of these tools, and perhaps the magnitude of changes across versions (i.e., the HCR-20\(^{V3}\) saw changes to content and process, whereas the VRAG Items 1 – 7 remained relatively unchanged).

The correlation between the HCR-20\(^{V3}\) and VRAG-R was fair. Regarding static measures, the HCR-20\(^{V3}\)’s H-Scale was significantly correlated with the VRAG-R, with a positive, moderate-to-good relationship. These correlations were produced using Kendall’s Tau statistic as a precautionary measure due to variations in normality of the data at the univariate level. When correlations were produced using Pearson’s
correlation, the strength of associations improved from moderate-to-good to a clear good relationship. All correlations were positive and significant. The strength of relationship between the HCR-20\textsuperscript{V2} and HCR-20\textsuperscript{V3} scales were all moderate to good. The strongest correlation was observed for the H-Scale, followed by the C-Scale and then the R-Scale. This is not surprising given the static nature of the H-scale.

7.3.3 Predictive Validity

It was hypothesised that the predictive validity of the HCR-20\textsuperscript{V3} and the VRAG-R for violence would be greater than chance. The HCR-20\textsuperscript{V3} Total scores produced an AUC of 0.70, a significant moderate effect size. The HCR-20\textsuperscript{V3} SRR’s produced an AUC of 0.77, a significant and large effect size. This is similar to previous research identifying SRR’s as a robust predictor, and indicating the superiority of SRR’s over Total scores (e.g., Douglas, Ogloff, & Hart, 2003; Guy, 2008). The predictive validity of the HCR-20\textsuperscript{V3}’s Relevance ratings was also explored. The Relevance rating Total score produced an AUC of 0.71. This was significant with a large effect size. It is interesting that the Relevance ratings produced a larger AUC than the Presence ratings, albeit similar and not significantly different. This may be a reflection of the entangling of the two measures, and capturing of similar content under both. Although further research is needed to understand the novel role of the Relevance ratings more clearly.

The indices of discrimination indicated that at a Moderate SRR, sensitivity = 1, specificity = 0.32, PPV = 0.60, NPV = 1. This means that the HCR-20\textsuperscript{V3} was able to correctly identify the presence of violence 100% of the time; and correctly identify the absence of violence 32% of the time. All
participants predicted to be violent were violent, while only 32% of participants predicted to be non-violent were in fact non-violent (i.e., 68% of non-violent participants were incorrectly predicted to be violent). These are encouraging results for the identification of violence, but also indicate that participants were being rated as posing a High risk of violence too often. Indeed, just under half of all participants were rated as High risk. However, erring on the side of caution, a reasonable balance was achieved between identifying the presence and absence of violence. At the Moderate SRR cut-off, the probability of participants with a positive test being violent was 60% and the probability of participants with a negative test not being violent was 100%.

At a High SRR, sensitivity = 0.68, specificity = 0.32, PPV = 0.74, NPV = 0.70. This means that, at the High SRR, the HCR-20\textsuperscript{v3} was able to correctly identify the presence of violence 68% of the time; and correctly identify the absence of violence 76% of the time. At the High SRR cut-off, the probability of participants with a positive test being violent was 74%; and the probability of participants with a negative test not being violent was 70%. No indices of discrimination were reported for the Low SRR as all cases are test-positive at this level.

The VRAG-R also significantly predicted violent recidivism (AUC = 0.71), with a large effect size. This was identical to the AUC produced by the VRAG-R Bins. This is understandable given that the Bin categories are derived from the Total scores. The discriminant capabilities of the VRAG-R at each risk Bin is hampered by low power relevant to the small number of cases falling within each risk Bin. Therefore the following interpretation is provided with caution. The VRAG-R appeared to work best at identifying all
violent persons and minimising incorrectly classifying non-violent persons in Bin 5, where it correctly identified the presence of violence 100% of the time; and correctly identified the absence of violence 34% of the time (PPV = 55%, NPV = 100%). In Bin 6, a compromise between identifying violent persons but not incorrectly classifying non-violent persons was achieved, where the VRAG-R was able to correctly identify the presence of violence 84% of the time, and correctly identify the absence of violence 47% of the time (PPV = 57%, NPV = 78%). Prediction at Bin Number 7 was probably the most balanced around the 50% sensitivity/specificity point. The VRAG-R was able to correctly identify the presence of violence 63% of the time; and correctly identify the absence of violence 60% of the time (PPV = 57%, NPV = 66%).

In comparing the predictive performance of the HCR-20\textsuperscript{V3} and VRAG-R, the largest AUC was observed for the HCR-20\textsuperscript{V3} SRR, followed by HCR-20\textsuperscript{V3} Relevance Ratings and VRAG-R Total score with equal AUC’s, and finally the HCR-20\textsuperscript{V3} Total score. However, all AUC’s were within a similar range (0.70 – 0.77), suggesting comparable outcomes in structured professional judgment and actuarial approaches to assessment. This is consistent with previous research by Guy (2008). This range of AUC’s is also similar to what has been observed in previous research utilising forensic psychiatric samples, both with the HCR-20\textsuperscript{V2} (Dolan et al., 2010; Gray et al., 2008; Snowden et al., 2010; Strand et al., 1999), the HCR-20\textsuperscript{V3} (de Vogel et al., 2014; Doyle et al., 2014; Strub et al., 2014), the VRAG (Doyle et al., 2002; Snowden et al., 2010; Thomson et al., 2008; Quinsey et al., 2006) and the VRAG-R (Rice et al., 2013).

In comparing current to previous iterations of the measures, it was also hypothesised that the levels of accuracy produced by the HCR-20\textsuperscript{V3} and
VRAG-R would be similar to levels of accuracy demonstrated by the HCR-20\textsuperscript{V2} and VRAG, respectively. For the HCR-20\textsuperscript{V3}, Total scores were used as the basis of comparison because no SRR’s were available in the HCR-20\textsuperscript{V2} datasets. The HCR-20\textsuperscript{V2} produced an AUC of 0.77, whereas the HCR-20\textsuperscript{V3} produced an AUC of 0.69. Both AUC’s were significant. A comparison of AUC’s indicated that the areas were not significantly different.

The VRAG Total score produced an AUC of 0.73, whereas the VRAG-R total score produced an AUC of 0.72. Similarly, the VRAG Bins produced an AUC 0.74, whereas the VRAG-R Bins produced an AUC of 0.72. Again, comparisons indicated that these differences were not significantly different.

### 7.3.4 HCR-20\textsuperscript{V3} Predictive Strength of Scales

It was hypothesised that the H-Scale would be the strongest predictor of post-release violence, however, when taking imminence of violence into account, the C-Scale and R-Scale would emerge as the strongest predictors of post-release violence.

Results indicated that the HCR-20\textsuperscript{V3} scales all significantly predicted violent recidivism. The strongest predictor was the C-Scale (AUC = 0.69), followed by the H-Scale (AUC = 0.65), and then the R-Scale (AUC = 0.62). A sequential binomial logistic regression revealed that, of the three scales, only the C-Scale significantly predicted post-release violent offence charges. It also emerged as the strongest predictor for general recidivism. When controlling for time at risk, the C-Scale remained the only significant predictor of violent recidivism. Consideration of the interactions between scales showed that the C-Scale and H-Scale x R-Scale interaction were the
strongest predictors, however the significance of the C-Scale discontinued when the HCR-20$^V3$ Total score was added to the analyses.

The finding of the C-Scale as the strongest predictor is an atypical finding within the literature as a whole. That is not to say that that superiority of the C-Scale is an absolute anomaly. This outcome has been observed in several other studies for the HCR-20$^V2$ in forensic psychiatric populations (e.g., Dolan et al., 2010; Douglas et al., 2003) and for the HCR-20$^V3$ (e.g., Doyle et al., 2014).

The superiority of the C-Scale could be related to the homogenous and high-risk nature of the sample. Almost half (46%) of the sample were assessed as High risk on the HCR-20$^V3$, and 54% had a HCR-20$^V3$ Total score greater or equal to 30. The H-Scale may have failed to emerge as a significant predictor due to less variability in the H-Scale, and greater variability in the C-Scale. It is considered that the superiority of the C-Scale may be associated with the psychiatric nature of the sample, particularly the level of psychopathology observed (i.e., 95% of the sample received a diagnosis of major mental illness upon discharge from TEH, with schizophrenia spectrum and other psychotic disorders as the primary diagnosis for 72%). The acuity of the sample is also a factor that may have played a role in the superiority of the C-Scale. Ninety-percent of the sample was discharged from acute and sub-acute units within TEH, which may speak to the level of acuity of their mental illnesses, and the importance of current active symptoms of mental illness (Douglas et al., 1999). It is also erroneous to assume that presumed stability at discharge is reflective of ongoing mental state post-discharge (Douglas et al., 1999). Certainly, mental illness is
dynamic in nature, and the C-Scale may correspond more closely to changes over time.

The superiority of the C-Scale may also be associated with the average time at risk. Survival curves showed a stark drop-off in survival in the early phases of follow-up. The H-Scale may have emerged as a stronger predictor if offending were more staggered across this longitudinal follow-up period, as historical factors are seen to contribute to a long-term risk state.

Notably, all three scales demonstrated predictive validity, with AUC’s falling within very similar ranges (0.62 – 0.69). Failure of the H-Scale to emerge as the strongest predictor does not discount that it was predictive, but is possibly a reflection of the high-risk, high acuity, psychiatric nature of the sample.

7.3.5 Discrimination

It was hypothesised that participants scoring above the median on the HCR-20\textsuperscript{V3} and VRAG-R would be significantly more likely than those scoring below the median to recidivate both generally and violently. For the HCR-20\textsuperscript{V3}, the odds of being charged with a criminal offence during the follow-up period was 3.89 times higher if the participant scored above the median as opposed to below. The odds of being charged with a violent offence during the follow-up period was 2.89 times higher if the participant scored above the median as opposed to below. For the VRAG-R, the odds of being charged with a criminal offence during the follow-up period was 3.30 times higher if the participant scored above the median as opposed to below, whilst the odds of being charged with a violent offence during the follow-up period was 2.60 times higher if the participant scored above the median as opposed to below.
Therefore, increases in the odds of violence was similar for both the HCR-20\textsuperscript{V3} and VRAG-R, across both general and violent recidivism. The odds observed are similar to odds reported in other studies (e.g., Doyle et al., 2012), although much higher odds have been seen within the literature (e.g., Douglas et al., 1999). Notably, the odds of being charged with a general offence was higher for both tools than the odds of being changed with a violent offence, which speaks to the utility of these tools for predicting recidivism more broadly. These findings support existing research indicating that higher scores are associated with higher probabilities of violence.

### 7.3.6 Survival Time

It was hypothesised that there would be statistically significant differences in survival time related to HCR-20\textsuperscript{V3} and VRAG-R scores (Total scores divided using a median split) and risk categories (HCR-20\textsuperscript{V3} SRR’s and constructed VRAG-R categories). For the HCR-20\textsuperscript{V3}, differences in survival times were clearly demonstrated in the diverging survival curves across the follow-up period. The ‘drop-off’ for High-risk persons was clearly greater in the earlier follow-up periods, more staggered for Moderate risk persons, and absent for Low risk persons. This illustrates the offending trajectories (first offence) that Low, Moderate and High risk participants followed post-release. The divergence between survival curves is encouraging and provides good evidence to support Total scores and SRR’s, not only for the presence of offending, but also the timing (i.e., High risk participants ‘dropped off’ early on in the follow-up period). Survival analyses demonstrated that participants deemed High risk were more likely to be violent, and to perpetrate violence sooner than Moderate or Low risk persons.
For general recidivism, the HCR-20\(^{V3}\)’s Total scores (based on groups of those scoring below and above the median) differed significantly in their survival distributions as did the HCR-20\(^{V3}\) SRR’s. This means that participants who scored above or below the mean, and participants rated Low, Moderate or High differed significantly in how long they ‘survived’ in the follow-up period, before being charged for an offence. As expected, the greatest distinction was between Low and High categories, but the Low category did not differ significantly from the Moderate category. The Moderate category differed significantly from the High category. This is not surprising given that the sample was considered a ‘high-risk’ group, with a median HCR-20\(^{V3}\) score of 30. The High category (and participants falling above the median) was clearly quite distinct from the two lower groups.

For violent recidivism, groups again differed in survival times. For HCR-20\(^{V3}\) Total scores, groups scoring below or above the median differed significantly in survival functions. This time however, there were significant differences between all SRR’s. Low differed significantly from Moderate and High groups, and Moderate differed significantly from High. This means that survival times based on SRR’s were more distinct for violent recidivism than general recidivism across the three groups.

Visually, the VRAG-R survival curves also showed good divergence in those scoring below and above the median, and for grouped risk categories of Low, Moderate, and High. However, statistical analyses revealed that persons scoring below and above the median differed significantly in survival time for general recidivism, but not for violent recidivism. This is in contrast to what was observed for the HCR-20\(^{V3}\) scores, where significant differences were detected for both offending types. This may stem from the VRAG-R
median score falling within the mid-range of all possible scores (i.e., $Md = 11.50$, in a possible range of -34 to +46), and participants being more scattered across the range of scores. Whereas the HCR-20$^\text{V3}$ median score fell in a much higher range ($Md = 30$, in a range of 0 to 40), with most participants falling above this score. The gradual increase in participants across the grouped VRAG-R Bins is likely to be related to why the grouped Bins did not differ significantly in general nor violent recidivism, as there was less of a stark contrast in the proportion of participants across groups, as observed in the HCR-20$^\text{V3}$ SRR’s.

Some caution is advised in interpreting the results of the Kaplan-Meier analyses as the presence of censoring was dissimilar in some cases, which can lead to incorrect conclusions about differences between groups (Hosmer et al., 2008). It is noted that the small sample size in group categories and lack of similarity in censoring may have contributed to a Type I or Type II error in finding the survival distributions significantly different overall, but not amongst the discrete categories.

7.3.7 Incremental Validity

7.3.7.1 Within Tool Incremental Validity

It was hypothesised that the HCR-20$^\text{V3}$’s Relevance ratings would demonstrate incremental validity over the Presence ratings. This hypothesis was not supported, as results indicated that the Relevance ratings did not improve upon prediction, and therefore, did not demonstrate incremental validity over the Presence ratings. This is likely because the two variables are capturing similar information, becoming redundant when considered simultaneously. The Relevance ratings were significant predictors in isolation.
Further analysis using both forward and backward stepwise logistic regression showed that Relevance ratings emerged as the most significant predictor. However, these results should be interpreted with caution because stepwise regression is still considered a controversial method due to its reliance on statistically-based decision-making, however, it was still included in the current research as an additional analysis because of the exploratory nature of this hypothesis (Field, 2009).

Inspection of the individual ratings on the 20 risk factors demonstrated significant correlations for all risk factors, and ratings often matched (i.e., participants rated ‘Yes’ on Presence were also rated high on Relevance; and participants rated ‘No’ on Presence were rated low on Relevance). This was the case for the vast majority of risk factors. For example, the percentage of participants rated ‘Yes’ for Presence ratings who also received a High Relevance reached as high as 85% (H1 Violence). Without fail, Participants rated ‘No’ for Presence always had the highest frequency of participants in the Low Relevance rating category. In fact, for all risk factors, participants with a ‘No’ Presence rating always had a Low Relevance rating. Indeed, ROC curves for the two variables produced identical AUC’s. This suggests that the Presence and Relevance ratings are capturing much of the same information, which leads to redundancy when both are analysed simultaneously.

Alternatively, it may also speak to biases amongst raters that a risk factor which is definitely present is then highly relevant, or that a risk factor which is absent cannot be relevant. The HCR-20\textsuperscript{V3} manual advises that absent risk factors may still be considered relevant (Douglas et al., 2013), yet inspection of individual ratings seemed to suggest that this was not the case in the current study.
7.3.7.2 Between Tool Incremental validity

It was hypothesised that the PCL-R would not add significant incremental validity to the HCR-20$^{V3}$ and VRAG-R. Three comparisons were made: PCL-R and HCR-20$^{V3}$ Total scores; PCL-R and HCR-20$^{V3}$ SRR’s; and PCL-R and VRAG-R Total scores. The hypothesis was not supported for the HCR-20 Total scores and SRR’s, as the PCL-R added significant incremental validity to these measures; however, the hypothesis was supported for the VRAG-R, as the PCL-R did not add significant incremental validity. The HCR-20$^{V3}$ also added incremental validity to the VRAG-R, but the VRAG-R did not add incrementally to the HCR-20$^{V3}$. This may be attributed to the static and dynamic nature of the HCR-20$^{V3}$, and the breadth of risk factors included.

These results need to be interpreted with caution for several reasons. Firstly, regarding the HCR-20$^{V3}$, because psychopathy is captured jointly with other personality disorders (i.e., Anti-social Personality Disorder and Dissocial Personality Disorder, amongst others) under H7 on the HCR-20$^{V3}$, psychopathy diagnoses could not be partitioned out and excluded from the analysis. This creates a problematic conflation issue. Presence ratings that informed the HCR-20$^{V3}$ Total scores (and ultimately, SRR’s) were a reflection of a range of all possible personality disorders, not limited to psychopathy. Given that the PCL-R captures a very broad range of personality traits, behaviours and affective dispositions which differ from features of Anti-social Personality Disorder and Dissocial Personality Disorder, it is reasonable that the PCL-R could add significantly to the model. Secondly, it is noted that the VRAG-R sample size ($N = 42$) is bordering on the bare minimum requirement of 15 cases per independent variable for logistic analyses (Leard Statistics, 2015).
7.3.8 Inter-rater Reliability

It was hypothesised that the HCR-20\textsuperscript{V3} and the VRAG-R would demonstrate acceptable inter-rater reliability. Results indicated that inter-rater reliability for the HCR-20\textsuperscript{V3} Total score was excellent, whereas agreement for SRR’s was good. This is consistent with previous research on the HCR-20\textsuperscript{V3} indicating that agreement is greater for Total scores than SRR’s (e.g., de Vogel et al., 2014; Douglas et al., 2014). Mathematically, it makes sense that SRR’s would yield lower levels of agreement than Total scores because SRR’s are a gross single-item rating, whereas scores are compared on a multi-item scale (Otto et al., 2010).

Regarding the HCR-20 scales, agreement for the H-Scale was excellent; the C-Scale was good and the R-Scale was moderate. The best agreement was observed for the H-Scale, whereas the least agreement was for the R-Scale. The finding of greater levels of agreement for the H-Scale on the HCR-20\textsuperscript{V3} has been found in other recent studies (e.g., Howe et al., 2014; Smith et al., 2014). High levels of agreement for the H-Scale is not surprising given the historical static nature of the scale. This is contrasted with the prospective nature of the R-Scale in forecasting future living and contextual factors. It is reasonable that agreement on past and current functioning may be greater than agreement on future functioning. Despite this, other recent HCR-20\textsuperscript{V3} research that has found superior levels of agreement for the R-Scale (e.g., Doyle et al., 2014; Kötter et al., 2014).

Moderate agreement was found for the Relevance ratings. This is promising given that the Relevance ratings are a new feature of the HCR-20\textsuperscript{V3}. The VRAG-R Total score had excellent inter-rater reliability, which is not surprising given the static nature of the tool. These are all single rater
outcomes, as tools are generally used in this way within an Australian context (i.e., mental health professionals generally do not complete violence risk assessment tools using averaged or consensus ratings).

7.4 Findings Overview and Clinical Implications

This study adds a noteworthy contribution to the limited body of research on the HCR-20\textsuperscript{V3} and VRAG-R. Outcomes for a range of basic psychometric properties were reported, including validity (predictive, concurrent and incremental) and inter-rater reliability. The HCR-20 and VRAG boast a large international evidence base in their ability to predict future violence. However, limited research has stemmed from Australian cohorts, despite the widespread use of the measures in Australia. Therefore, the current research sought to make novel contributions to the international evidence base more broadly, but particularly within Australia given the dearth of research available.

Overall, findings for the performance of the HCR-20\textsuperscript{V3} and VRAG-R in predicting both general and violent recidivism were promising and support the continued use of these tools in clinical practice. Both tools are significantly associated with violence, and were able to significantly predict violent recidivism, within AUC ranges of 0.70 – 0.77. These are moderate to large effects. Predictive accuracy is not significantly different to that of the HCR-20\textsuperscript{V2} or VRAG, although there was a discrepancy in observed AUC’s (i.e., HCR-20\textsuperscript{V2} AUC = 0.77, HCR-20\textsuperscript{V3} AUC = 0.69; VRAG AUC = 0.73, VRAG-R AUC = 0.72), suggesting differences in the discriminant ability of tools. Of course, such gross measures of performance cannot account for other advantages, such as improvements to risk management and intervention
planning. The best outcomes were observed for SRR’s on the HCR-20\textsuperscript{V3}, which supports the SPJ approach to violence risk assessment, however comparable outcomes were observed for both the SPJ and actuarial approach. Specifically, these findings suggest comparable performance of these two measures, and support the inclusion of clinical judgment in the violence risk assessment process, over and above actuarial scores, as suggested previously by Guy (2008).

The HCR-20\textsuperscript{V3} seems to have achieved its revision goal of “continuity of concept” (Douglas et al., 2013, p.18, see section 2.2.3 for a detailed explanation). Existing conceptualisations of risk assessment were maintained across iterations, where users of the HCR-20\textsuperscript{V2} would be able to learn and adapt reasonably quickly to the HCR-20\textsuperscript{V3}, due to a consistent framework (i.e., 20 risk factors spread diachronically across three scales, rated on a three-level system and resulting in a gross risk judgment). In doing so, the conceptual framework of the tools are similar and evaluators do not need to adopt a new approach to assessing risk (Douglas et al., 2014). Versions 2 and 3 are significantly correlated overall (Total score) and at scale levels. Given its static nature, it was unsurprising that the strongest correlation amongst HCR-20\textsuperscript{V3} scales was observed for the H-Scale. The most surprising finding was probably the superiority of the C-Scale in prediction of both general recidivism and violent recidivism. However, this is not an entirely unheard-of finding, and others have reported similar results (e.g., Dolan et al., 2010; Douglas et al., 2003; Doyle et al., 2014). The incremental validity hypotheses revealed interesting outcomes regarding the combination of tools in boosting predictive accuracy. The incremental validity added to the HCR-20\textsuperscript{V3} by the PCL-R suggested that there may be predictive value in continuing
to aid HCR-20\textsuperscript{V3} risk assessments with the PCL-R. Furthermore, the HCR-20\textsuperscript{V3} added incremental validity to the VRAG-R, which may possibly indicate that dynamic risk factors add to static factors in increasing predictive validity. Given that the AUC’s ranges are similar for both tools, advocacy for using both tools or the use of one tool over the other is not encouraged at this stage – more research is required. Choice for one tool over the other is also context dependent, relating to the training, resources, setting and purpose of the assessment.

The failure of the Relevance ratings to add incremental validity to Presence ratings is not surprising given they captured almost identical levels of variance and, in effect, cancelled each other out. Closer inspection at the Item-level showed that, across the 20 risk factors, Presence and Relevance ratings were often matched on ratings (i.e., majority of participants rated ‘yes’ on Presence were also rated ‘high’ on Relevance; and all participants rated ‘no’ on Presence were rated ‘low’ on Relevance). Indeed, previous research has suggested similar levels of predictive accuracy between the HCR-20\textsuperscript{V3}’s Presence and Relevance ratings, and patterns in the Presence and Relevance AUC’s across scales (Hogan et al., 2015). If Presence and Relevance ratings are indeed redundant, this outcome may have clinical implications, given the time and effort required to score the Relevance ratings. Given the limited number of studies available, more research is required. Despite Relevance ratings not improving upon predictions in this analysis, it should be noted that the success of the SRR in yielding the largest AUC is a reflection of the interactions between Presence ratings, Relevance ratings and formulations (Douglas et al., 2013). Furthermore, these analyses
do not evaluate the input of Relevance ratings in assisting with risk management planning.

The archival nature of the study may have also played a role in Relevance ratings not adding predictive validity. The relevance of risk factors as assessed through file review may differ from relevance as assessed by clinicians who work directly with patients. Such familiarity may bring a richer understanding of the relevance of risk factors to the person's perpetration of violence.

In general, future research is needed on the HCR-20V3 and VRAG, but particularly with regarding to novel elements such as the Relevance ratings, SRR’s (Future Violence/ Case Prioritisation, Serious Physical Harm and Imminence), and risk ‘scenarios’ (nature, severity, imminence, frequency/duration, likelihood) (see the HCR-20V3 Manual, Douglas et al., 2013, p. 61-65). Further research is needed to understand the novel role of Relevance ratings in the risk assessment and management process more clearly. This may need to draw on both quantitative and qualitative research in understanding tasks as complex and diverse as risk formulation.

Even if improvements in risk assessment stemming from the new tools and associated features are not reflected in the current research, this is not to say that they have no clinical benefits in risk formulation and management. It is a positive outcome that the HCR-20V3 has retained its ability to predict violent recidivism whilst adding to increased formulation. Likewise, the VRAG-R has retained predictive validity whilst providing broader applicability, a simpler administration process and less reliance on additional tools or diagnostic systems. Clinicians can also be reassured that in their use of
current as opposed to dated tools, they are practicing within the current evidence base and best practice guidelines.

The field of violence risk assessment has advanced significantly over the past five decades, since the era of unstructured clinical judgment. For society, while false negatives do still occur, risk assessments are more structured and transparent, resulting in observable stages in decision-making and increased accountability. As violence risk assessment tools are revised and evaluated, clinicians are able to access evidence-based tools that reflect current research and employ best practice approaches. While it is unlikely that violence risk assessment will ever reach perfect prediction, the upper-levels of accuracy have been observed in the body of research (i.e., AUC = .80) (e.g., Strand et al., 1999; Douglas et al., 1999; de Vogel et al., 2014). For patients, whilst false positives do still occur, the process of violence risk assessment has become far more tailored, fluid and transparent. Risk assessments, particularly SPJ assessments that consider dynamic risk factors, now hold value for assessment and risk management.

It is important to acknowledge that improvements in violence risk assessment do not only occur at a statistical level. Utility in clinical practice is also important. The HCR-20\textsuperscript{V3} has brought with it improved guidance, further refined idiosyncratic assessment and greater specification of risk management. Not only in the identification of the Presence and Relevance of risk factors and subsequent formulation, but also through consideration of other SRR’s such as Serious Physical Harm and Imminence of Violence, and associated scenario planning. Through these changes, the assessment process has become more systematic and detailed, which also promotes advantages for risk management in the long-term and in day-to-day clinical
care (Bjorkly et al., 2014). Although these elements of the HCR-20^V3^ have not been evaluated in the current research, they pose as key areas for future research.

In summary, these findings support the use of the HCR-20^V3^ and VRAG-R to predict violent recidivism in Australia within forensic psychiatric populations. This is consistent with other international literature on previous and current versions. This research will enable Australian mental health professionals using these tools within forensic psychiatric release-decision making contexts to do so with a degree of confidence in the tools’ applicability to Australian forensic psychiatric populations.

7.5 Limitations

There are a number of limitations in the study that should be considered when interpreting the findings. First, the study was limited by its pseudo-prospective design. Efforts to remain blind to outcomes were taken through the coding of predictors and scores prior to outcome data. On the positive side, however, the research adopted a ‘confidential inquiry approach’ (as recommended by Doyle et al., 2014), which increases the chances of capturing participants thought to be at greater risk of non-compliance, anti-sociality and violence, and therefore, less likely to consent (Doyle et al., 2014).

The reliance on file information means that there is a limit to the amount of information that can be extracted. In the case of the current study, an average of 8.75 files was reviewed per participant. The researcher found these files to be comprehensive in nature, as is reflected in the minimal amount of missing data. Despite this, the understanding of subjects and the
relevance of risk factors to their violence risk may have been limited by file-based data, and could have been enriched by the addition of interviews with the patients, staff or collaterals.

The study was also limited by a relatively small sample size. Hanley and McNeil (1982) have indicated that this sample size may be too small for significant differences between ROC curves to be detected. They provide recommendations around sample size requirements and suggest a sample size of 286 – 652 violent and non-violent participants (total participants being 572 – 1304) per ROC curve in the ranges of those described in this study (AUC = .70 - .77), to ensure that differences can be detected with 80% - 95% certainty. Whilst these guidelines are not binding, they do highlight the lack of statistical power in samples below this threshold, and the possibility of Type II error (Hanley et al., 1982).

The limited sample size also prevented analysis at the item-level and sub-item level, which may be an important undertaking in indicating which items attribute to the overall predictive validity observed. Previous research by Coid and colleagues (2009) has indicated that predictive validity may in fact be driven by a very limited number of predictors. For the VRAG, they found that only five of the 12 items were independently predictive (Psychopathy, Age at Index Offence, Non-violent Offence score, History of Alcohol Problems and Female Victim). For the HCR-20$^{V2}$, they found that only eight of the 20 items were independently predictive (Young Age at First Violent Incident, Substance Use Problems, Early Maladjustment, Prior Supervision Failure, Negative Attitudes, Impulsivity, Exposure to Destabilisers and Non-compliance with Remediation Attempts). Similar Item-level analyses on the VRAG-R and HCR-20$^{V3}$ may inform further refinements.
Furthermore, whilst HCR-20\textsuperscript{V3} and VRAG-R AUC’s were within similar ranges (0.70 – 0.77), it was impracticable to compare the HCR-20\textsuperscript{V3} and VRAG-R AUC’s using Hanley & McNeil’s (1983) method as the AUC’s were produced from different samples with varying sample sizes (i.e., HCR-20\textsuperscript{V3} $N = 100$; VRAG-R $N = 42$). This presented as a key limitation of the research, and future research could improve upon the current study by using the same sample and statistically comparing not only previous and current iterations, but also predictive accuracy between tools.

The gender composition of the sample also prevented reporting of gender specific outcomes. For the HCR-20\textsuperscript{V3} sample ($N = 100$), separate analyses by gender were not computed as the dataset consisted of only 27 females. Similarly, for the VRAG-R, a limited number of females ($N = 10$), prevented computation of separate statistical analyses for females. While several studies have began to shed light on gender differences regarding the HCR-20\textsuperscript{V3} (see Green et al., 2016; Strub, Douglas & Nicholls, 2016), the study could have been improved by employing a larger sample with representation from both males and females, enabling analysis and reporting of gender specific outcomes. This is necessary given the disparity in research informing management of females in forensic mental health and prison populations, as opposed to males (de Vogel & Nicholls, 2016). Therefore, a key focus for future research will be to provide evidence and knowledge to further inform the debate around gender differences for violence risk assessment and management.

This study only employed official police records, without reliance on self-report or collateral sources. It is probable that the base rate of violence recorded in this study represented an underestimate of the true level of
violence perpetrated by participants (see Doyle et al., 2012; Dolan et al., 2006; Douglas et al., 2003; Mulvey, et al., 1994). Additional sources of data to inform outcomes would have resulted in a greater detection of instances of violence and therefore increased base rates.

Other limitations related to both the primary researcher and inter-rater reliability research assistant not attending formal VRAG training. However, at the commencement of data collection, formal training was not available in Australia, neither were online webinar training opportunities. It is plausible that the number of raters may have also influenced the predictive accuracy indices observed. Both the HCR-20$^{V2}$ and VRAG represent the efforts of three individual raters (i.e., Campbell, Chu, and a research assistant from the Campbell project), whereas all HCR-20$^{V3}$ and VRAG-R ratings were made by the author. Therefore, there is less variability in scoring for the current iterations. This may affect the applicability of results to certain contexts where more than one individual completes violence risk assessment tools, such as in the case of multi-disciplinary teams or with averaged ratings. It is also noted that the author is in professional training in forensic psychology. Although she had used the HCR-20 in practice and attended the formal training, she was a novel user overall, particularly within a research context. However, the effects of this inexperience were buffered by professional training and guidance by senior research personnel. Regarding the inter-rater reliability methodology, two different raters completed the same versions of each tool. Raters varied greatly in experience, but it is encouraging that the inter-rater reliability was still good despite such differences.

Despite these limitations, the study fills a significant gap in the body of research on these newly developed tools, representing the first of its kind in
Australia. This represents a significant contribution, especially considering the widespread use of the HCR-20 in Australian forensic practice.

7.6 Future Research

Given these aforementioned limitations, a methodology for future research evaluating violence risk assessment tools for prediction of post-release community violence is discussed. An ideal methodology would employ a truly prospective design where researchers are able to remain blind to outcomes. Researchers should endeavor towards a confidential inquiry approach, with ratings made by an experienced evaluator who has access to formal training. Ratings should be made by several raters, to increase variability. Ratings should be made on the basis of several sources of information, including inpatient files, participant and staff interviews, and collateral information. However, it is noted that the confidential inquiry approach may be a hindrance to participant interviews. Preferably, outcomes should be based on convictions as opposed to charges, supplementing official records with self-report and collateral data as well, to increase the detected base rate. Future research may also wish to consider the use of fixed follow-up periods, which remains relatively uncommon. Immediate short-term risk (i.e., 1-7 days to a few weeks) and longitudinal follow-up periods to the degree used within this research (12 Years and 10 Months) is relatively rare within the literature. Most studies tend to fall within the middle-band of a few months to several years, with average times at risk reported as opposed to set incremental follow-up periods.

Should the study include comparisons of ROC curves, the sample size should reflect recommendations made by Hanley and McNeil (1982). It would
be advantageous to have representation of both males and females, with reporting of gender-specific predictive validity outcomes. Most past research has also stemmed from American and North American studies. Although studies from other countries have contributed to the growing evidence base, there is a need for greater demographic diversity, as well as higher proportions of females and CALD groups.

Given the recent release of the HCR-20\(^{V3}\) and VRAG-R, and the general lack of research exploring their psychometric properties, future research is required to consolidate the emerging evidence base pertaining to the psychometric properties of the HCR-20\(^{V3}\) and VRAG-R. Despite the abundance of research available on previous versions of these measures, the evidence-base for these newly updated tools is still in its infancy. Future research should evaluate a range of psychometric properties, such as the current research has attempted (e.g., predictive and concurrent validity, inter-rater reliability etc.), especially for novel aspects of the tool, such as the Relevance ratings, increased focus on formulation and different SRR’s (Case Prioritisation/ Future Violence, Serious Physical Harm and Imminence of Violence). A few initial studies have tapped into differences in three SRR’s (e.g., Hogan et al., 2016; Howe et al., 2016) however, replication is key. As mentioned previously, for research investigating Relevance ratings, the study could be improved upon if raters have some familiarity or knowledge of patients to inform decision-making around the relevance of risk factors. Further investigation of the understandings of Relevance ratings are also important. Strub and colleagues (2014) have posed the question of whether clinicians are using and defining Relevance in a consistent manner. Insight into clinicians’ understanding and interpretation of Relevance, and how they
are adopted in practice is required. Opinions of the author indicated that whilst Relevance ratings were useful in guiding and tailoring formulations of violence, greater guidance is required on understanding Relevance for some Items. For example, H1 History of Violence, how relevant is history of violence to future perpetration of violence? Some would argue that past violence is always relevant to future violence. This specific question has also been raised by others seeking greater clarification (see Douglas et al., 2014).

In addition to evaluating psychometric properties and novel aspects of the tools, little is known about the utility of these tools in daily practice, risk management activities. More research on the clinical use of these tools, and practical considerations such as user-friendliness, time requirements and labour-intensity would be valuable for clinical audiences. Doyle and Logan (2012) have noted that, unlike risk assessments geared towards establishing long-term risk and prevention strategies, “there are many circumstances where a closer link is required between risk assessment findings and risk management interventions; for example, where a practitioner in a prison or forensic mental health facility is required to engage with a client to understand and address specific relevant risk factors in order to manage risks on a day-to-day basis” (p. 408). Further research on the utility of these tools in such circumstances would be valuable.

7.7 Conclusion

This study explored validity and inter-rater reliability for the HCR-20\textsuperscript{V3} and VRAG-R within a sample of Australian forensic psychiatric patients. Results are promising, demonstrating psychometric properties that support the use of these tools within Australian forensic psychiatric practice. Violence
is a complex construct, and its prediction is unlikely to ever reach a point of absolute certainty (Mullen et al., 2009). Despite this, significant advancements have been made, and the evidence base now indicates predictive accuracies that are much higher than observed in unstructured clinical judgment approaches, and beyond what early proponents in the field envisioned possible.

But given the plateau of predictive indices seen in recent years, perhaps the field has reached the upper bounds of accuracy and the ceiling effect predicted earlier. Considerable improvement has been made over the past five decades, but consistent levels of reported outcomes indicates that the field may have reached the plateau in predictive accuracy (Davis, 2010). If statistical improvements are not seen (i.e., AUC’s are leveling-off and reaching the upper bounds of accuracy), advancements in clinical utility, applicability and user friendliness of tools have emerged. Contemporary violence risk assessment presents as a tailored practice, with descriptive outcomes far richer than the dichotomous predictions seen in decades prior. Regardless, the aim of violence risk assessment should be decision-making that is “conceptually, empirically and clinically sound” (Strub et al., 2014, p. 148). The current state of practice seems to reflect these three corner-stones.
References


*Corrections Act 1991 (VIC), s. 56, (Austl.).*


*Crimes (Mental Impairment and Unfitness to be Tried Act 1997) (VIC) (Austl.).*


http://dx.doi.org/10.1177/001112877201800407


http://dx.doi.org/10.1177/088626001016002005

doi: http://dx.doi.org/10.1177/009385480102800405


   http://dx.doi.org/10.1207/s15327752jpa5003_6


Mental Health Act 1986 (VIC) s. (Austl.).

Mental Health Act 2014 (VIC) s. (Austl.).

   doi: 10.1016/0160-2527(86)90055-5


*Sentencing Act 1991* (VIC) (Austl.).


http://dx.doi.org/10.1080/14999013.2010.501845


http://dx.doi.org/10.1080/14999013.2008.9914413


doi: 10.1080/02699200500297799


http://dx.doi.org/10.1080/14999013.2011.555934


Appendix A

Summary of Three Generations of Violence Risk Assessment*

<table>
<thead>
<tr>
<th>Generation</th>
<th>Characteristics of Decision Making</th>
<th>Nature of Accuracy of Outcomes</th>
</tr>
</thead>
</table>
| First      | • Unstructured clinical judgments employed without risk assessment tools  
            • Qualitative & subjective  
            • Dearth of available research  
            • Superior flexibility & idiosyncrasy  
            • Illusory correlations  
            • Fail to incorporate base rates  
            • Lack of known valid predictors  
            • Lack of knowledge about combining risk factors  
            • Do not incorporate situational/contextual factors | • Prediction of ‘dangerousness’  
• Lack of consistency in personal approaches to prediction (low inter-rater reliability)  
• Lack of transparency in decision-making  
• Binary yes/no outcomes  
• Low rates of accuracy, below chance  
• High rates of false positives  
• Tendency to over-predict violence  
• Criticism regarding the ethicality of prediction practice |
| Second     | • Influx of empirical research  
            • Focus on statistical risk prediction and actuarial decision-making  
            • Quantitative & objective  
            • Emergence of empirically identified risk factors  
            • Increased focus on situational & contextual variables  
            • Increased focus on specified populations and short-term predictions  
            • Limited number of risk factors included  
            • Cannot account for ‘broken legs’ | • Increased levels of accuracy, above chance  
• Clear transparency in decision-making  
• Prescribed systematic integration of risk-relevant information  
• Provide baseline level of risk  
• Insensitive to changes in risk over time due to emphasis on static risk factors  
• Not applicable to populations dissimilar to construction sample |
| Third      | • Significant growth in the body of research  
            • Conceptual changes from ‘risk prediction’ to ‘risk assessment’  
            • Focus on Structured Professional Judgment  
            • Amalgamation of two previous generations – tempering of actuarial judgments  
            • Accounting for both static and dynamic risk factors | • Continuation of higher levels of accuracy, far exceeding chance  
• Risk assessment outcome encapsulates risk factors, harm, and level of risk  
• Probabilistic outcomes (e.g., Low, Moderate, High)  
• From risk assessment to risk management  
• Sensitive to changes in risk over time due to inclusion of dynamic risk factors |

*Adapted from Ogloff & Davis (2005).
Appendix B

Violence Risk Factors and Coding Scheme for the HCR-20 (Version 2)

Historical

H1  Previous Violence
(0)  No previous violence
(1)  Possible/less serious previous violence (one or two acts of moderately severe violence)
(2)  definite serious previous violence (three or more acts of violence, or any acts of severe violence)

H2  Young Age at First Violent Incident
(0)  40 years and older at first known violent act
(1)  Between 20 and 39 years at first known violent act
(2)  Under 20 years at first known violent act

H3  Relationship Instability
(0)  relatively stable and conflict-free relationship pattern
(1)  Possible/less serious unstable and/or conflictual relationship pattern
(2)  Definite/serious unstable and/or conflictual relationship pattern

H4  Employment Problems
(0)  No employment problems
(1)  Possible/less serious employment problems
(2)  Definite/serious employment problems

H5  Substance Use Problems
(0)  No substance use problems
(1)  Possible/less serious substance use problems
(2)  Definite/serious substance use problems
H6  **Major Mental Illness**

(0)  No major mental illness  
(1)  Possible/less serious major mental illness  
(2)  Definite/serious major mental illness

H7  **Psychopathy**

(0)  Nonpsychopathic. Score of under 20 on the PCL-R, or under 13 on the PCL:SV.  
(1)  Possible/less serious psychopathy. Score of 20–29 on the PCL-R, or 13–17 on the PCL:SV.  
(2)  Definite/serious psychopathy. Score of 30–40 on the PCL-R, or 18–24 on the PCL:SV.

H8  **Early Maladjustment**

(0)  No maladjustment  
(1)  Possible/less serious maladjustment  
(2)  Definite/serious maladjustment

H9  **Personality Disorder**

(0)  No personality disorder  
(1)  Possible/less serious personality disorder  
(2)  Definite/serious personality disorder

H10  **Prior Supervision Failure**

(0)  No supervision failure(s)  
(1)  Possible/less serious supervision failure(s)  
(2)  Definite/serious supervision failure(s)
Clinical

C1 Lack of Insight
(0) No lack of insight
(1) Possible/less serious lack of insight
(2) Definite/serious lack of insight

C2 Negative Attitudes
(0) No negative attitudes
(1) Possible/less serious negative attitudes
(2) Definite/serious negative attitudes

C3 Active Symptoms of Major Mental Illness
(0) No active symptoms of major mental illness
(1) Possible/less serious active symptoms of major mental illness
(2) Definite/serious active symptoms of major mental illness

C4 Impulsivity
(0) No impulsivity
(1) Possible/less serious impulsivity
(2) Definite/serious impulsivity

C5 Unresponsive to Treatment
(0) Responsive to treatment
(1) Possible/less serious unresponsiveness to treatment
(2) Definite/serious unresponsiveness to treatment
**Risk Management**

**R1**  Plans Lack Feasibility

(0) Low probability that plans will not succeed  
(1) Moderate probability that plans will not succeed  
(2) High probability that plans will not succeed

**R2**  Exposure to Destabilisers

(0) Low probability of exposure to destabilisers  
(1) Moderate probability of exposure to destabilisers  
(2) High probability of exposure to destabilisers

**R3**  Lack of Personal Support

(0) Low probability of lack of personal support  
(1) Moderate probability of lack of personal support  
(2) High probability of lack of personal support

**R4**  Noncompliance with Remediation Attempts

(0) Low probability of noncompliance with remediation attempts  
(1) Moderate probability of noncompliance with remediation attempts  
(2) High probability of noncompliance with remediation attempts

**R5**  Stress

(0) Low probability of stress  
(1) Moderate probability of stress  
(2) High probability of stress
## Appendix C

### Violence Risk Factors and Sub-Items on the HCR-20 (Version 3)

<table>
<thead>
<tr>
<th>Item</th>
<th>Sub-Items</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Historical - History of Problems with:</strong></td>
<td></td>
</tr>
<tr>
<td>H1 Violence</td>
<td>(H1a) under 12</td>
</tr>
<tr>
<td></td>
<td>(H1b) 12-17</td>
</tr>
<tr>
<td></td>
<td>(H1c) 18+</td>
</tr>
<tr>
<td>H2 Other Antisocial Behaviour</td>
<td>(H2a) under 12</td>
</tr>
<tr>
<td></td>
<td>(H2b) 12-17</td>
</tr>
<tr>
<td></td>
<td>(H2c) 18+</td>
</tr>
<tr>
<td>H3 Personal Relationships</td>
<td>(H3a) Intimate Relationships</td>
</tr>
<tr>
<td></td>
<td>(H3b) Non-intimate Relationships</td>
</tr>
<tr>
<td>H4 Employment</td>
<td></td>
</tr>
<tr>
<td>H5 Substance Use</td>
<td></td>
</tr>
<tr>
<td>H6 Major Mental Illness</td>
<td>(H6a) Psychotic Disorders</td>
</tr>
<tr>
<td></td>
<td>(H6b) Major Mood Disorders</td>
</tr>
<tr>
<td></td>
<td>(H6c) Other</td>
</tr>
<tr>
<td>H7 Personality Disorder</td>
<td>(H7a) Antisocial, Psychopathic and Dissocial</td>
</tr>
<tr>
<td></td>
<td>(H7b) Other</td>
</tr>
<tr>
<td>H8 Traumatic Experiences</td>
<td>(H8a) Victimisation/ Trauma</td>
</tr>
<tr>
<td></td>
<td>(H8b) Adverse Childrearing Experiences</td>
</tr>
</tbody>
</table>
Violent Attitudes

Treatment or Supervision Response

Clinical - Recent problems with:

C1 Insight (C1a) Into Mental Disorder (C1b) Into Violence Risk (C1c) Into Need for Treatment

C2 Violent Ideation or Intent

C3 Symptoms of Major Mental Disorder (C3a) Psychotic Disorders (C3b) Major Mood Disorders (C3c) Other Major Mental Disorders

C4 Instability

C5 Treatment or Supervision Response (C5a) Compliance (C5b) Responsiveness

Risk Management – Future problems with:

R1 Professional Services

R2 Living Situation

R3 Personal Support

R4 Treatment or Supervision Response (R4a) Compliance (R4b) Responsiveness

R5 Stress and Coping
Appendix D

Violence Risk Appraisal Guide (VRAG) Scoring Rubric

Item 1: Lived with both biological parents to age 16
Score no if offender did not live continuously with both biological parents until age 16 except for death of one or both parents. In case of death of parent(s), score as for yes.

Yes = -2
No = +3

Item 2: Elementary school maladjustment (up to and including Grade 8)
Options are no problems; slight or moderate discipline or attendance problems; or severe (i.e., frequent or serious) behaviour or attendance problems (e.g., truancy or disruptive behaviour that persisted over several years or resulted in expulsion).

No problems = -1
Slight or moderate problems = +2
Severe problems = +5

Item 3: History of alcohol problems
Allot one point for each of the following: Alcohol abuse in biological parent, teenage alcohol problem, adult alcohol problem, alcohol involved in prior offence, alcohol involved in index offence.

0 points = -1
1 or 2 points = 0
3 points = +1
4 or 5 points = +2

Item 4: Marital status (at time of index offence)
Ever married includes lived common law in the same home for at least 6 months.

Ever married = -2
Never married = +1
Item 5: Criminal history score nonviolent offences
Convictions and charges for nonviolent offences prior to the index offence (from the Cormier-Lang System)

\[
\begin{align*}
0 &= -2 \\
1 - 2 &= 0 \\
> 3 &= +3
\end{align*}
\]

Item 6: Failure on prior conditional release
Includes parole violation or revocation, breach of or failure to comply with recognizance or probation, bail violation; and any new charges, including the index offence, while on conditional release.

\[
\begin{align*}
\text{No} &= 0 \\
\text{Yes} &= +3
\end{align*}
\]

Item 7: Age at index offence (at most recent birthday)

\[
\begin{align*}
\geq 39 &= -5 \\
34 - 38 &= -2 \\
28 - 33 &= -1 \\
27 &= 0 \\
\leq 26 &= +2
\end{align*}
\]

Item 8: Victim Injury
Index offence only. Most serious injury is scored.

\[
\begin{align*}
\text{Death} &= -2 \\
\text{Hospitalised} &= 0 \\
\text{Treated and released} &= +1 \\
\text{None or slight (includes no victim)} &= +2
\end{align*}
\]

Item 9: Any female victim (for index offence)

\[
\begin{align*}
\text{Yes} &= -1 \\
\text{No (includes no victim)} &= +1
\end{align*}
\]

Item 10: Meets DSM-III criteria for any personality disorder

\[
\begin{align*}
\text{No} &= -2 \\
\text{Yes} &= +3
\end{align*}
\]
Item 11: Meets DSM-III criteria for Schizophrenia

Yes = -3
No = +1

Item 12: Hare Psychopathy Checklist – Revised Score

≤ 4 = -5
5 – 9 = -3
10 – 14 = -1
15 – 24 = 0
25 – 34 = +4
≥ 35 = +12
Appendix E

Violence Risk Appraisal Guide – Revised (VRAG-R) Scoring Rubric

Item 1: Lived with both parents (except for death of parent)
Score no if offender did not live continuously with both biological parents until age 16 except for death of one or both parents. In case of death of parent(s), score as for yes.

Yes = -2
No = +2

Item 2: Elementary school maladjustment
Slight or moderate discipline or attendance problems, or severe (i.e., frequent or serious) behaviour or attendance problems (e.g., truancy or disruptive behaviour that persisted over several years or resulted in expulsion).

No problems = -3
Slight or moderate problems = +1
Severe problems = +4

Item 3: History of alcohol or drug problems
Allot one point for each of the following that can be answered yes: Alcohol problem before age 18, illicit drug problem before age 18, alcohol involved in a prior offence, illegal drug involved in a prior offence, alcohol involved in a current offence, illegal drug involved in a current offence. Prorate score if required then convert points to score as follows:

< 3 points = -2
3 points = 0
4 points = +1
> 4 points = +4

Item 4: Marital status
Ever married includes lived common law in the same home for at least 6 months.

Ever married = -1
Never married = +1
**Item 5: Criminal history score nonviolent offences**
Convictions and charges for nonviolent offences prior to the index offence (from the Cormier-Lang System)

\[
\begin{align*}
0 & = -3 \\
1 - 2 & = -1 \\
3 - 7 & = +1 \\
\geq 8 & = +5
\end{align*}
\]

**Item 6: Failure on prior conditional release**
Includes parole violation or revocation, breach of or failure to comply with recognizance or probation, bail violation; and any new charges, including the index offence, while on conditional release.

\[
\begin{align*}
\text{No} & = -2 \\
\text{Yes} & = +4
\end{align*}
\]

**Item 7: Age at index offence**

\[
\begin{align*}
> 40 & = -6 \\
33 - 40 & = -2 \\
29 - 32 & = 0 \\
22 - 28 & = +1 \\
< 22 & = +2
\end{align*}
\]

**Item 8: Criminal history score violent offences**
Convictions and charges for violent offences prior to the index offence (from the Cormier Lang System)

\[
\begin{align*}
0 & = -2 \\
1 - 5 & = +1 \\
6 - 12 & = +2 \\
\geq 13 & = +4
\end{align*}
\]

**Item 9: Prior admissions to correctional institutions**
Admission length one day or more

\[
\begin{align*}
0 & = -2 \\
1 & = +2 \\
2 \text{ or } 3 & = +3 \\
\geq 4 & = +6
\end{align*}
\]
**Item 10: Conduct disorder**
Score 1 point for each of the following before age 15: Repeated truancy, school suspension or expulsion, delinquency brought to the attention of juvenile court, running away from home at least twice, persistent lying, repeated sexual intercourse within a casual relationship, repeated drunkenness or substance abuse, thefts, vandalism, school grades markedly below expectations in relation to estimated or known IQ, chronic violations at home and/or school other than truancy, involvement in fights. Then convert points to scores as follows:

- 0 points = -2
- 1 point = 0
- 2 to 4 points = +4
- ≥ 5 points = +5

**Item 11: Sex offending**
- No known hands-on sex offences = -2
- Sex offence(s) exclusively against girls under the age of 14 = -1
- Sex offender but not known to fit another category = +1
- At least one sex offence against a female victim 14 or older = +3

**Item 12: Anti-sociality**
Convert PCL-R Facet 4 Total score to score as follows:

- 0 = -6
- ≥ 1 and < 1.5 = -4
- ≥ 1.5 and < 2.5 = -2
- ≥ 2.5 and < 3.5 = +1
- ≥ 3.5 and < 7.5 = +3
- ≥ 7.5 = +6
Appendix F

Ethics Approval Notices and Letters of Support

SHR Project 2014/056 Swinburne Ethics Clearance

Keith Wilkins
Sent: Thursday, 17 April 2014 2:28 PM
To: Michael Daffern, Delene Brookstein
Cc: RES Ethics

To: Prof Michael Daffern/Ms Delene Brookstein, CFBS/FHAD

Dear Michael and Delene

SHR Project 2014/056 For Better or Worse: The Predictive Validity of the Historical-Clinical-Risk Management 20 (Version 3; HCR-20 V3) and the Violence Risk Appraisal Guide (Revised; VRAG-R) in Community Settings
Prof Michael Daffern (FHAD); Ms Delene Brookstein et al
Approved Duration: 04/04/2014 to 31/01/2015 [Adjusted]

Ethical review of the above Swinburne supervised student project protocol was undertaken by Swinburne’s Human Research Ethics Committee (SUHREC) at its Meeting 03/2014 held 4 April 2014.

I am pleased to advise that the Committee approved the application, as submitted, agreeing that the proposed supervised student research was substantially in the public interest. The standard conditions for on-going ethics clearance (as applicable) are outlined further below. The overall Forensicare project chief investigator and the other organisations/HREC involved may need to be apprised of the Swinburne ethics clearance here issued.

In approving this project, the Ethics Committee noted that:

(a) the application involved the discrete use (as distinct from collection or disclosure) of identifiable health information within an overall Forensicare project already given ethics clearance by the Department of Justice HREC and accepted by Monash University whilst supervised student research was being conducted under Monash University auspices prior to candidature transfer;
(b) the discrete Swinburne supervised student project involved no recording of personally identifying information; and
(c) that the Chair and Secretary will finalise the reasoning that will be reported to the Health Services Commissioner as regards applicable Statutory Guidelines issued under the Health Records Act (Vic). In this regard, the reasoning submitted by the Swinburne chief investigator/supervisor pertaining to the overall Forensicare project will be incorporated as appropriate.

I also acknowledge receipt of evidence of support from the Department of Health concerning access to the RAPID database as emailed on 1 April 2014.

Standard Conditions
- All human research activity undertaken under Swinburne auspices must conform to Swinburne and external regulatory standards, including the current National Statement on Ethical Conduct in Human Research and with respect to secure data use, retention and disposal.
- The named Swinburne Chief Investigator/Supervisor remains responsible for any personnel appointed to or associated with the project being made aware of ethics clearance conditions, including research and consent procedures or instruments approved. Any change in chief investigator/supervisor requires timely notification and SUHREC endorsement.
- The above project has been approved as submitted for ethical review by or on behalf of SUHREC. Amendments to approved procedures or instruments ordinarily require prior ethical appraisal/

https://outlook.swin.edu.au/owa/?ae=Item&te=IPM.Note&id=RgAAAAADyRGnWt%2... 22/04/2014
clearance. SUHREC must be notified immediately or as soon as possible thereafter of (a) any serious or unexpected adverse effects on participants and any redress measures; (b) proposed changes in protocols; and (c) unforeseen events which might affect continued ethical acceptability of the project.

- At a minimum, an annual report on the progress of the project is required as well as at the conclusion (or abandonment) of the project. (Reports and requests made to VPHREC also being submitted to Swinburne Research for processing/endorsement may suffice.)

- A duly authorised external or internal audit of the project may be undertaken at any time.

Please contact the Research Ethics Office if you have any queries about Swinburne on-going ethics clearance, citing the project number. A copy of this email should be retained as part of project record-keeping.

Best wishes for the transferred project.

Yours sincerely

Keith

---------------------------------------------

Keith Wilkins
Secretary, SUHREC & Research Ethics Officer
Swinburne Research (H68)
Swinburne University of Technology
P O Box 218
HAWTHORN VIC 3122
Tel +61 3 9214 5218
Fax +61 3 9214 5267

https://outlook.swin.edu.au/owa/?ae=Item&it=IPM.Note&id=RgAAAAADyRGnWt%2...  22/04/2014
21 September 2012

Dear Dr McCarthy

Re: An Evaluation of the Problem Behaviour Program: A community based model for the assessment and treatment of problem behaviours

The Forensicare Operational Research Committee has given operational approval for your research to be conducted at Forensicare. This approval is subject to approval by the Department of Justice Human Research Ethics Committee.

You may not commence the research until you provide a letter of approval from the Department of Justice Human Research Ethics Committee, and you receive a letter from the Forensicare Research Committee acknowledging receipt of the approval letter.

Approval is given for the period between the anticipated commencement and completion dates as set out in the documentation. If the study has not been completed by the nominated completion date, an application for extension will be required.

To enable the Committee to meet its obligations in relation to monitoring Forensicare’s research program, you are required to provide a report within 12 months or on completion of your project, whichever is earlier.

Forensicare must report ongoing research activities to the Minister of Mental Health quarterly. As such you may be asked to provide information on the progress of your research. Failure to comply may lead to the withdrawal of consent for the research to be conducted at Forensicare.

Please ensure that the Operational Research Committee is notified of any matter that arises that may affect the conduct of the approved program.

Should you have any queries please don’t hesitate to contact Ms Mitali Gupta on 99472543 or email mitali.gupta@forensicare.vic.gov.au.

Yours sincerely,

[Signature]

Mr Danny Sullivan
Acting Clinical Director
Forensicare
Professor James Ogloff
Director, Centre for Forensic Behavioural Science
Forensicare
505 Hoddle Street
Clifton Hill 3068

11 February 2013

Dear Professor Ogloff,

Corrections Victoria (CV) has considered your recent application to the Corrections Victoria Research Committee for the project An Evaluation of the Problem Behaviour Program: A Community Based Model for the Assessment and Treatment of Problem Behaviours.

CV provides support for the project. CV believes the project has the potential to contribute valuable information to ensure that the Problem Behaviour Program is using best practice in assessing and treating problem behaviours in offenders, which will assist in reducing recidivism.

If you have any queries regarding this correspondence, please contact Laura Wilson, Research and Evaluation Officer, Research and Evaluation on 868 46567.

Yours sincerely,

Leanne Barnes
Director
Strategic Policy and Planning
28 March 2013

Dr Jennifer McCarthy
Victoria Institute of Forensic Mental Health
Forensicare

Re: An Evaluation of the Problem Behaviour Program: A community based model for the assessment and treatment of problem behaviours

Dear Dr McCarthy,

The Department of Justice Human Research Ethics Committee (JHREC) considered your application in relation to the project An Evaluation of the Problem Behaviour Program: A community based model for the assessment and treatment of problem behaviours at its meeting on 26 March 2013 and granted full approval for the duration of the investigation. The Department of Justice reference number for this project is CF/13/3430. Please note the following requirements:

- To confirm JHREC approval sign the Undertaking form attached and provide both an electronic and hardcopy version within ten business days.
- The JHREC is to be notified immediately of any matter that arises that may affect the conduct or continuation of the approved project.
- You are required to provide an Annual Report every 12 months (if applicable) and to provide a completion report at the end of the project (see the Department of Justice Website for the forms).
- Note that for long term/ongoing projects approval is only granted for three years, after which time a completion report is to be submitted and the project renewed with a new application.
- The Department of Justice would also appreciate receiving copies of any relevant publications, papers, theses, conferences presentations or audiovisual materials that result from this research.
- All future correspondence regarding this project must be sent electronically to ethics@justice.vic.gov.au and include the reference number and the project title. Hard copies of signed documents or original correspondence are to be sent to The Secretary, Justice Human Research Ethics Committee, PO Box 4356, Melbourne, Victoria, 3001.

If you have any queries regarding this application you are welcome to contact me on (03) 8684 1514 or email: ethics@justice.vic.gov.au.

Yours sincerely,

Ms Nicole Wilson
Secretary,
Department of Justice Human Research Ethics Committee
6 February 2014

Dr Jennifer McCarthy
Victorian Institute of Forensic Mental Health, Forensicare

Re: An Evaluation of the Problem Behaviour Program: A community based model for the assessment and treatment of problem behaviours

Dear Jennifer,

The Department of Justice Human Research Ethics Committee (JHREC) considered your request for amendment to the project An Evaluation of the Problem Behaviour Program: A community based model for the assessment and treatment of problem behaviours at its meeting on 29 January 2014 and granted full approval for the amendment for the duration of the investigation. The Department of Justice (DOJ) reference number for this project is CF/13/3430.

Please ensure that the JHREC is notified immediately of any matter that arises that may affect the conduct or continuation of the project. To enable the JHREC to fulfill its reporting obligations you are asked to provide an Annual Report every 12 months (if applicable) and to report on the completion of your project. Annual Report and Completion of Research forms are available on the Justice Human Research Ethics website.

All future correspondence regarding this project must be sent electronically to ethics@justice.vic.gov.au and include the DOJ reference number and the project title. Hard copies of signed documents or original correspondence may be sent to The Secretary, Justice Human Research Ethics Committee, PO Box 4356, Melbourne, Victoria, 3001.

If you have any queries regarding this application, you are welcome to contact me on (03) 8684 1514 or email: ethics@justice.vic.gov.au.

Yours sincerely,

Ms Nicole Wilson
Secretary
Department of Justice Human Research Ethics Committee
21 JAN 2013

Dr Jennifer McCarthy
Manager, Problem Behaviour Program
Victorian Institute of Forensic Mental Health
SOS Hoddle Street
CLIFTON HILL VICTORIA 3068

Dear Dr McCarthy

Thank you for your letter on 27 December, 2012 seeking support from the Department of Health for your research evaluation of the Problem Behaviour Program (PBP).

The department can provide you with the data from the RAPID database to assist in the client matching required by your research and hence I am providing formal confirmation of the department’s support. The support is conditional on approval given by the Department of Justice Human Research Ethics Committee (HREC) or by another HREC which is constituted in accordance with the National Statement on Ethical Conduct in Human Research (2007).

Please contact Mr Lachlan Rimes on 9096 6134 to discuss the data requirements of the project.

I would be pleased to receive a copy of your findings when the study is complete.

Yours sincerely

Paul Smith
A/Executive Director
Mental Health, Drugs and Regions Division
18 October 2012

Dr Jennifer McCarthy
Victorian Institute of Forensic Mental Health
505 Hoddle St
Clifton Hill, Vic 3068

Dear Dr McCarthy,

Re: Application to the Research Coordinating Committee for An Evaluation of the Problem Behaviour Program: A community based model for the assessment and treatment of problem behaviours

I write to advise you that the Victoria Police Research Coordinating Committee (RCC) has approved your request to undertake the above research involving Victoria Police.

This approval is conditional on:
- Evidence of approval from the Department of Justice HREC, and
- The Research Organisation signing a Research Agreement outlining the conditions governing the conduct of research involving Victoria Police.

Victoria Police Corporate Statistics has costed your request at $1,672 (Inclusive of GST).

You will need to ensure the completion of the Research Agreement and return it to Victoria Police before the research can commence. The Research Agreement will be forwarded to you electronically in due course.

If you have any queries or require further clarification please contact the RCC Secretariat on the contact details above.

Yours sincerely,

Dr David Ballek
Secretariat, Research Coordinating Committee
11 April 2014

Ms Nicole Wilson
Secretary
Human Research Ethics Committee
Department of Justice
GPO Box 43566QQ
MELBOURNE VIC 3002

Dear Ms Wilson,

NCIS – APPLICATION FOR NCIS ACCESS (PROJECT NUMBER M0390 / CT/13/3430)

Organisation: Victorian Institute of Forensic Health (Forensicare)

Project Title: An Evaluation of the Problem Behaviour Program: A community based model for the assessment and treatment of problem behaviours (FHEC Approved, 2013)

Principal Researcher: Dr Jennifer McCarthy

The above organisation’s application for access to NCIS data was considered and approved by the NCIS Research Committee (NRC) on 11 April 2014 for the above mentioned research.

Approval is subject to the following conditions:

i) That the researchers do not access the NCIS directly via online access; secure electronic data transfer of case information will be facilitated by the NCIS Unit.

ii) That the researchers sign a Confidentiality Undertaking prior to accessing NCIS data.

iii) That the research agency enters into an Access Agreement with the Department of Justice (via the NCB Unit), prior to accessing NCIS data.

iv) That no identifying information sourced from the NCIS will be disseminated.

Please find attached an Amendment Request Form submitted on behalf of the Victorian Institute of Forensic Health (Forensicare) for the Committee’s consideration.

Yours sincerely,

[Signature]

JOANNA COTSONIS,
(ACTING) CORONIAL LIAISON OFFICER.
Appendix G

Offence Categories for Criminal Versatility

PCL-R Item 20 (Hare, 2003)

1. Theft

Theft
Break and enter (with intent to commit an indictable offence)
Possession of housebreaking tools
Possession of stolen property
Possession of items obtained by crime
Loitering at night
Possession of stolen credit card
Retaining stolen property
Shopbreaking
Shoplifting
Auto theft
Unlawfully dwelling in house
Theft by wrongful possession

2. Robbery

Robbery
Armed Robbery
Robbery with violence
Extortion

3. Drug Offences

Possession of narcotics
 Trafficking in a narcotic
 Importing narcotics
 Possession for the purpose of trafficking
 Cultivation of a narcotic

4. Assault

Assault causing bodily harm
Threatening
Common assault
Aggravated assault
Grievous bodily harm
Wounding

5. Murder

First degree murder
Second degree murder
Manslaughter
Causing bodily harm with intent to endanger life
Discharging a firearm with intent to endanger life
Wounding with intent

6. Possession of Weapon

Possession of a weapon
Posession of explosives
Carrying a concealed weapon
Dangerous use of firearm
Pointing a firearm
Using a firearm (during commission of an indictable offence)

7. Sex Offences

Indecent assault
Rape
Incest
Buggery
Bestiality
Carnal knowledge
Committing an indecent act (in public)
Indecent Exposure
Gross indecency
Prostitution
8. Major Driving Offences

- Criminal negligence
- Driving while intoxicated
- Hit and run
- Dangerous driving
- Careless driving
- Driving while ability impaired
- Driving with more than 80 mgs of alcohol in blood
- Failure to provide a breath sample
- Failure to remain at the scene of an accident
- Failure to stop (at scene of an accident)

9. Fraud

- Fraud
- Forgery
- False pretenses
- Personation
- Uttering (a forger document)
- Dealing with a forged document
- Failure to given name and address (with intent)
- Fraudulently obtaining food or lodging
- Making a false statement
- Obtaining food by fraud
- Using stolen credit cards

10. Escape

- Escape (lawful custody)
- Unlawfully at large
- Breaking out of prison
- Failure to appear
- Failure to attend court
- Breach of recognizance
- Breach of bail
- Failure to comply (with probation order)
- Breach of probation

11. Kidnapping

- Unlawful confinement
- Forcible seize
- Hijacking
- Abduction

12. Arson

13. Obstruction of Justice

- Perjury
- Assaulting a police officer
- Obstructing a peace officer
- Resisting arrest
- Contempt of court

14. Crimes Against the State

- Treason
- Espionage
- Smuggling
- Evasion of income tax

15. Miscellaneous

- Vandalism
- Causing a disturbance
- Mischief
- Willful damage
- Driving whilst disqualified
- Driving while license suspended
- Driving while prohibited
- Vagrancy
- Living off the avails of prostitution
- Violation of immigration laws
- Bookmaking
- Disguised with intent

16. Stalking

Stalking is an additional offence category that was not included in the original offence categories by Hare (2003).
## Cormier-Lang System for Quantifying Criminal History

### Cormier-Lang Criminal History Score for Violent Offences

<table>
<thead>
<tr>
<th>Offence</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homicide (murder, manslaughter, criminal negligence causing death)</td>
<td>28</td>
</tr>
<tr>
<td>Attempted murder, causing bodily harm with intent to wound</td>
<td>7</td>
</tr>
<tr>
<td>Kidnapping, abduction, and forcible confinement</td>
<td>6</td>
</tr>
<tr>
<td>Aggravated assault, choking, administering a noxious substance</td>
<td>6</td>
</tr>
<tr>
<td>Assault causing bodily harm</td>
<td>5</td>
</tr>
<tr>
<td>Assault with a weapon</td>
<td>3</td>
</tr>
<tr>
<td>Assault, assaulting a police officer</td>
<td>2</td>
</tr>
<tr>
<td>Aggravated sexual assault, sexual assault causing bodily harm</td>
<td>15</td>
</tr>
<tr>
<td>Sexual assault with a weapon</td>
<td>12</td>
</tr>
<tr>
<td>Sexual assault, gross indecency (vaginal or anal penetration, victim forced to fellate offender)</td>
<td>10</td>
</tr>
<tr>
<td>Sexual assault (attempted rape, indecent assault)</td>
<td>6</td>
</tr>
<tr>
<td>Gross indecency (offender fellates or performs cunnilingus on victim)</td>
<td>6</td>
</tr>
<tr>
<td>Sexual assault (sexual interference, invitation to sexual touching)</td>
<td>2</td>
</tr>
<tr>
<td>Armed robbery (bank, store)</td>
<td>8</td>
</tr>
<tr>
<td>Robbery with violence</td>
<td>5</td>
</tr>
<tr>
<td>Armed robbery (not a bank or store)</td>
<td>4</td>
</tr>
</tbody>
</table>
## Cormier-Lang Criminal History Score for Non-Violent Offences

<table>
<thead>
<tr>
<th>Offence</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robbery (bank, store)</td>
<td>7</td>
</tr>
<tr>
<td>Robbery (purse snatching)</td>
<td>3</td>
</tr>
<tr>
<td>Arson and fire setting (church, house, barn)</td>
<td>5</td>
</tr>
<tr>
<td>Arson and fire setting (garbage can)</td>
<td>1</td>
</tr>
<tr>
<td>Threatening with a weapon, dangerous use of or pointing firearm</td>
<td>3</td>
</tr>
<tr>
<td>Threatening (uttering threats)</td>
<td>2</td>
</tr>
<tr>
<td>Theft over(^a) (includes car theft and possession of stolen property over(^a))</td>
<td>5</td>
</tr>
<tr>
<td>Mischief to public or private property over(^a)</td>
<td>5</td>
</tr>
<tr>
<td>Break and enter and commit indictable offence (burglary)</td>
<td>2</td>
</tr>
<tr>
<td>Theft under(^b) (includes possession of stolen property under(^b))</td>
<td>1</td>
</tr>
<tr>
<td>Mischief to public or private property under(^b)</td>
<td>1</td>
</tr>
<tr>
<td>Break and enter (includes break and enter with intent to commit an offence)</td>
<td>1</td>
</tr>
<tr>
<td>Fraud (extortion, embezzlement)</td>
<td>5</td>
</tr>
<tr>
<td>Fraud (forged check, impersonation)</td>
<td>1</td>
</tr>
<tr>
<td>Possession of a prohibited or restricted weapon</td>
<td>1</td>
</tr>
<tr>
<td>Procuring a person for or living on the avails of prostitution</td>
<td>1</td>
</tr>
<tr>
<td>Trafficking in narcotics</td>
<td>1</td>
</tr>
<tr>
<td>Dangerous driving, impaired driving (driving while intoxicated)</td>
<td>1</td>
</tr>
<tr>
<td>Obstructing a peace officer (including resisting arrest)</td>
<td>1</td>
</tr>
<tr>
<td>Causing a disturbance</td>
<td>1</td>
</tr>
<tr>
<td>Wearing a disguise with the intent to commit an offence</td>
<td>1</td>
</tr>
<tr>
<td>Indecent exposure</td>
<td>2</td>
</tr>
</tbody>
</table>

*Note. \(^a\) Roughly equivalent to grand larceny based on the value of the stolen property – as of 2003, over $5,000.00
\(^b\) Roughly equivalent to larcency based on the value of the stolen property – as of 2003, under $5,000.00*
Appendix I

*PCL-R Facet 4 (Antisociality) Scoring Rubric*

(Hare, 2003)

Item 10 (Poor Behavioural Controls) & Item 12 (Early Behavioural Problems):

(2) The item applies to the individual; a reasonably good match in most essential respects; behaviour is generally consistent with the flavor and intent of the item.

(1) The item applies to a certain extent but not to the degree required for a score of 2; a match in some respects but with too many exceptions or doubts to warrant a score of 2; uncertain about whether or not the item applies; conflicts between interview and file information that cannot be resolved in favour of a score of 2 or 0.

(0) The item does not apply to the individual; does not exhibit the trait or behavior in question, or exhibits characteristics that are the opposite of, or inconsistent with, the intent of the item.

**Item 18: Juvenile Delinquency**

(2) has a history of serious offences as an adolescent, including murder, attempted murder, manslaughter, rape, serious assault, robbery, auto theft, or other major theft, kidnapping, arson, fraud, trafficking drugs, major driving violations (e.g., drive while intoxicated, hit and run, dangerous driving), escape, and so forth.
(1) has a history of minor offences as an adolescent, including possession of drugs, minor theft, possession of stolen property, simple assault, mischief, causing a disturbance, minor driving violations (e.g., driving without a license), breach of probation, and so forth.

(0) Has no history of arrests for antisocial behavior as an adolescent.

**Item 19: Revocation of Conditional Release**

(2) Has had major violations or escapes including revocation of parole or mandatory supervision, breach of bail, escape, and so forth.

(1) Has had minor violations only, including suspension of parole or mandatory supervision, charges or convictions for fail to appear or breach of court order, and so forth.

(0) Has had no violations or escapes.

**Item 20: Criminal Versatility (see Appendix G for offence categories)**

(2) Has committed 6 or more types of offences

(1) Has committed 4 or 5 types of offences

(0) Has committed 3 or fewer types of offence
### Appendix J

**The Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE) Statement Checklist**  
Singh et al., 2014

<table>
<thead>
<tr>
<th>Section</th>
<th>Item</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>1</td>
<td>Include a structured abstract describing the study</td>
<td>xxi</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Identify the article as a risk assessment study in which predictive validity is measured</td>
<td>xxi</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Identify the risk assessment instrument(s) whose predictive validity is measured</td>
<td>xxi</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>State the nature of the principle outcome (e.g., violence, sexual violence, criminal offending, institutional misconduct)</td>
<td>xxi</td>
</tr>
<tr>
<td>Introduction</td>
<td>5</td>
<td>Provide the rationale and a summary of the scientific/theoretical background for the study</td>
<td>Ch 3 &amp; p.47</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>State the research questions and/or study aims</td>
<td>178-181</td>
</tr>
<tr>
<td>Method</td>
<td>7</td>
<td>Report the sample size</td>
<td>Ch 4</td>
</tr>
<tr>
<td>Participants</td>
<td>8</td>
<td>Report the sex/gender composition of the sample</td>
<td>185</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Report the average age at assessment (with dispersion parameter)</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Report the index offence composition of the sample</td>
<td>221</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>Report the characteristics of groups that underwent subgroup analysis</td>
<td>NA</td>
</tr>
<tr>
<td>Instrument Design</td>
<td>12</td>
<td>Report the acronym(s) and full names(s) of the instrument(s) under investigation with appropriate reference to source document</td>
<td>36 &amp; 24</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Report the number of items on the instrument(s) under investigation</td>
<td>61 &amp; 71</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>Report the approach by which the assessment information from the instrument(s) under investigation is organised into an overall evaluation of risk</td>
<td>Ch. 2</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>Report the population for which the instrument(s) under investigation was intended to be used</td>
<td>Ch.2</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>Report the outcome(s) that the instrument(s) under investigation was intended to assess</td>
<td>Ch.2</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>Report the length of follow-up for which manual-recommended probability estimates of risk were derived for instrument(s) under</td>
<td>178</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Report the cut-off score(s) and/or risk categories that the instrument(s) under investigation was designed to use to classify risk level</td>
<td>Ch.2</td>
<td></td>
</tr>
</tbody>
</table>
| 19 | Instrument Administration  
Report whether risk assessments were conducted in the context of research or practice                                                                                                                                                   | 185       |
| 20 | Identify when risk assessments occurred (e.g., pre-admission, admission, release, post-release)                                                                                                                                                           | 187       |
| 21 | Report the number of assessors in the study as well as their training in the administration of the instrument(s) under investigation                                                                                                                      | 183 & 199 |
| 22 | Identify the sources(s) of information used to administer the instrument(s) under investigation                                                                                                                                                           | 198       |
| 23 | Study Design  
Describe any modifications made to the instrument(s) under investigation                                                                                                                                                                                 | NA        |
| 24 | Report the geographical location and the clinical setting in which risk was assessed                                                                                                                                                                             | xxi & 173 |
| 25 | Describe the methods(s) used to recruit participants                                                                                                                                                                                                               | 184       |
| 26 | Identify the temporal design of the study (prospective or quasi-prospective)                                                                                                                                                                                     | 176       |
| 27 | Identify the setting in which participants were followed to ascertain whether outcome(s) of interest had occurred                                                                                                                                                 | 177       |
| 28 | Report the average length of follow-up and time at risk (with dispersion parameter, if not fixed), including a description of periods subtracted from follow-up time (e.g., incarceration and/or hospitalisation)                                                     | 213 & 211 |
| 29 | Predicted Outcome  
Specify the event(s) coded as meeting outcome criteria (e.g., assault, rape, homicide)                                                                                                                                                         | 190 & 194 |
| 30 | Identify the type (e.g., arrest, charge, conviction, incarceration) and source (e.g., criminal records, self-report. collateral) of information used to detect outcome occurrence                                                                                                    | 184       |
| 31 | Statistical Analysis  
Describe statistical methods used to conduct all analyses and report the purpose of each analysis                                                                                                                                               | Ch. 5     |
<p>| 32 | Report whether risk scores and/or categories of the instrument(s) under investigation were used as an independent variable in analyses                                                                                                                                 | Ch. 6     |
| 33 | Identify the statistical significance level used                                                                                                                                                                                                               | Ch. 6     |</p>
<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>Describe any subgroup level analysis planned <em>a priori</em></td>
<td>Ch. 6</td>
</tr>
<tr>
<td>35</td>
<td>Report inter-rater reliability for administration of the instrument(s) under investigation (if conducted). If inter-rater reliability was not assessed, clarify why not</td>
<td>315</td>
</tr>
<tr>
<td>36</td>
<td>Include at least one discrimination performance indicator when measuring predictive validity</td>
<td>234</td>
</tr>
<tr>
<td>Participant Outcomes</td>
<td>Report the rate of attrition</td>
<td>NA</td>
</tr>
<tr>
<td>37</td>
<td>Report the outcome occurrence rate for the entire sample as well as for the relevant subgroups</td>
<td>317</td>
</tr>
<tr>
<td>Predictive Validity</td>
<td>Report predictive validity performance indicators for each outcome of interest as specified in the methods with associated dispersion parameters</td>
<td>263-278</td>
</tr>
<tr>
<td>39</td>
<td>Report the number of participants with each risk score and/or in each risk category and how many went on the engage in the outcome(s) of interest</td>
<td>318</td>
</tr>
<tr>
<td>40</td>
<td>Report the results of subgroup analyses planned <em>a priori</em> as specified in the methods</td>
<td>Ch. 6</td>
</tr>
<tr>
<td>41</td>
<td>Describe and report the findings of any <em>post hoc</em> analyses conducted</td>
<td>Ch. 6</td>
</tr>
<tr>
<td>Discussion</td>
<td>Provide a summary of the principal findings, including a discussion of their relevance in the context of the current literature</td>
<td>318 &lt;</td>
</tr>
<tr>
<td>43</td>
<td>Discuss limitations of the study</td>
<td>338</td>
</tr>
<tr>
<td>44</td>
<td>Discuss the generalisability of the study findings</td>
<td>224</td>
</tr>
<tr>
<td>45</td>
<td>Discuss future research directions based on study findings</td>
<td>342</td>
</tr>
<tr>
<td>Disclosures</td>
<td>Report any commercial interests and/or source(s) of funding as well as their roles(s) in the conduct of the study</td>
<td>NA</td>
</tr>
<tr>
<td>47</td>
<td>Report whether an author or translator of the risk assessment instrument(s) under investigation was also a study author</td>
<td>NA</td>
</tr>
<tr>
<td>48</td>
<td>Report whether the study presented in the article has been published in an alternative form (e.g., government report)</td>
<td>NA</td>
</tr>
<tr>
<td>49</td>
<td>Report whether the sample or portion thereof has been studied in other publications</td>
<td>184</td>
</tr>
</tbody>
</table>

N/A = Not applicable to study.
Appendix K

Hanley & McNeil (1983) Correlation Coefficients

This image is unable to be reproduced online.
Please consult print copy held in the Swinburne Library.