**An Investigation of How Change in Dynamic Risk and Protective Factors
Affects the Prediction of Imminent Criminal Recidivism**

By

Simon T. Davies

A thesis submitted to the Victoria University of Wellington in fulfilment of the
requirements for the Degree of Doctor of Philosophy

Victoria University of Wellington

2019

Acknowledgements

I have been very fortunate to receive amazing help and support from a number of people throughout the process of completing this PhD. It is entirely due to these people and their support that my experience of completing a PhD has been so fulfilling and enjoyable.

There are a few people who deserve special mention. First, I want to thank the New Zealand Department of Corrections for making this research possible. In particular, I would like to thank the members of the Research and Evaluation team at National Office. Thanks to Jo and Paula for providing the dataset that was the centrepiece of this thesis and putting up with my barrage of questions when that dataset was being extracted. Thanks also to Marianne and Jill for checking my conference talks, and to Peter for his oversight of it all.

Second, thanks to all the members of the forensic, criminal justice, cellular psychology, and romantic relationships labs that I have been a part of during my PhD, as well as my fellow clinical students. Doing a PhD could easily be a lonely and frustrating experience but having a community of staff and students to run ideas past, practice talks on, and talk about things other than my research, has ensured that this was certainly not the case for me. Thanks especially to Julia, Tadhg, Jess, and Danica who have sat next to me and shared the beautiful view looking south out of EA623. Special thanks also to Wendy and Sue for leaving me a few easy crossword clues to do each day, and to Al for being such a tech whiz.

Third, thanks to my friends outside University, who have provided some much-needed balance to my life during this time. In particular, thanks to everyone I have played football and been on the committee with at VUW AFC. The club has been a big part of my life throughout my thesis and has been central to

my attempt to have a life away from my thesis. Thanks also to my family, who have been a constant source of support throughout not only this thesis, but throughout my life, and are the reason I have reached this point in my studies. You may not know what my thesis is about yet, but you will soon...only 271 pages to go. And of course, a very special thanks to Ellie. You have been a terrific sounding board, never afraid to tell me if my ideas are terrible (which they often are), have sat through my confused attempts to explain statistics, put up with and given feedback on countless talks, and kept me grounded throughout the whole process (a very challenging task at times). The fact that this thesis has gone so smoothly and that you came along just as this thesis was starting is certainly not a coincidence.

Finally, thanks to my three supervisors: Devon Polaschek, Caleb Lloyd, and Garth Fletcher. Garth, thanks for finding this research orphan a new home; the romantic relationships' lab was an unusual, but very welcoming landing spot. Caleb, I can't say enough about the help and support you have given me. No detail has ever been too small and no question too trivial. The level of thought and effort you put into everything you do is amazing to see and extremely difficult to emulate! Devon, my typically-last-minute search for an Honours supervisor could not have turned out any better. Even from a distance, you have provided amazing support throughout this process, pushing me to think in new and challenging ways, and selflessly imparting your seemingly-eternal wisdom in this area. You have shaped my future in many ways and I can't thank you enough for that.

Abstract

Dynamic risk and protective factors are changeable, psychosocial variables associated with an increased or decreased likelihood of future criminal behaviour. These variables have an important role in correctional psychology. In particular, they are increasingly central to the management and supervision of individuals released from prison. The changeable nature of these variables means that, with frequent reassessment, the likelihood of recidivism can be monitored during the release period, and intervention can be more carefully targeted to an individual's needs. However, research has yet to clearly demonstrate that reassessment of dynamic risk and protective factors can accurately track the likelihood of recidivism over time. Further, relatively little is known about how these variables change over time, and how change is associated with recidivism.

This thesis set out to investigate whether reassessment of a dynamic risk assessment tool—the Dynamic Risk Assessment for Offender Re-entry (DRAOR; Serin, 2007; Serin, Mailloux, & Wilson, 2012)—would enhance the prediction of imminent recidivism among a large sample of high-risk men ($n = 966$) released from prison on parole in New Zealand. The analyses addressing this question were divided into three primary sections: 1) an investigation of whether a single proximal assessment was a more accurate predictor of imminent recidivism than a single baseline assessment completed shortly after release; 2) an investigation of whether a single proximal assessment was a more accurate predictor of recidivism than a series of aggregated measures across multiple time points, and; 3) an investigation of whether several different measures of intra-individual change demonstrated incremental predictive validity over the most proximal

assessment. This approach represented a replication and extension of the framework set out by Lloyd (2015) in a recent thesis for testing whether reassessment of dynamic risk and protective factors enhances the prediction of imminent recidivism.

Across all three sections, results provided consistent evidence that the most proximal assessment was the most accurate predictor of imminent recidivism. The most proximal assessment was a significantly more accurate predictor than a baseline assessment, and neither aggregation nor measures of intra-individual change clearly improved predictive accuracy. These results highlight the importance of reassessment for monitoring changes in the likelihood of recidivism over time and have important implications for community correctional agencies who are tasked with managing individuals released from prison, particularly those deemed to be the highest risk of recidivism. The results also have theoretical implications for the concepts of dynamic risk and protective factors and their role in the process leading to recidivism. A better understanding of the recidivism process should lead to intervention strategies that are more effective at reducing recidivism.

Table of Contents

Acknowledgements	iii
Abstract	v
Table of Contents	vii
List of Tables	xiii
List of Figures	xvii
Introduction	1
Chapter 1: The Importance of Reassessment	7
What is Criminal Risk Assessment and Reassessment?	9
Risk Assessments and Reassessments in the Criminal Justice System	10
The Concept of Dynamic Risk Factors	14
The Concept of Protective Factors	21
Conclusion	25
Chapter 2: The Existing Empirical Evidence on Reassessment	27
Pre- and Post-Treatment Designs	28
Retrospective Designs	29
Prospective Multi-Wave Designs	31
A large, prospective New Zealand study.	37
The Current Research	39
Chapter 3: The Dataset	41
Correctional Environment	42
Parole.	42
Risk assessment.	44

Static risk.	44
Dynamic risk.	44
Psychometric Properties of DRAOR Data in Previous Research	45
Factor structure.	46
Predictive validity.	47
Choosing the Sample	49
Moving from the Raw Data to the Final Sample	51
Step 1: Men released more than once during the follow-up.	52
Step 2: Participants with no DRAOR assessments.	55
Step 3: Errors with participants and DRAOR assessments.	55
Step 4: Defining baseline.	56
Step 5: Defining recidivism.	58
Step 6: Ensuring prospective prediction.	59
Step 7: Gaps of more than six weeks between assessments.	60
Step 8: Truncating the follow-up.	61
The Final Sample	62
Participants.	63
Recidivism.	63
DRAOR assessments.	65
Psychometric Properties of Baseline DRAOR Assessments	65
Factor structure.	65
Predictive validity.	68
Analytical Technique	72
Structuring the Data for Analysis	75
The implications of defining a discrete unit of time.	75

Weeks with assessments and recidivism events. _____	76
Weeks with multiple assessments. _____	77
Weeks with no assessments. _____	78
Retrospective imputation. _____	78
Prospective imputation. _____	79
Imputation strategies for carrying forward assessments. _____	80
Final dataset after imputation. _____	83
A Final Note About Men with Only One Non-Imputed Assessment _____	84
Summary _____	86
Chapter 4: Does Reassessment Enhance Prediction? _____	89
Descriptive Statistics _____	90
Life table. _____	90
Inter-individual change. _____	93
Intra-individual change. _____	95
Correlations and internal reliability. _____	100
Longitudinal measurement invariance. _____	103
General Regression Assumptions _____	109
Normality and univariate outliers. _____	109
Homoscedasticity and linearity. _____	112
Multivariate outliers. _____	113
Multicollinearity. _____	117
Assumptions Specific to Cox Regression and Discrete-Time Hazard Models _____	117
Equivalence of censored and remaining cases. _____	117
Choosing a representation of time. _____	120

Proportionality of hazards. _____	123
Baseline vs. Reassessment Models _____	129
Univariate baseline and reassessment models. _____	130
Multivariate baseline and reassessment models. _____	138
Discussion _____	146
Summary of findings. _____	146
Conclusions. _____	148
Next steps. _____	150
Chapter 5: Does Aggregation Enhance Prediction? _____	151
Theoretical Background _____	152
Empirical Research _____	156
The Current Study _____	159
Preparation for Analyses _____	160
Results _____	163
Discussion _____	166
Chapter 6: Intra-Individual Change and Imminent Recidivism ____	171
Introduction _____	173
Is intra-individual change associated with recidivism? _____	173
Is intra-individual change <i>uniquely</i> associated with recidivism? ____	178
Why might intra-individual change be uniquely associated with recidivism? _____	183
Dynamic risk assessments are not sufficiently proximal to the outcome. _____	186

Risk assessment tools do not include all variables associated with recidivism. _____	192
Summary. _____	194
Conclusion. _____	195
The Current Study _____	196
Defining and calculating intra-individual change. _____	197
Results _____	203
Was intra-individual change associated with imminent recidivism? _____	203
How did change scores differ for recidivists and non-recidivists? _____	203
Did change scores predict imminent recidivism? _____	208
Was intra-individual change <i>uniquely</i> associated with imminent recidivism? _____	212
Discussion _____	216
Summary of results. _____	216
Comparison with previous research. _____	218
Theoretical implications and practical applications. _____	220
Limitations and future research. _____	223
Conclusion. _____	225
Chapter 7: Alternative Measures of Intra-Individual Change _____	227
Variability Scores _____	228
Mean Scores _____	232
The Current Study _____	235
Defining and calculating intra-individual change. _____	235

Measures of variability. _____	235
Mean scores. _____	238
Results _____	238
Measures of variability. _____	238
How did variability scores differ for recidivists and non- recidivists? _____	238
Were variability scores associated with imminent recidivism? _____	239
Were variability scores uniquely associated with imminent recidivism? _____	241
Mean scores. _____	245
Discussion _____	247
Chapter 8: General Discussion _____	253
Theoretical Implications _____	257
The concepts of dynamic risk and protective factors. _____	257
The distinction between stable and acute variables. _____	260
Practical Implications _____	265
Community supervision. _____	265
Other areas of correctional practice. _____	268
General Limitations and Future Research _____	270
Summary _____	272

List of Tables

Table 3.1 <i>Area Under the Curve (AUC) Statistics and Hazard Ratios from Cox Regression Analyses for Baseline Measures of Static (RoC*RoI) and Dynamic (DRAOR) Risk Predicting General Recidivism Over Follow-up Periods of 6 Weeks and 6 Months.</i>	71
Table 3.2 <i>Model Fit Indices (AIC) for Discrete-Time Hazard Models Using DRAOR Subscales to Predict Time to Recidivism, Comparing Six Dataset-Building Strategies.</i>	82
Table 4.1 <i>Life Table Showing When High-Risk Parolees Recidivated or Were Censored Over the First 26 Weeks After Release from Prison.</i>	91
Table 4.2 <i>Means, Standard Deviations and Percentage of Non-Imputed Scores for DRAOR Subscales for Each Week of the 26-Week Follow-Up.</i>	94
Table 4.3 <i>Proportion of Men Showing Weekly Change for DRAOR Subscales for Each Week of the 26-Week Follow-Up.</i>	97
Table 4.4 <i>Internal Reliability (Cronbach's Alpha), Inter-Subscale Correlations, and Correlations with Static Risk (RoC*RoI) for Each Week of the 26-Week Follow-Up.</i>	101
Table 4.5 <i>Number of Components Obtained from Parallel Analysis, and Measures of Fit from Confirmatory Factor Analyses (CFA) Testing the Original Three-Factor Structure of the DRAOR for Each Week of the 26-Week Follow-Up.</i>	105
Table 4.6 <i>Tests for Measurement Invariance Between Scores from the Beginning and Middle, and the Beginning and End of the 26-Week Follow-Up Using Confirmatory Factor Analysis.</i>	108
Table 4.7 <i>Measures of Normality—Skewness Ratios, Kurtosis Ratios, and Number of Extreme Outliers—for Each Week of the 26-Week Follow-Up.</i>	110

Table 4.8 <i>Results for Fitting Four Discrete-Time Hazard Models Predicting Criminal Recidivism Including Quadratic Predictor Terms.....</i>	114
Table 4.9 <i>Independent Samples t-Tests Comparing the Static Risk (RoC*RoI) and DRAOR Subscale Scores of Censored and Uncensored Non-Recidivists for Each Week of the 26-Week Follow-Up.</i>	118
Table 4.10 <i>Comparison of Different Representations of Time for a Discrete-Time Hazard Model.....</i>	121
Table 4.11 <i>Results for Fitting Four Discrete-Time Hazard Models Predicting Criminal Recidivism Using Score by Time Interaction Terms.....</i>	128
Table 4.12 <i>Results from Three Discrete-Time Hazard Models Predicting Time to Recidivism Using Scores from the DRAOR Stable Risk Subscale.</i>	131
Table 4.13 <i>Results from Three Discrete Time Hazard Models Predicting Time to Recidivism Using Scores from the DRAOR Acute Risk Subscale.....</i>	132
Table 4.14 <i>Results from Three Discrete Time Hazard Models Predicting Time to Recidivism Using Scores from the DRAOR Protective Factor Subscale.....</i>	133
Table 4.15 <i>Results from Three Discrete Time Hazard Models Predicting Time to Recidivism Using Scores from the RoC*RoI and the DRAOR Subscales.....</i>	141
Table 5.1 <i>Fictional Example of a Single Individual's Score on the DRAOR Acute Subscale Across the First 13 Weeks of the Follow-up, and the Rolling Mean, and Moving Rolling Mean Scores Calculated from those Observed Scores. ...</i>	162
Table 5.2 <i>Model Fit (AIC) for Discrete-Time Hazard Models and Effect Size for Cox Regression with Time-Varying Predictor Models for Rolling Mean and Moving Rolling Mean Scores for the Three DRAOR Subscales.</i>	165

Table 6.1 <i>Fictional Example of a Single Individual's Scores on the DRAOR Acute Subscale Across the First 13 Weeks of the Follow-up, and the Change Scores Calculated from those Observed Scores.</i>	199
Table 6.2 <i>Results from Two Discrete Time Hazard Models Predicting Time to Recidivism Using Scores from the DRAOR Acute Subscale.</i>	202
Table 6.3 <i>Mean Change Scores for the DRAOR Subscales Over Different Time Periods Immediately Prior to Recidivism.</i>	205
Table 6.4 <i>Parameter Estimates for Discrete-Time Hazard Models, and Effect Sizes for Cox Regression with Time-Varying Predictor Models Using Scores from a Baseline Assessment and Change Scores Across Different Time Periods on the DRAOR Subscales.</i>	210
Table 6.5 <i>Parameter Estimates for Discrete-Time Hazard Models, and Effect Sizes for Cox Regression with Time-Varying Predictor Models Using Scores from the Single Most Proximal Assessment and Change Scores Across Different Time Periods on the DRAOR Subscales.</i>	214
Table 7.1 <i>Fictional Example of a Single Individual's Scores on the DRAOR Acute Subscale Across the First 13 Weeks of the Follow-up, and the Rolling Total Mean, Change Score, and Variability Scores Calculated from those Observed Scores.</i>	237
Table 7.2 <i>Regression Coefficients of Variability Measures in Discrete-Time Hazard Models Predicting Imminent Recidivism After Controlling for Baseline DRAOR Score.</i>	240
Table 7.3 <i>Regression Coefficients of Variability Measures in Discrete-Time Hazard Models Predicting Imminent Recidivism After Controlling for Proximal DRAOR Score.</i>	242

Table 7.4 <i>Effect Sizes for Cox Regression with Time-Varying Predictor Models Using Scores from the Single Most Proximal Assessment and Measures of Variability Across Different Time Periods on the DRAOR Acute Subscale.</i>	243
Table 7.5 <i>Parameter Estimates for Discrete-Time Hazard Models, and Effect Sizes for Cox Regression with Time-Varying Predictor Models Using Scores from the Single Most Proximal Assessment and Mean Scores Across Different Time Periods on the DRAOR Subscales.</i>	246

List of Figures

<i>Figure 3.1.</i> Data flow chart showing decision steps made to obtain final sample from raw data.....	53
<i>Figure 4.1.</i> Hazard rate (proportion of men who recidivated) for each week of the 26-week follow-up.	92
<i>Figure 4.2.</i> Cumulative proportion of men remaining crime-free in the community across the 26 weeks of the follow-up.....	92
<i>Figure 4.3.</i> Logit-transformed hazard rate (proportion of men who recidivated) for each week of the 26-week follow-up, with superimposed linear and quadratic lines of best fit.	122
<i>Figure 4.4.</i> Logit-transformed hazard rates for each week of the 26-week follow-up for RoC*RoI scores above and below the median, with superimposed linear lines of best fit.	126
<i>Figure 4.5.</i> Logit-transformed hazard rates for each week of the 26-week follow-up for scores approximately 1 SD above and below the mean for the DRAOR stable risk subscale, with superimposed linear lines of best fit.	126
<i>Figure 4.6.</i> Logit-transformed hazard rates for each week of the 26-week follow-up for scores approximately 1 SD above and below the mean for the DRAOR acute risk subscale, with superimposed linear lines of best fit.....	127
<i>Figure 4.7.</i> Logit-transformed hazard rates for each week of the 26-week follow-up for scores approximately 1 SD above and below the mean for the DRAOR protective subscale, with superimposed linear lines of best fit.....	127
<i>Figure 4.8.</i> Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for proximal and baseline scores on the DRAOR	

stable subscale. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid. 135

Figure 4.9. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for reassessment and baseline scores on the DRAOR acute subscale. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid..... 136

Figure 4.10. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for reassessment and baseline scores on the DRAOR protective subscale. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid..... 137

Figure 4.11. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for RoC*RoI scores and a model including RoC*RoI scores and baseline scores on the three DRAOR subscales. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid..... 143

Figure 4.12. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for RoC*RoI scores and a model including RoC*RoI scores and proximal scores on the three DRAOR subscales. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid..... 144

Figure 4.13. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for multivariate models including RoC*RoI scores and either baseline or proximal scores on the three DRAOR subscales. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid. 145

Figure 6.1. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for the single most proximal assessment score and for a model including the proximal score and 3-week change scores on the DRAOR acute subscale. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid..... 215

Introduction

Assessments of how likely it is that an individual will engage in criminal behaviour in the future have a central role in the criminal justice system. These *risk assessments* are particularly important for decisions related to prison release. The outcome of a risk assessment can impact on when individuals are released from prison, what conditions they are required to adhere to upon release, and how they are managed by correctional agencies during their release, including the extent and focus of rehabilitation and reintegration efforts. Given their importance, these assessments must accurately reflect the likelihood of criminal behaviour occurring, or at least a measure of detected criminal behaviour known as *recidivism*. Assessments that are not accurate predictors of recidivism could, among other consequences, result in unjustified restrictions on individual liberty or place the community at risk of harm.

Risk assessments usually involve rating a series of different psychosocial variables associated with either an increased—*risk* factors—or decreased—*protective* factors—likelihood of recidivism. Risk and protective factors can be further classified on the basis of whether they can change over time. Unchangeable variables are known as *static* factors, whereas changeable variables are known as *dynamic* factors. The capacity for change means dynamic variables have several advantages over static variables. Dynamic variables can be targets for intervention, with the idea that effecting change in relevant dynamic variables will impact on an individual's likelihood of recidivism. Dynamic variables can also be used to monitor risk over time, which should allow for more carefully-targeted intervention and better-informed supervision in the community.

To monitor risk over time, or to establish that intra-individual change is associated with recidivism, dynamic variables must be assessed on multiple occasions, or *reassessed*. In this context, reassessment refers to a repeated assessment of the same risk and protective factors. A single reassessment following rehabilitation treatment programmes is already common practice in correctional psychology, and regular reassessment during community re-entry is becoming more common. Just as with initial assessments, any reassessment *score* must accurately predict recidivism. In fact, a reassessment score should predict recidivism more accurately than a previous assessment score because the new assessment is more *proximal* to the outcome and should account for any changes that might have occurred, leading to more accurate measurement of the relevant variables. More accurate measurement should lead to more accurate prediction. The requirement that reassessment improve prediction is therefore both inherent to the concept of dynamic risk and protective factors and essential for how these variables are used in practice.

This thesis is an examination of how reassessment impacts on the prediction of recidivism. More specifically, this thesis is focused on *imminent* recidivism. The term imminent is used to refer to the short amount of time (e.g., days or weeks) between an assessment and the measurement of recidivism. The focus on imminent recidivism reflects the fact that reassessments are most relevant to behaviour occurring shortly after an assessment. In theory, something either internal or external to the individual (e.g., the motivation or opportunity to offend) must change in the immediate lead-up to criminal behaviour for that behaviour to occur. Additionally, in practice, the risk of immediate criminal behaviour is the most important consideration for those

tasked with supervising individuals in the community, so studying imminent recidivism has a clear and important practical application.

The primary research question in this thesis can be framed in two alternate ways. Initially, we¹ frame the question as to what extent reassessment of dynamic risk and protective factors impacts on the prediction of imminent recidivism during re-entry into the community after imprisonment. This question highlights an essential characteristic of dynamic risk and protective factors: reassessment of truly dynamic predictors should enhance the prediction of recidivism. Therefore, one of the primary aims of this thesis is to test the theoretical concepts of dynamic risk and protective factors. To answer that research question, this thesis starts with a replication of another recent thesis (Lloyd, 2015) that proposed a three-step framework for testing whether reassessment of dynamic risk and protective factors enhances the prediction of imminent recidivism, and more broadly, testing the concepts of dynamic risk and protective factors. That thesis represents one of the few previous studies to address the primary research question using reassessment data collected during re-entry and, using that data, to attempt to establish the link between the *timing* of reassessments and recidivism.

The replication of that thesis has been broken down into several stages. In Chapter 1, we explore the practical and theoretical rationale for why it is important to study reassessments; specifically, why should we examine the impact of reassessment on prediction? In Chapter 2, we present the existing

¹ First person plural pronouns (i.e., “we”, “our”, “ours”) are used throughout this thesis. The use of the plural form is an acknowledgement of the fact that, although the work in this thesis is my own, I drew on the support and advice of many other people, particularly my supervisors, in conducting the research.

empirical evidence relevant to that question and highlight the gaps in the existing literature that our study aims to fill. In Chapter 3, we describe the dataset used for the analyses, including a detailed description of the dynamic risk assessment measure used, why the specific data were chosen, and how the data were structured in preparation for the analyses. In Chapter 4, we test the first step in Lloyd's (2015) framework. The analyses in Chapter 4 start with an extensive set of tests of the model assumptions, and then move to the substantive analyses testing to what extent a single proximal reassessment score is a more accurate predictor of imminent recidivism than a score from an assessment completed shortly after release.

In Chapter 5, we explore and test the second and third steps in Lloyd's (2015) framework. Those steps focus on whether aggregation of reassessments can further enhance the prediction of imminent recidivism. The chapter begins with a short review of the literature related to measurement aggregation and considers how that literature applies to dynamic risk and protective factors and the prediction of imminent recidivism. That review provides a detailed examination of the theoretical rationale behind the second and third steps in Lloyd's framework, and questions whether aggregation should ever benefit prediction. We then present a series of analyses to examine the extent to which a single proximal reassessment score is a more accurate predictor of imminent recidivism than an average score from multiple prior assessments.

Framing the research question around the value of reassessment is one possibility. The alternative framing is to what extent intra-individual change in dynamic risk and protective factors is associated with imminent recidivism. A finding that reassessment enhances prediction also provides evidence that intra-

individual change in dynamic risk and protective factors is associated with imminent recidivism, because a reassessment score accounts for any intra-individual change that has occurred. However, when the question is framed around intra-individual change, there is a risk that results can be misinterpreted. For example, a finding that intra-individual change is associated with recidivism can be misinterpreted as evidence that the mere fact of change is what matters, when the more likely explanation is that change matters because it allows calculation of the current, or *most proximal*, value of relevant dynamic variables.

In Chapter 6, we present a detailed discussion of these conceptual issues, and examine the limited existing empirical evidence in this area. We then outline and test a method for examining the relationship between intra-individual change and imminent recidivism. Specifically, we propose a method for testing to what extent intra-individual change in dynamic risk and protective factors is associated with imminent recidivism *after* controlling for the single most proximal assessment score. That method represents a revision and extension of the framework proposed by Lloyd (2015) for testing whether reassessment enhances the prediction of imminent recidivism. The revised framework aims to provide a comprehensive approach for finding *the best way* in which reassessment data should be used to predict imminent recidivism.

In Chapter 7, we examine the same questions posed in Chapter 6, but this time using two alternative ways of operationalising *intra-individual change*. Specifically, we explore several different measures of variability within an individual sequence, the relationship between those measures and imminent recidivism, and the extent to which those measures enhance the prediction of imminent recidivism alongside the single most proximal assessment score. We

also re-examine the average measures from Chapter 5, this time looking at whether the measures demonstrated significant incremental predictive validity over the most proximal score, and the extent to which these measures enhanced the prediction of imminent recidivism.

Finally, in Chapter 8, the thesis concludes with a recap of the main findings and discussion of the explanations, implications, and limitations of those findings. In particular, we discuss the theoretical implications of the results for the concepts of dynamic risk and protective factors and the conceptual distinction between acute and stable dynamic variables. We also discuss the practical implications of the findings for community supervision officers and correctional agencies tasked with supervising and managing individuals in prison and in the community.

Chapter 1: The Importance of Reassessment

In New Zealand, research suggests over 70% of people released from prison will be reconvicted within 5 years (Nadesu, 2009). Many other countries report similarly high recidivism rates (Fazel & Wolf, 2015). Of course, recidivism rates generally only include offending behaviour that gets detected and results in an outcome such as arrest or conviction. Research indicates that the clear majority of offending goes undetected (Farrington et al., 2006; Farrington et al., 2003; Farrington, Jolliffe, Loeber, & Homish, 2007), suggesting reoffending rates are much higher than measured recidivism rates. High rates of reoffending contribute substantially to the victimisation and harm in the community that occurs because of criminal behaviour. Reoffending is also extremely expensive; chronic criminal careers can cost taxpayers millions of dollars (M. A. Cohen, Piquero, & Jennings, 2010; Piquero, Jennings, & Farrington, 2013; Welsh et al., 2008) and place substantial additional strain on the limited resources available in the criminal justice systems of most jurisdictions.

The social and fiscal impact of high rates of reoffending has led to an increased focus on successfully reducing recidivism. Correctional rehabilitation programmes form a crucial part of these efforts. The effectiveness of these human service interventions, once in doubt (Martinson, 1974), is now widely accepted. Research shows that interventions that are carefully designed and delivered with integrity, can have a substantial impact on recidivism (Andrews, Zinger, et al., 1990; Dowden & Andrews, 1999a, 1999b, 2000; Hanson & Morton-Bourgon, 2009; Lipsey & Cullen, 2007; McGuire, 2002). There is also growing evidence that rehabilitation-focused supervision conducted by well-trained community correctional officers can reduce recidivism (Chadwick, Dewolf, &

Serin, 2015). Despite this progress, there remains considerable room for improvement. More effective techniques for reducing recidivism are needed.

One way of improving interventions is to advance the understanding of the concepts that inform those interventions (Taxman & Caudy, 2015). Dynamic risk and protective factors are two concepts that are central to correctional rehabilitation and management. However, these concepts are poorly defined and theoretically under-developed (Serin, Chadwick, & Lloyd, 2016; Ward, 2016, 2017); therefore, they represent a clear avenue for improving interventions. A better understanding of what these concepts are and how they are associated with recidivism could lead to improvements in how they are used in efforts to reduce recidivism.

In this chapter, we look at the role of dynamic risk and protective factors in the criminal justice system, particularly in risk assessment, risk management, and correctional rehabilitation. We also examine the concepts of dynamic risk and protective factors, explore their role in explanations of criminal behaviour, and identify aspects of these concepts that would benefit from further theoretical development. Ultimately, this chapter presents the argument that more empirical research is needed to advance existing practice and our current understanding of the concepts of dynamic risk and protective factors. In particular, more research is needed that examines the accurate classification of variables as dynamic risk or protective factors. As we will show, this type of research could advance the understanding of several existing conceptual issues with dynamic risk and protective factors. This conceptual progress could provide guidance for the development of more effective methods of reducing recidivism.

What is Criminal Risk Assessment and Reassessment?

A criminal risk assessment generally involves measuring a series of psychosocial variables that are associated with future criminal behaviour. Most risk assessments follow the principle that risk is determined by a combination of variables; no single variable provides a strong indication of risk on its own (Mann, Hanson, & Thornton, 2010). The combination of offending-related variables is used to provide an estimate of the likelihood that an individual will offend at some point in the future. In correctional psychology, where the individuals have usually been convicted of at least one prior offence, the assessment focuses on the likelihood of that individual *reoffending*. The measurable form of reoffending is known as recidivism and can refer to several outcomes including rearrests or reconvictions. Research suggests that structured methods of assessing and combining risk-related variables (e.g., actuarial risk tools) are more accurate at predicting recidivism than less-structured approaches that rely heavily on professional judgment (Andrews, Bonta, & Wormith, 2006; Hanson & Morton-Bourgon, 2009).

The variables measured in a risk assessment take many different forms. They can include demographic characteristics, psychological traits, behaviour patterns, and external circumstances or features of the surrounding environment. Variables that are associated with an increased likelihood of recidivism are referred to as *risk factors*. Increasingly, risk assessments also measure *protective factors*, which are variables that are theoretically associated with a reduced likelihood of recidivism (de Vries Robbé, de Vogel, & Stam, 2012). The variables can also be categorised by whether they change over time (Bonta, 1996; Zamble & Quinsey, 1997). Some variables such as age, ethnicity or criminal

history cannot change and thus are commonly described as *static factors*.² Variables that, theoretically, can change over time are described as *dynamic factors*. Dynamic risk factors have been further classified into slowly changing *stable* factors such as antisocial attitudes, and more rapidly changing *acute* factors such as negative mood and intoxication (Hanson, Harris, Scott, & Helmus, 2007; Hanson & Harris, 2000).

Criminal risk *reassessment* is rarely defined. Instead, it is implicit that a reassessment is an assessment of the same individual using the same method of assessment, most commonly the same risk assessment tool (Brown, St. Amand, & Zamble, 2009; Jones, Brown, & Zamble, 2010; Vose, 2016). The reason for conducting the reassessment is to ascertain whether any of the dynamic variables have changed since the previous assessment. The implication is that any changes in those variables will indicate that the individual's overall likelihood of recidivism has changed since the previous assessment (Andrews et al., 2006). To what extent that implication is accurate is a central focus of this thesis.

Risk Assessments and Reassessments in the Criminal Justice System

Risk assessments are now used at almost every point in the criminal justice process. A risk assessment can inform decisions on the granting of bail, sentence length, sentence type (e.g., imprisonment or a community sentence),

² Static factors are sometimes divided into *fixed* markers such as ethnicity and *variable* markers such as age and criminal history (Tonry, 2014). Fixed markers cannot possibly change over time, whereas variable markers can change, but usually only with the passage of time. The important common feature of fixed and variable markers is the inability to be altered through any purposive intervention effort (e.g., it is impossible to alter an individual's age, ethnicity, or their criminal history). So, while some can technically change, as a rule, static factors are not useful as targets for intervention (Mann et al., 2010).

security classification in prison, early release from prison (i.e., parole), access to rehabilitation, targets of any rehabilitation, and the conditions or restrictions imposed upon release (James, 2015). As the use of risk assessment has grown in recent years (Monahan & Skeem, 2014), the implications of that use have come under increased scrutiny. In some areas, particularly sentencing, risk assessment has been criticised for resulting in discriminatory outcomes for minorities and other disadvantaged groups in society (Tonry, 2014). However, these criticisms apply mostly to risk assessments that use static risk factors. When assessments focus on dynamic factors, and particularly when the goal is to identify who should receive available rehabilitation and how they should be rehabilitated, it is strongly recommended that risk assessment should guide decision making in every case (Bonta & Andrews, 2016; Tonry, 2014).

Predominantly, this recommendation to use risk assessment comes from the risk-need-responsivity (RNR; Andrews, Bonta, & Hoge, 1990) model of offender rehabilitation. The RNR model is the dominant framework for guiding correctional interventions (Polaschek, 2012). The three main principles of this model are *risk*, *need*, and *responsivity*. These three principles stipulate, respectively, that (a) more intensive intervention should be provided to higher risk individuals, (b) intervention should target changeable variables that are associated with recidivism, and (c) interventions should be delivered in a way that is consistent with both the research evidence on effective models for changing human behaviour (e.g., a cognitive behavioural approach), and the learning style and ability of the individual (Bonta & Andrews, 2016). There is now a substantial body of research demonstrating that interventions that adhere to these principles result in significant reductions in recidivism (Bonta &

Andrews, 2016). Adherence, particularly to the risk and need principles, requires valid and reliable risk assessment. Therefore, the research makes clear that correctional agencies aiming to reduce recidivism should be conducting risk assessment and using those assessments to guide their practice.

It is also implicit from the RNR model that correctional agencies should be routinely *reassessing* criminal risk. The need principle states that interventions should target dynamic risk factors and attempt to “move [them] in the direction of becoming strengths” (Andrews & Bonta, 2010, p. 46). A major reason that dynamic risk factors were added to risk assessments was to provide changeable treatment targets (Bonta, 1996). The only way to evaluate whether change has occurred, and consequently, whether the risk of recidivism has changed, is by reassessing risk. Consequently, some correctional agencies (e.g., the New Zealand Department of Corrections; Polaschek & Kilgour, 2013) now routinely reassess dynamic risk factors before and after a treatment programme to evaluate change in dynamic risk. It is also increasingly common for these assessments to include protective factors or strengths (de Vries Robbé & Willis, 2017). Reassessments can provide important information about whether an individual’s risk of recidivism has changed over the course of treatment (Beggs, 2010; Serin, Lloyd, Helmus, Derkzen, & Luong, 2013).

Reassessment of dynamic risk and protective factors is also used to assist the supervision of individuals serving sentences in the community. In community settings, correctional agencies use risk assessment to guide important decisions about ongoing supervision (Serin, 2007; Serin et al., 2012). For example, training programs for community corrections officers aiming to reduce the likelihood of recidivism instruct them to focus their rehabilitation

efforts on dynamic risk factors identified during assessment as being most relevant for the individual they are supervising (Bonta et al., 2011). In addition to rehabilitation, community correctional agencies can use reassessment information to assist with risk management. At its most extreme, reassessment information might be used as part of a decision to recall individuals to prison because they pose an undue risk to the community (in New Zealand, see the Parole Act 2002, s61). These decisions have serious and immediate consequences: underestimating risk can compromise public safety, while overestimating risk can unfairly restrict individual liberty.

Justification for the growing use of risk assessment in criminal justice decision-making rests predominantly on the premise that a risk assessment provides accurate information about an individual's likelihood of recidivism. Several decisions in the criminal justice system rely on the *reassessment* of dynamic risk and protective factors. For reassessments, it is not enough that scores predict recidivism because there will already be an existing assessment score that should give an indication of the likelihood of recidivism. A reassessment score must go further; it must provide information about the likelihood of recidivism that was not provided by the previous assessment score. From a practical perspective, whether that means assessing rehabilitation progress or revising management strategies, a reassessment that does not provide new information is a waste of the valuable, limited resources available for correctional agencies.

A reassessment that can improve the prediction of *imminent* recidivism would be particularly valuable. The circumstances of individuals released from prison often change rapidly, and recidivism shortly after release is common

(Burnett, 2009; Nadesu, 2009). Therefore, short time periods are very meaningful for correctional agencies tasked with supervision and risk management in the community. It would be immensely helpful for those agencies if reassessments could account for the rapid changes that are occurring and indicate how likely those changes are to signal recidivism that may be imminent (Vose, 2016).

In summary, reassessment of dynamic risk and protective factors is already an important part of practice in correctional psychology. Reassessment is central to gauging rehabilitation progress and assisting with ongoing supervision decisions. Correctional practice must be evidence-based to maintain public confidence and produce positive outcomes for everyone involved (Latessa, Cullen, & Gendreau, 2002). Consequently, it is important that reassessments are accurately capturing changes in risk and identifying the individuals whose likelihood of recidivism has changed in a meaningful way. More specifically, to justify their use in practice, reassessments must provide information that was not provided by an earlier assessment. A reassessment score must be a more accurate predictor of recidivism than a previous assessment score. A reassessment score that can improve the prediction of *imminent* recidivism could further enhance practice. Therefore, research examining the predictive accuracy of reassessments of dynamic risk and protective factors is essential.

The Concept of Dynamic Risk Factors

Before looking at the empirical research on dynamic risk and protective factors, first we need to look more closely at why, from a theoretical perspective, these variables are associated with recidivism, and why reassessment of those

factors might improve the prediction of recidivism. In short, what are dynamic risk and protective factors and are they causally related to recidivism or merely correlates? It is important to understand the underlying theory in order to evaluate the way reassessments are used currently, and to have the best chance of improving the way they are used in the future.

A variable must meet three criteria to classify it as a dynamic risk factor (Brown et al., 2009; Caudy, Durso, & Taxman, 2013; Jones et al., 2010). First, as a *risk* factor, the variable must have a significant association with increased recidivism. Second, as a *dynamic* factor, the variable must show significant change over time. Third, combining the previous two criteria, to be a *dynamic risk* factor, there must be a significant association between change and recidivism (i.e., as risk factors increase, the risk of recidivism increases, and vice versa). Take employment as an example. Employment would be classified as a dynamic risk factor if, (a) recidivism was higher for individuals who were unemployed than those who were employed; (b) individuals moved from being employed to being unemployed; and (c) the likelihood of recidivism increased for individuals who moved from being employed to being unemployed. Any test of the concept of dynamic risk factors must start with these three criteria.

Based on this definition, some theorists argue that the existing concept of dynamic risk factors is predictive rather than explanatory (Heffernan & Ward, 2015; Ward, 2016; Ward & Beech, 2015; Ward & Fortune, 2015). They argue that a variable can classify as a dynamic risk factor without a plausible explanation for *why* that variable is statistically associated with recidivism. Consequently, they point out that the existing definition is adequate if prediction of recidivism is the only goal. However, as the previous section showed, dynamic risk factors

are not used only for prediction; they are used to guide several other decisions, most notably intervention. As the theorists note, to assist intervention, risk assessments should contribute to a comprehensive theoretical explanation of criminal behaviour (Mann et al., 2010; Ward & Beech, 2015). Therefore, although the definition of dynamic risk factors does not strictly require an explanatory component, it is useful to determine whether these factors can be used for explanatory purposes, in addition to prediction.

Some theorists have suggested that dynamic risk factors may be “a theoretical dead-end” (Ward & Beech, 2015, p. 100). This argument is based on issues arising from the assumption that dynamic risk factors are *causes* of reoffending. These theorists argue that dynamic risk factors are not causes of reoffending. A cause is a necessary and sufficient condition for an outcome to occur (Kraemer et al., 1997). It is very unlikely that an individual variable will meet these criteria for an outcome as complex as criminal behaviour. It is more likely that criminal behaviour will be *caused* by the interaction of multiple causal mechanisms (Ward & Beech, 2015). Furthermore, just because a variable is statistically associated with recidivism, does not mean it should be viewed as causal; there may be several alternative explanations for that statistical association (Polaschek, 2012; Ward & Beech, 2015). Thus, even though research shows that targeting dynamic risk factors reduces recidivism (Bonta & Andrews, 2016), this evidence is not sufficient for the factors to be considered causes of reoffending.

Theorists have also pointed to the conceptual inconsistency of dynamic risk factors as evidence of their limited theoretical utility. Ward and colleagues (Heffernan & Ward, 2015; Ward, 2016) argue that the concept of dynamic risk

factors is not a single coherent theoretical construct. Rather, it is a composite construct, referring to a mixture of descriptive internal states, external contextual factors, and (possibly) causal mechanisms. Similarly, individual dynamic risk factors (e.g., antisocial cognitions) can be broken down into several separate components (e.g., the distorted beliefs, the problematic thoughts, and the verbalisations of those thoughts; Heffernan & Ward, 2015), demonstrating how individual dynamic risk factors are also composite constructs. A comprehensive explanation of reoffending needs to be clear about any concepts that are used as part of the explanation. These theorists argue that theoretical progress requires the reconceptualisation of dynamic risk factors into a more coherent construct (Ward, 2016).

There are, however, several reasons to think that dynamic risk factors can be used in comprehensive theoretical explanations of reoffending. Firstly, dynamic risk factors were not selected simply because of their statistical association with recidivism; they were drawn from broader theories of human behaviour including psychodynamic theory, social learning theory, strain theory, differential association theory, and other criminological theories (Bonta & Andrews, 2016). For example, the identification of criminal peers as a central dynamic risk factor is a result of the importance of that factor in theories such as differential association (Sutherland, 1947) and social learning (Bandura, 1977), both of which provide explanations for why individuals with pro-criminal peers are more likely to reoffend. Secondly, although it is always right to caution against treating correlation as causation, evidence of a dynamic association between a variable and an outcome (i.e., an association between change in that variable and recidivism) provides stronger evidence that the variable may be

involved in the causal process than static evidence (i.e., a one-off measurement of the variable and outcome). Kraemer et al. (1997) refer to variables that meet the dynamic criteria as *causal risk factors*; use of this term provides an indication of the strength of this evidence. Thirdly, even Ward and colleagues suggest that dynamic risk factors *may* be markers of underlying causal mechanisms (Ward & Beech, 2015), or contain a causal component (Ward & Fortune, 2015). Thus, evidence suggests that dynamic risk factors are likely to have a role in the causal process leading to reoffending, but the nature of that role needs to be more clearly identified.

The existing concepts of *acute* and *stable* dynamic risk factors may help to achieve that goal. Acute factors are defined as variables that can change rapidly and signal imminent reoffending; stable factors are variables that are likely to change more slowly and are more relevant to reoffending over longer time periods (Hanson et al., 2007; Hanson & Harris, 2000; Zamble & Quinsey, 1997). These concepts have considerable theoretical and practical potential. Theoretically, information about the temporal connection of a variable to recidivism (i.e., whether it is more of a stable or an acute factor), could improve what we know about the period leading up to criminal behaviour. The understanding of how different risk factors interact prior to imminent recidivism remains one of the biggest gaps in current theoretical explanations of criminal behaviour (Heffernan & Ward, 2015; Mann et al., 2010). Acute and stable risk factors could help to address that gap. From a practice perspective, Hanson and Harris (2000) suggest that stable factors should be the primary target of interventions, while acute factors need to be monitored and managed because a recent change might suggest reoffending is imminent. Therefore, community

correctional agencies would be able to focus on acute factors in the short-term and stable factors in the long-term.

The utility of acute and stable factors for theoretical development can be illustrated by considering the general personality and cognitive social learning (GPCSL) theory of criminal conduct (Bonta & Andrews, 2016). This theory states that crime is the result of individuals, who, because of their genetic and environmental background, have developed characteristics (including anti-social attitudes and associates) that are supportive of criminal behaviour. Crime occurs when these individuals are presented with situations where, due to their characteristics and the internal states that arise in those situations, they perceive criminal action to be the best or only option. The concepts of acute and stable factors map neatly onto this theory. Beech and Ward (2004) suggested acute factors could be conceptualised as either situational factors (e.g., job loss), or internal states (e.g., anger), whereas stable factors could be relabelled as individual traits or dispositions; Mann et al. (2010) and Thornton (2015) preferred the term *propensities*. Therefore, this theory indicates one way in which stable and acute factors could be integrated into a comprehensive explanation of offending behaviour.

The challenge for utilising these promising concepts is to accurately identify which factors should be classified as acute and which should be classified as stable. Despite some theoretical limitations, the three criteria outlined at the start of this section remain a useful starting point for determining whether a variable is a dynamic risk factor. To determine whether the variable can be further classified as either an acute or stable dynamic risk factor, it is necessary to look more closely at how the variable changes and how change

impacts on recidivism. Change in acute factors should signal an *imminent* risk of reoffending (Hanson et al., 2007; Hanson & Harris, 2000). Change in stable factors must also be associated with a simultaneous change in the likelihood of recidivism, otherwise those variables would not meet the criteria for dynamic risk factors. However, in the short term, existing theory would suggest the association between change and recidivism is likely to be weaker for stable factors than it is for acute factors.

As the defining characteristic that distinguishes acute factors from stable factors, the issue of imminent recidivism requires further examination. The term “imminent” suggests an event is likely to happen within a very short space of time. The concept of acute dynamic risk factors, as introduced by Hanson and Harris (2000), did not include an explicit definition of a short space of time. However, we can infer that it should be equivalent to the amount of time it takes for acute factors to change. For example, if the internal temporary state of anger is an acute dynamic risk factor, theoretically, recidivism will occur during the period in which an individual remains angry. Once the anger has subsided, criminal behaviour resulting from that internal state should be much less likely to occur. Based on this theoretical rationale, imminent recidivism would refer to a matter of hours, days, or weeks, since this is the period over which it is suggested acute factors are likely to change (Hanson et al., 2007; Hanson & Harris, 2000). In contrast, imminent recidivism is only likely to follow an increase in a stable dynamic risk factor (e.g., a favourable shift in attitudes towards offending behaviour) if an acute factor changes simultaneously or shortly after the change in the stable factor (e.g., a situational change presents an opportunity to offend). Hence, the association between change and imminent

recidivism for stable factors might be present but it is likely to be weaker than the association with acute factors.

In summary, we argue that the concept of dynamic risk factors is not a theoretical dead-end. Dynamic risk factors, and specifically the concepts of acute and stable dynamic risk factors, hold considerable potential for advancing theoretical explanations of reoffending. We agree with Ward and Beech (2015, p. 103) that “simply referring to a list of dynamic stable and acute risk factors is insufficient” for developing comprehensive explanations of criminal behaviour and case formulations. Further theoretical development of these concepts is undoubtedly needed. That work should include explanation of how acute and stable factors interact prior to reoffending, how individual stable factors interact with other stable factors, how acute factors interact with other acute factors, and whether the acute and stable categories are meaningful (Mann et al., 2010; Ward, 2016; Ward & Fortune, 2015). However, we argue that this theoretical development must be preceded by better conceptual specification. Before we can discuss the interaction of acute and stable factors, we need to clearly establish which variables should be classified as dynamic risk factors, and whether those variables can be further classified as stable or acute dynamic risk factors, as they are currently defined. Empirical research needs to evaluate the three criteria outlined at the start of this section, and the relationship between the timing of changes in those variables and recidivism, to see if they appear to be stable or acute dynamic risk factors.

The Concept of Protective Factors

The criteria that should be used to classify a variable as a protective factor are less clear than the criteria for dynamic risk factors. Several different

conceptualisations of protective factors have been used within forensic and correctional practice (Ward, 2017). Each one suggests slightly different classification criteria should be used. Some scholars argue that protective factors are not conceptually distinct from risk factors (Baird, 2009; Harris & Rice, 2015). Based on existing evidence, they suggest that protective factors are inverse risk factors. For example, the risk factor *impulsivity* is the same variable as the protective factor *self-control*; it has merely been relabelled and the measurement scale has been inverted. The more common view of researchers in this field is that protective factors are conceptually distinct from risk factors (Farrington, Ttofi, & Piquero, 2016; Lösel & Farrington, 2012; Serin et al., 2016). However, considerable uncertainty remains about *how* protective factors are distinct from risk factors. Various suggestions have been put forward, including that protective factors may interact with risk factors (de Vries Robbé, de Vogel, & Douglas, 2013; MacDonald, 2016; Serin et al., 2016), but no consensus has emerged in this area yet.

It is likely that the term *protective factor*, in its broadest sense, includes several different concepts. For example, some variables considered to be protective factors may be the inverse of risk factors, some may interact with risk factors to reduce risk, and some may directly reduce the likelihood of recidivism, separate from any risk factors. For this reason, the concept of protective factors has also been described as a composite construct (Ward, 2017), analogous to the composite construct of dynamic risk factors. As with dynamic risk factors, clear identification of the different components of this composite construct would lead to advancements in theoretical explanations of reoffending. For example, knowing whether a variable directly reduces the likelihood of recidivism or

whether it is only protective in the presence of certain risk factors would provide clearer information about the role of that variable in the causal pathway leading to reoffending. This important theoretical work is already underway, starting with the introduction of separate terms to differentiate between the different concepts (e.g., "promotive factors", "risk-based protective factors"; Farrington et al., 2016; Jones et al., 2010).

The one feature that researchers agree protective factors must have is an empirical association with decreased likelihood of recidivism (Polaschek, 2017). Any variable that is consistently found to have no association with decreased recidivism should not be classified as a protective factor. There is one other feature that defines protective factors: change over time. In the risk assessment literature, risk factors are almost always classified as being either static or dynamic (and then occasionally further classified as acute or stable). There is at least one risk assessment instrument, the Structured Assessment of Protective Factors (SAPROF; (de Vogel, de Vries Robbé, de Ruiter, & Bouman, 2011), that includes static protective factors—intelligence and secure attachment in childhood—but this instrument appears to be an exception. In the growing literature on protective factors, these variables are almost always assumed to be dynamic factors, without being explicitly described as such.³ Consequently, to classify as a protective factor, a variable must meet the same three criteria as a dynamic risk factor: association with (decreased) recidivism, change over time, and association between change and recidivism.

³ In this thesis, from this point on, whenever we use the term “protective factors”, we are referring to dynamic protective factors. The term “static protective factors” will be used otherwise.

The argument for accurately identifying variables that meet these criteria is essentially the same as for dynamic risk factors. Knowing that changes in certain variables leading up to recidivism are meaningful can provide important information about the causal process that led to that behaviour. For example, a reduction in social support that precedes recidivism would be evidence to suggest that social support had previously been making recidivism less likely. Once again, we must be careful not to suggest that this is evidence of social support *causing* reoffending (Ward, 2017). However, evidence of a dynamic association provides stronger evidence that a variable is involved in the causal process than a static association (Kraemer et al., 1997). Therefore, it would be valuable to conduct research that establishes whether protective factors are (negatively) associated with recidivism, that they demonstrate change over time, and that change is associated with recidivism in the expected direction.

Just as with dynamic risk factors, it would be valuable to further classify protective factors according to their temporal relationship with recidivism. To date, the distinction between stable and acute dynamic variables has not been applied to protective factors. However, existing conceptualisations of protective factors would suggest they are closer in kind to stable factors than acute factors. Ward (2017, p. 26) suggested that “protective factors are the internal and external capacities and personal priorities that enhance individuals’ well-being and reduce the likelihood that they will harm others or themselves”. This terminology is very similar to the language used to describe stable dynamic risk factors (e.g., “propensities”; Mann et al., 2010; Thornton, 2015). Serin et al. (2016, p. 157) described protective factors as a “blend of personal and social capital”. This terminology suggests that these variables build up slowly over

time. The variables are not thought to change rapidly in the way that we would expect from acute factors. It is still possible that some variables, in theory, may be *acute* protective factors. For example, an external event occurring that decreases the likelihood of imminent recidivism such as joining the military (Sampson & Laub, 2005) might meet this definition. Empirical evidence establishing the conceptual similarity between stable dynamic risk factors and protective factors would be beneficial.

In summary, although there remains debate about the concept of protective factors, there are important areas of consistency. Protective factors should be associated with decreased recidivism and are generally expected to behave dynamically. Empirical evidence that can clearly establish which variables meet these criteria, and whether protective factors behave more like acute or stable dynamic factors, would represent an advancement in our understanding of the concept of protective factors. A better understanding of this concept could lead to more comprehensive explanations of reoffending.

Conclusion

In this chapter, we have examined the practical and theoretical reasons for conducting empirical research with reassessments of dynamic risk and protective factors. From a practical perspective, the case for empirical research is simple. Reassessments of dynamic risk factors and (increasingly) protective factors are already an important component of the criminal justice system. Empirical research needs to establish that these risk assessments are accurate. If a reassessment score is used as an indication of future recidivism, and important decisions are made on that basis, then empirical evidence should support that decision. The practical benefits are particularly promising for management and

supervision of individuals in the community, where recidivism can occur at any time. The use of reassessment to track a changing likelihood of imminent recidivism in real time could have a major impact on these areas of practice. Most promisingly, this knowledge could enhance efforts to reduce and prevent reoffending.

Theoretically, dynamic risk factors and protective factors are both concepts with several limitations, but they hold considerable promise for advancing explanations of reoffending. The concepts need to be more carefully defined, and we need to be clearer about which variables should be classified under these conceptual headings. Initial conceptual development should start with the three simple criteria required of all dynamic predictors: an association with the outcome, change over time, and an association between change and outcome. Empirical research looking at whether reassessment of dynamic predictors improves the prediction of recidivism would appear to represent a promising method for evaluating whether variables meet these criteria.

Ultimately, the practical and theoretical rationales outlined in this chapter share the same goal of reducing reoffending. High rates of recidivism are all too common, particularly for individuals who are released from prison back into the community. Careful evaluation of existing practice and theoretical development that enhances explanations for why individuals reoffend would be important steps towards the development of more effective methods of reducing reoffending. Empirical research looking at whether reassessments of dynamic risk and protective factors improve the prediction of imminent recidivism has the potential to achieve both aims.

Chapter 2: The Existing Empirical Evidence on Reassessment

There is strong evidence that dynamic risk factors are empirically associated with recidivism. Several meta-analyses, looking at both individual dynamic risk factors (Bonta & Andrews, 2016; Gendreau, Little, & Goggin, 1996), and risk assessment tools comprised of combinations of dynamic risk factors (Campbell, French, & Gendreau, 2009; Gendreau, Goggin, & Smith, 2002; Hanson & Morton-Bourgon, 2009; van den Berg et al., 2017; Yang, Wong, & Coid, 2010), have found a significant association with increased recidivism. The empirical evidence for protective factors is considerably more limited, but several recent studies have found a significant association between protective factors and decreased recidivism (de Vries Robbé, de Vogel, & de Spa, 2011; de Vries Robbé et al., 2013; de Vries Robbé, de Vogel, Koster, & Bogaerts, 2015; Ullrich & Coid, 2011; Yesberg & Polaschek, 2015).

Most of this research has examined dynamic variables in a static way (i.e., using a single assessment), with a long follow-up period after the assessment. Dynamic variables must be assessed on at least two occasions to establish whether reassessment improves the prediction of imminent recidivism. Ideally, the variables will be assessed on *more than* two occasions. This type of research—empirical research examining dynamic predictors of recidivism assessed in a dynamic way (i.e., across multiple time points)—remains relatively uncommon. In this chapter, we look at the limited existing empirical research that has investigated whether reassessment improves the prediction of imminent recidivism.

The chapter is divided into three primary sections. Each section looks at a different approach that has been taken in examining the predictive value of

reassessment. The results obtained using these different approaches are discussed along with the strengths and weaknesses of each approach. The approaches are presented in the order of increasing importance. After identifying the gaps in the existing research, at the end of the chapter, we set out the first research question and hypotheses for the current thesis.

Pre- and Post-Treatment Designs

The simplest design for examining the impact of reassessment is to compare the predictive accuracy of a single earlier assessment against a single later assessment of the same variables. If reassessment improves prediction, the second assessment should be more strongly associated with later recidivism than the first assessment. Although it is not their primary purpose, psychological treatment programs where risk assessments are completed before and after treatment are well-suited to this kind of research. Studies that have taken this approach have found some support for the hypothesis that reassessment improves prediction, with post-treatment dynamic assessment scores demonstrating stronger or incremental predictive validity over pre-treatment scores (Beggs & Grace, 2010; de Vries Robbé, de Vogel, Douglas, & Nijman, 2015; Hudson, Wales, Bakker, & Ward, 2002; Lewis, Olver, & Wong, 2013; Pettersen et al., 2015; Wakeling, Beech, & Freemantle, 2013). However, the effects have not been strong, and a meta-analysis of attitudes supportive of sexual offending, (Helmus, Hanson, Babchishin, & Mann, 2013), found no significant difference in the predictive accuracy of assessment scores from before and after treatment.

For several reasons, this research can only provide limited guidance about whether reassessment of dynamic factors improves the prediction of imminent recidivism. First, the majority of the studies cited in the previous

paragraph were focused on sex offenders, and mostly on sexual recidivism, minimising the generalisability of those findings to the wider offender population and general recidivism. Second, because these rehabilitation programmes are usually completed in prison or forensic psychiatric institutions, the dynamic factors assessed cannot include situational or contextual variables such as housing or employment that can only be accurately assessed upon release. Research indicates that these variables are related to recidivism (Dickson & Polaschek, 2014; Kroner, Gray, & Goodrich, 2013; Scoones, Willis, & Grace, 2012; Zamble & Quinsey, 1997), so an assessment that excludes these variables is likely to be omitting important information. Finally, like most other dynamic factor research, these studies generally use long follow-up periods of several months or years. Therefore, the studies provide very little evidence either way for the claim that reassessments of dynamic factors can improve the prediction of *imminent* recidivism in the community.

Retrospective Designs

Retrospective interviews or file reviews of recidivists can provide information about dynamic variables that could improve the prediction of recidivism through reassessment. A study by Zamble and Quinsey (1997) is often regarded as a pioneer of this approach. They interviewed 311 men in the Canadian province of Ontario who had been reconvicted and returned to prison within a year of a previous term of imprisonment. The men were asked about their life after release and specifically about the events in the month prior to the recidivism event that resulting in them returning to prison. Results indicated that dynamic variables such as negative cognitions, negative mood, difficult life events, and substance use had all got worse prior to recidivism. Interestingly, the

authors note that there was little evidence to suggest that any of the men were involved in deliberate planning or imminent preparation for the offence. A comparison group of non-recidivists was also interviewed, but with only 36 men, little could be inferred from the comparison.

Hanson and Harris (2000) used a similar methodology in their study of 409 sex offenders in Canada. A combination of interviews with community supervision officers and file reviews was used to provide information about (a) the 6 months prior to recidivism, and (b) the 1 month prior to recidivism, for the 208 offenders who committed a new sexual offence. A comparison group of 201 non-recidivists was also interviewed. An extensive list of dynamic variables was assessed at both time points, with researchers aiming to identify whether the variable had ever been a problem, and whether it had got worse in the month prior to recidivism. Results suggested that many dynamic risk factors—anger, distress, substance use, social problems, victim access, victim blaming, self-management, engagement and co-operation with supervision—got worse in the month before recidivism. It should be noted that some variables (e.g., attitudes to sexual offending) that were strongly associated with recidivism, were assumed to be stable, rather than acute, so there was no measure of whether those variables got worse in the month prior to recidivism.

These studies demonstrate two important points. First, by highlighting dynamic variables that get worse in the period immediately prior to recidivism, they indicate that reassessment of certain variables could improve the prediction of imminent recidivism. Second, they provide evidence that some variables might be able to be classified as either acute or stable, and that those concepts may be useful in predicting imminent recidivism. However, the retrospective nature of

these studies, and the pre-selection of known recidivists, limit how much can be drawn from these findings. The results could easily have been influenced by recall that was inaccurate or subject to confirmation bias. These findings need to be replicated by studies with prospective designs before they should be considered evidence that reassessment of dynamic variables can improve the prediction of imminent recidivism.

Prospective Multi-Wave Designs

The best approach for investigating whether reassessment improves the prediction of imminent recidivism is to examine multiple, prospective assessments of dynamic factors completed during a time when recidivism can occur at any point. There are several benefits to this type of design. A prospective design ensures there is no possibility of pre-selection of the assessments that are known to be the most proximal to recidivism, but the proximity of assessments and recidivism is retained, and will be enhanced if reassessments occur frequently enough. The use of more than two assessments ensures that dynamic variables have more opportunity to demonstrate change. Although only a small group of studies have used this design (Babchishin, 2013; Brown et al., 2009; Greiner, Law, & Brown, 2015; Hanson et al., 2007; Howard & Dixon, 2013; Jones et al., 2010; Lloyd, 2015; Viljoen et al., 2017), many of these studies are very recent, indicating that researchers are increasingly aware of the benefits of this research approach.

One of the first major prospective multi-wave studies with dynamic factors was conducted by Hanson et al. (2007), using a large sample of 997 sexual offenders in Canada and the United States. Dynamic variables, classified as either stable or acute risk factors, were regularly assessed by community

supervision officers as part of their normal practice. The predictive validity of stable scores from an initial assessment was compared with the validity of scores for the same variables assessed 6 months later. Results indicated that for several different measures of recidivism, there was no significant difference in predictive accuracy of the two assessments. The acute variables were assessed considerably more often, with an average of 9.5 assessments per person. This frequent assessment allowed the authors to compare the predictive validity of the assessment score that was most proximal to recidivism (within 45 days) to the next most recent assessment (within 90 days). When these analyses were completed, the authors found no evidence to suggest that a single assessment closer in time was a better predictor of recidivism than a single earlier assessment. The authors suggested that limited change between assessments, for both acute and stable variables, may have impacted the findings.

Babchishin (2013) re-analysed the acute scores from the dataset used by Hanson et al. (2007) and found slightly stronger evidence that reassessment might improve the prediction of recidivism. Reassessment scores of the combined acute variables consistently predicted general and sexual (but not violent) recidivism after controlling for the initial acute scores. However, the scores used for prediction were model-derived and prediction models did not consider the timing of recidivism, so the occurrence of a recidivism event during the assessment period did not result in that individual being excluded from the analysis. Therefore, these results also only provide weak evidence that a reassessment of dynamic variables enhanced the prediction of recidivism compared to an earlier assessment of the same variables.

Research by Howard and Dixon (2013) represents by far the largest study conducted using a prospective, multi-wave design. Their study looked at dynamic risk assessments completed over 4 years for a sample of 196,293 individuals being supervised in the community in England and Wales. The follow-up period was divided into 3-month units; however, most individuals were not reassessed during each 3-month period. Reassessment of the same dynamic variables was found to significantly improve the prediction of violent recidivism, but with such a large sample size, a significant effect was not surprising. The predictive effect sizes observed, which were similar for initial assessment and reassessment, suggested reassessment did not meaningfully improve prediction.

Two studies by Brown et al. (2009) and Jones et al. (2010) provide slightly stronger evidence of reassessment improving prediction. Brown et al. used a three-wave design with 136 men released from prison in Canada. The three comprehensive interview assessments took place within 45 days prior to release, 1 month after release, and 3 months after release, and recidivism was measured for an average of 10 months after release. The static and dynamic variables were all assessed at time 1, and dynamic variables were reassessed twice more unless recidivism—defined as return to prison—had already occurred. A prediction model that incorporated updated dynamic scores using Cox regression with time-varying predictors was a better predictor of recidivism than a prediction model which included only pre-release dynamic scores, but this difference was not statistically significant. A model which included both static and time-varying dynamic factors was found to have the highest predictive validity of the models that were tested.

Jones et al. (2010) examined the same dataset but added an extra wave of assessment at 6 months after release, extended the follow-up time to over 6 years, and used reassessments carried out by community supervision officers in addition to the assessments by the researchers. Once again, the model that incorporated reassessment scores was a better predictor of recidivism than the baseline model, and this time the difference was approaching significance. Interestingly, whether the reassessments were conducted by the supervision officers or by the researchers did not make a significant difference to the predictive accuracy of the models. Together, these studies provide tentative evidence that reassessment of dynamic variables can improve the prediction of recidivism. However, the effects they observed were very small and the dynamic factors selected for inclusion in their prediction models were different for the baseline and reassessment models. Therefore, it remains unclear whether reassessment of the same dynamic variables improves the prediction of recidivism.

Greiner et al. (2015) also used a four-wave design in their study of 497 adult women released from prison in Canada. The follow-up period was divided into four 6-month periods following release, and each woman was required to have an assessment (or have been reconvicted) during each period to remain in the sample. Where an individual had more than one assessment in the same period, the later one was used for prediction analyses. Seven different dynamic variables were assessed: employment, family, associates, substance abuse, community, personal/emotional, and attitudes. Although there was overlap in predictive validity, results indicated a strong trend in favour of reassessment improving prediction, both at the univariate and multivariate level. Similar to

Brown et al. (2009) and Jones et al. (2010), different dynamic variables were included in the baseline and reassessment multivariate models, limiting the validity of these findings.

Finally, Viljoen et al. (2017) conducted a series of analyses exploring the impact of reassessment on the prediction of recidivism. Assessments on two structured risk assessment tools—the Structured Assessment of Violence Risk in Youth (SAVRY), and the Youth Level of Service/Case Management Inventory (YLS/CMI)—were completed by research assistants every 3 months for 1 year with a sample of 156 youth on probation. Two recidivism outcomes—charges for any new offending and charges for new violent offending—were measured during each of the 3-month intervals, and up to 2 years. The findings, for both assessment tools, indicated that reassessment did not enhance predictive accuracy, with initial assessment scores predicting recidivism outcomes with similar accuracy to reassessment scores. Similar to Hanson et al. (2007), the authors noted that the assessment scores may not have changed sufficiently for reassessment to have an impact on predictive accuracy. However, the amount of change in scores was not reported, making it difficult to evaluate that hypothesis.

In summary, a few studies have used a prospective, multi-wave design to examine whether reassessment of dynamic variables can improve the prediction of recidivism. Despite some promising findings, the studies discussed in this section have not found strong evidence to support the value of reassessment. These studies have also not found clear evidence to support the conceptual distinction between stable and acute factors. Studies where these concepts have been measured separately (Brown et al., 2009; Hanson et al., 2007; Jones et al.,

2010) have not observed substantive differences in how reassessment impacts the predictive validity of these variables.

The approaches taken by these previous studies may help to explain the mixed findings. Firstly, assessments in these studies were often several months, or even years apart. To reliably capture change, particularly for acute variables, assessments must occur more frequently. This type of intensive monitoring is ethically and practically very challenging, so longer time intervals between assessments must be accepted,⁴ even if they risk missing changes in some variables (Douglas & Skeem, 2005). However, if it were possible to complete assessments more frequently than has been done in the previous research, that could improve the validity of the findings. Secondly, with the exception of Hanson et al. (2007), the follow-up periods in these studies were also several months or years after the assessments. As we explained in the previous section, the value of reassessment is likely to lie in predicting *imminent* recidivism. Therefore, research with a shorter period between assessments and recidivism would be beneficial.

In addition to issues around timing, each of the studies cited above had several other idiosyncratic features that could explain the variation in results. For example, Greiner et al. (2015) used only women, Hanson et al. (2007) and Babchishin (2013) used only sexual offenders, and Viljoen et al. (2017) focused exclusively on youth. There was also inconsistency in whether studies used

⁴ A study by Vasiljevic, Berglund, Öjehagen, Höglund, and Andersson (2017) examined daily ratings of acute risk factors obtained using automated telephone assessments for 30 days after release from prison, showing that more frequent assessment may not be impossible, but the validity of assessments completed using this approach is questionable, and this type of research must overcome several ethical problems.

assessments completed by the researchers themselves, or assessments completed by practitioners. Clearly, further research is required before any conclusions can be drawn about the predictive value of reassessment.

A large, prospective New Zealand study.

In a recent thesis, (Lloyd, 2015) aimed to address several of the issues with existing prospective, multi-wave, dynamic factor research. His study looked at frequent reassessments of dynamic risk and protective factors conducted by community supervision officers for a large sample of offenders released from prison in New Zealand ($n=3694$). The dynamic variables assessed were divided into subscales of stable risk, acute risk, and protective factors. Two aspects of his design were particularly unique from previous research. First, reassessments were more frequent than they have been in previous studies. An assessment was recorded in every week of the 1-year follow-up period, or at least up until recidivism occurred. Second, recidivism was defined as a new conviction where the offence date occurred within 6 weeks of the most recent assessment. If a new offence occurred more than 6 weeks after the final assessment on record, the individual was not classified as a recidivist. Therefore, this research was focused on predicting *imminent* recidivism.

The primary research question in Lloyd's study was whether reassessment of dynamic variables improves the prediction of short-term recidivism. He compared the predictive accuracy of a single baseline score from an assessment completed around the time of release from prison, to a predictive model that used reassessment scores. The results provided mixed support for the predictive value of reassessment. Although the reassessment scores consistently demonstrated significant incremental predictive validity over the

baseline scores, effect sizes were small and confidence intervals showed the predictive accuracy of the two models overlapped for most of the follow-up period. The study found some evidence that reassessment might improve prediction more for acute risk factors than for stable risk or protective factors; the difference in predictive accuracy of the baseline and reassessment models was greater for a set of putatively acute factors than for a set of stable or protective factors.

Despite the mixed findings, Lloyd's research represented an advancement on previous research in this area. The increased frequency of assessment and the restricted definition of recidivism provided a more robust test of the relevant concepts and more closely matched existing practice than any previous research. The use of assessments conducted by supervision officers in their regular practice provided the study with strong ecological validity. The statistical technique used by Lloyd—a discrete-time hazard model—was also unique. Although this technique is very similar to the Cox regression survival analysis with time-varying predictors model used by several previous studies in this area (Brown et al., 2009; Greiner et al., 2015; Howard & Dixon, 2013; Jones et al., 2010; Viljoen et al., 2017), the discrete-time hazard model has several advantages over the Cox model such as the ability to more easily test model diagnostics. The effect size estimates used by Lloyd—Xu and O'Quigley's (1999) R^2 and Heagerty and Zheng's (2005) c -index, which are described in more detail in the next chapter—were also relatively unique in this area and have the potential to provide easily interpretable measures of the extent to which reassessment improves the prediction of recidivism.

The Current Research

Existing research indicates that reassessment of dynamic variables *may* improve the prediction of imminent recidivism. While pre-post and retrospective designs can provide some insights, the best way of reliably addressing this question is through research utilising a prospective, multi-wave design. The few studies using this approach have found some evidence that reassessment improves prediction, but these findings have been inconsistent and effect sizes have been small. For several reasons, the research by Lloyd (2015) represents an advancement in this field that could better evaluate the predictive value of reassessment and serve as a more valid test of the concepts of dynamic risk and protective factors. However, Lloyd's results, while stronger than previous studies, were also inconclusive. More research is needed using the unique framework he laid out, to see whether consistent results will be obtained using the same approach with a different sample.

For that reason, the first aim of this thesis was to complete a replication of the work by Lloyd (2015). We obtained a large dataset of prospective, repeated assessments of dynamic risk and protective factors completed by community supervision officers working with individuals immediately after release from prison in New Zealand. These assessments were completed using the same risk assessment tool—the Dynamic Risk Assessment for Offender Re-Entry (DRAOR; (Serin, 2007; Serin et al., 2012)—used in Lloyd's study, ensuring that the measure of dynamic risk and protective factors was consistent. The dataset is described in extensive detail in the following chapter.

Consistent with previous research, we expected that incorporating reassessments would significantly improve the prediction of imminent

recidivism. As we described in Chapter 1, dynamic risk and protective factors are conceptualised to be changeable, and for changes to be associated with changes in the likelihood of recidivism. Reassessment must improve the prediction of recidivism for the variables being measured to meet these criteria. The updated information provided by reassessment about an individual's functioning and circumstances should lead to enhanced prediction. However, based on the existing research, observed effects were expected to be relatively small.

This study also set out to test whether the impact of reassessment on predictive accuracy was consistent for stable, acute, and protective factors. Theoretically, reassessment of acute factors would be expected to have the largest impact on predictive accuracy, since these are the variables expected to change most rapidly and frequently (Hanson et al., 2007; Hanson & Harris, 2000). Change on acute factors is theorised to signal imminent recidivism; changes in stable or protective factors should also be associated with imminent recidivism but not to the same extent. The limited existing research examining this point has found minimal evidence to support these conceptual distinctions. However, the strongest evidence comes from Lloyd (2015), who found that reassessing acute factors improved the prediction of imminent recidivism more than stable and protective factors. Given that we were replicating Lloyd's approach, it was expected that we would observe a similar difference.

Chapter 3: The Dataset

This thesis is a replication and extension of the work by Lloyd (2015). In order to best replicate that prior research, this thesis used a very similar dataset. The same dynamic risk assessment tool— the Dynamic Risk Assessment for Offender Re-entry (DRAOR; Serin, 2007; Serin et al., 2012)—was used as the measure of dynamic risk and protective factors, and the same type of sample was used: individuals released into the community after a long period of imprisonment in New Zealand. The New Zealand Department of Corrections extracted the initial data from existing electronic file records, which included DRAOR assessment scores, a record of reconvictions, demographic information, and information about the prison sentence that was served prior to release. Ethical approval for the release of the data and for the research more generally were provided by both the Department of Corrections Research and Evaluation Governance Committee and the Victoria University of Wellington School of Psychology Human Ethics Committee. From the raw data, a dataset was selected that was used for the analyses completed in the following four chapters.

In this chapter, we provide a detailed description of that dataset. We start by presenting relevant background information to help readers understand the context in which the data were collected. Information is provided about (a) the correctional environment (i.e., New Zealand), (b) the dynamic assessment measure used, and (c) the reasons for choosing the particular sample group. Following that background information, the rest of this chapter describes how the final dataset was obtained from the raw data. We provide a step-by-step breakdown of the decisions made about the raw data, the main statistical technique—discrete-time hazard models—used for analyses throughout the

thesis, the reasons for choosing that statistical technique, and the steps taken to structure the dataset in a way that allowed that technique to be used.

Correctional Environment

Parole.

In New Zealand, at the discretion of the New Zealand Parole Board (NZPB), individuals who are sentenced to two years or more in prison can be released *on parole* prior to the end of their sentence. Unless a longer minimum parole period was imposed at sentencing, prisoners are eligible to be released after they have served one third of their sentence, but typically are not released until much later (Gluckman, 2018). Individuals who receive an indefinite prison sentence—either a life sentence or preventive detention—become eligible for parole once they have served their minimum parole period, which is specified by the court at their sentencing.

Upon release, all parolees must adhere to a set of standard release conditions, which are set out in legislation (Parole Act 2002, s14). The standard conditions include restrictions on where parolees can live and work, whom they can associate with, and how frequently they must report to a community supervision officer.⁵ Reporting frequency is determined by the supervision officer and generally ranges from biweekly to monthly, depending on factors such as the parolee's risk level and compliance. The NZPB can impose other

⁵ In New Zealand, supervision officers are referred to as “probation officers”. However, different names for this role are used internationally. In some jurisdictions, probation officers only supervise individuals serving community sentences, whereas individuals released from prison are supervised by “parole officers”. To avoid confusion, the broader terms “community supervision officers” or “supervision officers” are used throughout this thesis.

special conditions such as restrictions on certain areas the parolee is able to visit (Parole Act 2002, s15).

Some individuals serving sentences of two years or more do not technically get released *on parole*. The NZPB may decide that an individual poses an undue risk to the community and thus cannot be released prior to the end of his or her sentence. However, barring any additional order being made (e.g., a Public Protection Order, which can result in further detainment in a secure facility), those individuals will still be released once they have served their full sentences, and will be subject to the same standard release conditions—and special conditions, if applicable—as those who were released early.

Regardless of whether release occurs at sentence end or earlier, release conditions must be imposed for at least 6 months. For individuals released prior to their sentence end date, this period will often be longer than 6 months because it will include the remaining time that they would have spent in prison plus the additional 6 months. Up until their sentence end dates, individuals released early can be recalled to prison for failing to adhere to their release conditions (i.e., without having been convicted of an offence). Individuals who are released at the end of their sentence can only return to prison if they are convicted (or remanded) for a new offence; they cannot be recalled for failing to adhere to their release conditions. Anyone who receives an indefinite imprisonment sentence has release conditions imposed for life and can be recalled to prison at any point for failure to adhere to those conditions.

For the purposes of this thesis, both individuals who were released at the very end of their prison sentence *and* those released prior to that point are considered to be *on parole* and referred to as *parolees*.

Risk assessment.

Static risk.

All parolees in New Zealand are assessed on a static risk instrument known as the Risk of Reconviction*Risk of Re-imprisonment (RoC*RoI; Bakker, Riley, & O'Malley, 1999). Scores on this measure represent the likelihood of someone returning to prison within five years (e.g., RoC*RoI score of 0.5 = 50% chance of returning to prison). The RoC*RoI is scored automatically by a computer and is generally only updated following a new conviction and sentence. Important variables include age and several different measures of criminal history. Previous research has indicated that RoC*RoI scores predict recidivism with a high level of accuracy (Bakker et al., 1999; Nadesu, 2009).

Dynamic risk.

All parolees in New Zealand are regularly assessed on a dynamic risk assessment tool known as the Dynamic Risk Assessment for Offender Re-entry (DRAOR; Serin, 2007; Serin et al., 2012). The DRAOR consists of 19 items that are divided into stable risk (6 items), acute risk (7 items), and protective factors (6 items). The six stable items are peer associations, attitudes to authority, impulse control, problem-solving, sense of entitlement, and attachment with others. The seven acute items are substance use, anger/hostility, opportunity/access to victims, employment, interpersonal relationships, and living situation. The six protective items are responsivity to advice, prosocial identity, high expectations (for reintegration and rehabilitation success), costs/benefits, social support, and social control.

A three-point scoring format (0, 1, 2) is used for each individual item. For the risk subscales (stable and acute), a score of 0 indicates the item is “not a

problem”, 1 indicates a “slight or possible problem”, and 2 indicates a “definite problem”. The scoring is reversed for the protective items, with 2 indicating that the item is a “definite asset” and a 0 indicating that the item is “not protective”. Thus, scores on the stable and protective subscales can range from 0 to 12; scores on the acute subscale can range from 0 to 14. Higher scores on the risk subscales are expected to indicate a higher likelihood of recidivism; higher scores on the protective subscale are expected to indicate a lower likelihood of recidivism.

The DRAOR is completed by community supervision officers, who are required to assess parolees on the DRAOR after every reporting session. For parolees, this requirement means that the first DRAOR assessment should be completed immediately following the first supervision session after release. In some cases, supervision officers meet with prisoners shortly *before* release and complete a DRAOR assessment following that meeting. Also note that while it is recommended that supervision officers assess the *acute* items after every session, officers are only instructed to keep a “watching brief” on the stable and protective items (Serin et al., 2012, p. 4). Scores are entered on a computer and recorded on the parolee’s electronic file.

Psychometric Properties of DRAOR Data in Previous Research

Multiple different research projects have now investigated the psychometric properties of the DRAOR. This body of research includes several theses and draws on data provided by correctional agencies in New Zealand and Iowa. The two major psychometric properties relevant to the current research are factor structure, since all analyses in this thesis will be conducted using the three subscales rather than individual items, and predictive validity. Intra-

individual change over time is also an important psychometric property and one that is necessary for the substantive analyses in this thesis. Research examining that property is explored in more detail in later chapters, where the results are compared with Lloyd (2015), the other major study to examine change in DRAOR scores over time.⁶

Factor structure.

Multiple studies have investigated the factor structure of the DRAOR to see if it accurately corresponds to the three named subscales: stable risk, acute risk, and protective factors. A principal component analysis (PCA) by Hanby (2013) found general support for the three subscales, although there was some cross-loading between stable and acute items. Yesberg and Polaschek (2015) also ran a principal component analysis, following a confirmatory factor analysis that found the three-subscale structure was not a good fit for their data. The PCA indicated a better model fit was obtained when the acute items were divided into internal and external subscales, creating a four-subscale structure, with the stable and protective subscales remaining largely intact. Finally, Chadwick's (2014) exploratory factor analysis found support for a two-factor solution comprising either risk—stable and acute—or protective factors. These mixed findings provide some support for the three-factor structure but suggest another factor analysis with the current dataset is warranted.

Each of the above findings was made using DRAOR scores taken from shortly after release from prison or just after the beginning of the supervision

⁶ Hanby (2013) also completed several analyses looking at change in DRAOR scores over time but that study used the same dataset as Lloyd (2015), so we have chosen to focus primarily on a comparison with Lloyd's results.

period. Lloyd (2015) argued that factor invariance over time or *longitudinal measurement invariance* is a more important measure of factor structure, because it shows that the DRAOR is being scored consistently across time. One method of testing consistency is to run analyses using the same data measured at different time points (i.e., a baseline DRAOR and a DRAOR at 6 months). Longitudinal measurement invariance is considered in more detail in the next chapter.

Predictive validity.

Every existing study using DRAOR data has examined some measure of predictive validity. Overall, this research has established that the DRAOR subscale scores can distinguish between eventual recidivists and non-recidivists with a moderate degree of accuracy. These studies have also demonstrated that DRAOR scores can predict recidivism for a range of specific sub-groups including high-risk parolees (Yesberg & Polaschek, 2015), female offenders (Scanlan, 2015), youth offenders (Ferguson, 2015), sex offenders (Smeth, 2013) and general offenders of all risk categories, whether they are on parole (Hanby, 2013), or serving a community sentence (Chadwick, 2014).

The most common method for testing predictive validity has been to take scores from a single assessment and see how well those scores distinguish between offenders who will eventually recidivate and those who will not recidivate. However, there have been two important variations in methodology that appear to have impacted on the results: recidivism outcome and follow-up period. DRAOR scores are a good predictor of general recidivism (any new conviction including technical violations). The evidence is slightly weaker for the prediction of only technical violations, or of new convictions excluding technical

violations (Chadwick, 2014; Ferguson, 2015; Hanby, 2013; Scanlan, 2015; Yesberg & Polaschek, 2015), and DRAOR scores do not appear to be a reliable predictor of sexual recidivism (Smeth, 2013).

The second important variation in methodology is the amount of time between the DRAOR assessment and the outcome. Most studies have used roughly the same follow-up period: an average of 2 years for general recidivism, and 9 months for technical violations (offenders can only receive a technical violation while they are on supervision, hence the shorter follow-up period). One major exception is Yesberg and Polaschek (2015), who used a fixed 6-month follow-up period. However, while the follow-up lengths have been largely consistent, the point at which the DRAOR assessment is taken from has varied across studies. Most studies take scores from an assessment made shortly after the offender enters the community. This marks the beginning of the period in which the offender becomes at risk of reoffending. However, three studies (Ferguson, 2015; Hanby, 2013; Scanlan, 2015) have utilised the fact that the DRAOR is frequently reassessed, and have looked at the predictive validity of the score closest in time to a new offence, or in the case of a non-recidivist, the final assessment before the end of the supervision or follow-up period. Compared to studies using a DRAOR assessment from the beginning of the follow-up period, these *proximal* assessment scores have generally demonstrated a stronger association with recidivism. These studies provide some evidence that *reassessing* the DRAOR subscales can improve the prediction of recidivism but are limited by the fact that the proximal assessment was chosen retrospectively (i.e., tested because it occurred immediately prior to recidivism). The analytical

technique used by Lloyd (2015) and replicated in this thesis provides a more valid approach for examining the value of reassessment on predictive accuracy.

Previous research suggests the three subscales predict recidivism with a comparable level of accuracy. A couple of studies (Ferguson, 2015; Hanby, 2013) have found that scores on the acute subscale are better predictors of recidivism than scores on the other two subscales. Interestingly, both of these studies used proximal assessment scores to examine predictive validity. These findings support the hypothesis that acute variables are expected to be most strongly associated with imminent recidivism (Hanson et al., 2007; Hanson & Harris, 2000). However, the differences between subscales have been too small to draw definitive conclusions about differences in predictive accuracy.

Choosing the Sample

As highlighted earlier, the sample chosen for this thesis was very similar to the sample used by Lloyd (2015). Like Lloyd, we decided to focus solely on parolees, even though individuals serving other community-based sentences in New Zealand are also regularly assessed on the DRAOR. The transition back into the community after serving a long prison sentence is extremely difficult for most people, with very high rates of recidivism (Nadesu, 2007). Prioritising parole-based research can improve our understanding of what leads people to reoffend during this period and assist prevention efforts, in addition to expanding the understanding of the nature of dynamic risk and protective factors. The jurisdictional requirement that everyone must be on parole for at least six months also ensures that everyone in the sample can be compared for a minimum period of time.

The sample chosen for the current research differed from Lloyd's (2015) sample in three important ways. Firstly, his sample was drawn from individuals released from prison between April 2010 and April 2012. In contrast, we chose to use individuals released from prison between September 2012 and August 2015, the most recent 2-year period up to the start of this research. Drawing from a later time period ensured the samples would be different (although it is possible that some men in Lloyd's study were also included in our study following a different prison sentence). The later time period may have resulted in an increase in the integrity of the DRAOR assessments. Lloyd's sample used assessments from immediately after implementation of the DRAOR in New Zealand in April 2010. Since the assessments used for the current research were drawn from more than two years after that point, increasing familiarity with the tool may have led to improvements in the integrity of assessment (Flores, Lowenkamp, Holsinger, & Latessa, 2006). However, the possibility of fidelity drift over time should also be acknowledged (Gearing et al., 2011), and may have offset or minimised any benefits from increasing familiarity.

Secondly, Lloyd's sample included parolees regardless of their static risk level, whereas we chose to examine only individuals considered high-risk on the static risk measure, defined as those with a RoC*RoI of 0.65 or higher. There were several advantages and one major disadvantage of choosing to study only high-risk parolees. The major disadvantage was that it was likely to reduce the discriminative validity of the risk instruments. Evidence shows that some measures of predictive accuracy are reduced when the range of scores is limited (Hanson, 2008), as was the case in this research. However, the advantages of using only high-risk parolees outweigh that disadvantage. By definition, high-

risk parolees should have higher recidivism rates than parolees from lower-risk groups. This research investigates *imminent* recidivism, and thus, to observe any effect of reassessment on predictive accuracy, it is preferable to have a sample where there is a high rate of imminent recidivism. Also, as Lloyd noted, high-risk offenders are generally required to report for supervision more frequently than lower-risk offenders and thus are assessed more frequently on the DRAOR. By having more frequent DRAOR assessments, it was more likely that we would capture change in the dynamic risk and protective factors being measured.

Thirdly, Lloyd's sample included men *and* women—albeit only 7% were females—whereas we chose to include only males in our sample. Although research suggests that the DRAOR is a valid tool for use with women (Scanlan, 2015), it was decided to limit this sample to men to enable easier comparison with a concurrent programme of research examining *high-risk men* on parole in New Zealand. That project, known as the New Zealand parole project (NZPP; Polaschek, Bell, Casey, Dickson, & Yesberg, 2014), is a longitudinal study of nearly 300 men on parole in New Zealand. The project involved the collection of extensive quantitative and qualitative information through interviews shortly before and after release, as well as through file reviews. By also focusing on high-risk men in New Zealand in our study, it was hoped that we might be able to draw on some of the findings from the NZPP to add explanatory depth to the interpretation of the purely quantitative analyses conducted in this thesis.

Moving from the Raw Data to the Final Sample

The raw data provided by the New Zealand Department of Corrections included all DRAOR assessment scores, all reconvictions, and demographic and other relevant information (e.g., criminal history) for all men with a RoC*RoI of

0.65 who were released on parole in New Zealand between September 2012 and August 2015. That initial data included records for 1012 men and a total of 44,852 DRAOR assessment scores. Several features of that raw data made it unsuitable to use for the analyses in this thesis. A large number of assessments in the raw data were duplicates, several men had never been assessed on the DRAOR, and file review indicated a few men were not on parole. There was also a long time between the release date and the date of the first DRAOR assessment for many men, making it unclear if the assessment had occurred during the parole period (as opposed to during a different sentence).

In this section, we highlight the features of the raw data that were unsuitable for analysis and describe how those issues were addressed, leaving us with the final sample, which is described in the next section. We describe how we chose to define the *baseline* assessment score and *recidivism*. Those decisions had implications for which assessments from the raw data were retained for the analyses (e.g., assessments occurring prior to baseline, after a recidivism event, or after the end of the 6-month follow-up period were excluded from the analyses). These steps are described in detail below and a visual summary of the information in this section is presented in Figure 3.1.

Step 1: Men released more than once during the follow-up.

The assessments in the raw data were for all men released on parole between September 2012 and August 2015. However, 96 of the 1012 men were released twice, and 9 men were released three times during this period. Multiple releases might occur either because the man had been recalled to prison on his current parole sentence and re-released, or because he had been reconvicted and re-imprisoned for a new offence and then re-released after that new sentence.

Figure 3.1. Data flow chart showing decision steps made to obtain final sample from raw data.

In Lloyd's (2015) sample, there were also several individuals who had been released more than once. He chose to retain assessments following a second or third release date, resulting in the same individual being included multiple times in his dataset. Thus, sequences of assessments, rather than individuals, served as his unit of analysis. He noted that normally it would be problematic to include non-independent sequences. However, this problem was largely mitigated by the fact that the key difference between two release dates for the same person—the number of prior incarcerations and reconvictions—could be controlled for by using an updated static risk score.⁷ Importantly, his dataset included an updated static risk score for each different release date.

The raw data for the current research did not include updated static risk scores; first and subsequent releases had the same static risk score. Also, 78 out of the 114 multiple releases had no record of receiving an imprisonment sentence between releases. In some of these 78 cases ($n = 19$), it appeared that the second release followed a period on remand, because an offence was committed between releases, but conviction for that new offence did not result in imprisonment. In the absence of additional information, it would appear that the remaining 59 men must have been either recalled to prison or remanded in custody (and then re-released) for an act for which, ultimately, they were not convicted. Overall, the lack of clarity about why so many of this group had returned to prison provided further justification for excluding multiple release dates from the analysis. Although Lloyd ultimately included them, the fact that excluding multiple release sequences from the analysis made no substantial

⁷ Prior convictions and incarcerations are both important variables in the RoC*RoI calculation.

differences to his results or conclusions also added weight to our decision to exclude them in this research.

For all these reasons, we decided to remove from the dataset any release dates that were not first in time for the same man. At this point, a large number of duplicated assessments were also removed from the dataset; for technical reasons, when the data were extracted, the exact same assessments were repeated twice in the initial dataset where the same man had been released twice and repeated three times where a man had been released three times.

Step 2: Participants with no DRAOR assessments.

Seven of the initial 1012 men were removed from the dataset because they did not have any DRAOR assessments. Six of these seven men committed an offence very soon after release and were sentenced to imprisonment. It is likely that a DRAOR assessment was not completed because these men never reported to a supervision officer due to being on remand, awaiting trial, or sentence for a new offence. It is not clear why the seventh man did not have any assessments.

Step 3: Errors with participants and DRAOR assessments.

Three men and all their DRAOR assessments ($n = 180$ assessments) were removed from the dataset because they were not actually on parole. Reviews of their files showed that one man was transferred onto extended supervision immediately upon release, one was released at sentence end date with no parole due to a successful appeal of his sentence and the final man had no record of serving parole. There were also three assessments where the date was incorrect (e.g., year 0201) and one assessment that had no date provided. These four assessments were removed from the dataset.

Step 4: Defining baseline.

A baseline level of dynamic risk and protective factors needed to be defined to answer the research question about the extent to which reassessment scores are better than a baseline DRAOR score at predicting imminent recidivism. This step was also essential for ensuring that only assessments associated with the relevant release date (as opposed to an earlier or later sentence) remained in the dataset.

For 318 men, a DRAOR assessment was completed on the day of release, providing an ideal baseline measure. For the rest of the men, an assessment needed to be chosen to act as the baseline measure. This research used the same definition of baseline as Lloyd: if there was no assessment on the day of release, the baseline score was drawn from the period of up to 30 days before release. If there was also no assessment in that period, the baseline score was drawn from an assessment no more than 28 days after release. Ultimately, the baseline assessment was drawn from prior to the release date for 58 men, and after release for 590 men.

Lloyd provided extensive reasoning for choosing this definition of baseline. Briefly, an assessment may be completed in advance of release by a supervision officer preparing for release. Therefore, assessments from prior to release are suitable to use as a baseline measure. However, if the assessment is too far before release, it is possible that the assessment was completed during a previous period of supervision or in preparation for an earlier release that did not eventuate. One month, or 30 days, prior to release was deemed to be an acceptable threshold for determining that an assessment was sufficiently proximal to release.

All DRAOR assessments prior to 30 days before release were removed from the dataset. Where there was an assessment on the day of release, all assessments prior to release were removed even if they were within 30 days ($n = 33$ assessments). For the single parolee who had two assessments prior to release and within 30 days of release, the assessment closest to release was chosen and the other, earlier assessment was removed. As mentioned earlier, the raw data included all DRAOR assessments on record since 2010, sometimes as many as four or five years before the single relevant release date used in this research (i.e., the first release date between September 2012 and August 2015). Therefore, this step resulted in a substantial reduction in the number of assessments remaining in the dataset. Once those assessments were removed, there were 3 men who did not have any assessments left in the dataset, suggesting all the assessments recorded for those men must have been from an earlier sentence. These men were all convicted of an offence shortly after the release date provided and subsequently reimprisoned. Thus, they were probably not seen by a supervision officer and not assessed on the DRAOR because they were already back in prison.

Where there was no assessment either before, or on the day of release, an assessment from within 28 days after release was used as the baseline. The 28-day (or four-week) threshold ensures that the baseline measure is not too long after release. From a practical perspective, there are several reasons why the first assessment might not be completed until a later date. Parolees are required to report to a community supervision officer within 72 hours of release. Failure to report in that time can result in being breached and potentially a rapid return to prison. However, men who were reconvicted *before* their first DRAOR

assessment were removed from the dataset at step 6 (see below). For the other men, whose first DRAOR assessment was more than 72 hours after release, they must not have been breached for some reason, possibly either because they had a legitimate reason for not reporting immediately, or because the supervision officer failed to complete an assessment after their first meeting.

Eighteen men did not have any assessments within 28 days of release (i.e., there was a gap of more than 1 month after release until the man was first assessed). In most cases ($n = 12$ men), the gap appeared to be the result of a post-release conviction for an offence committed prior to starting the prison sentence for the index offence. It is likely these men were back in custody shortly after release: hence the absence of DRAOR assessments. Two men were breached shortly after release, raising the possibility they had not reported to their supervision officer for the first month, although this is speculation. It was unclear why the other four men had no DRAOR assessments within the first 28 days. In any case, without a relevant assessment to use as a baseline score, these men and their assessments had to be removed from the dataset.

Step 5: Defining recidivism.

Recidivism was defined as the first offence committed after release that resulted in a conviction. All types of offences, including breaches of the conditions of parole, were included in this definition. However, since it was necessary to match the date of the new offence to the date of a DRAOR assessment, any new offence where a date was not provided was excluded from the dataset. Of 5772 recorded recidivism events in the initial dataset, there were 957 with no offence date. Of these 957 offences, 951 were defaults on overdue fine payments. This particular offence generally involves the court replacing

outstanding debts with an equivalent number of hours of community work, so even the 'default on fines' convictions which do have an offence date are not meaningful (i.e., they do not specify that an individual has committed an offence on that date). For this reason, all reconvictions for defaults on fine payments were excluded from the dataset.

There were several instances where individuals were convicted *after release* for an offence committed *before* going to prison on the sentence they were subsequently released on parole. Unless the new offence resulted in a sentence of imprisonment, these recidivism events were excluded, on the basis that the man had not actually committed a new offence since his release.⁸ It is possible that learning of these prior offences may have led to an increase in risk and decrease in protective factor scores, because supervision officers believe the offender should now be considered higher risk. However, the evidence did not support this suggestion: in the week immediately after conviction for these offences, DRAOR scores did not change in a way that was inconsistent with the average amount of change observed in other weeks.

Step 6: Ensuring prospective prediction.

It was important to ensure that the assessments used to predict recidivism had occurred prior to the recidivism event (i.e., that prediction was *prospective*). If there was no assessment prior to a new offence, there was no possibility of prospective prediction. There were eight men whose first new offence was before their first DRAOR assessment ($M = 8$ days between offence

⁸ In the three instances in which a new offence committed before release did result in a new sentence of imprisonment, these offences were also later excluded because they were not sufficiently proximal to a DRAOR assessment under the specified proximity described in step 8.

and assessment). There were another seven men whose first DRAOR assessment was on the same day as their first new offence. Where this occurred, to avoid doubt about whether the assessment or the offence occurred first on that day, these men were removed from the sample. For recidivists with a DRAOR prior to the date of a new offence, all DRAOR assessments from after the date of their first new offence (except defaults on fines) were removed from the dataset.

Step 7: Gaps of more than six weeks between assessments.

In the raw data, there were often long gaps in time between two consecutive assessments for the same individual. The two most likely causes of these gaps were a supervision officer failing to complete an assessment or the parolee not reporting for supervision. In the latter situation, a supervision officer is effectively relying on the assessment from the previous week to determine the parolee's current risk of recidivism. This research attempted to recreate that process by carrying forward, or *imputing*, prior assessment scores where there was an extended gap between assessments. (The details of how imputation was performed is described later in this chapter.) However, given the expectation that dynamic risk and protective factors will change over time, there must be a limit on the period of time for which assessments can be imputed.

The threshold used in this study—identical to the one used by Lloyd (2015)—was a gap of more than 6 weeks between assessments.⁹ In practice, parolees need to report to a supervision officer at least once a month and usually

⁹ All references to a *gap of more than 6 weeks* refers to the gap between discrete weeks since release, as opposed to the number of days. For example, an assessment 4 days after release and another assessment 48 days after release are not *more than 6 weeks* apart because they occur in the first week and the seventh week after release, despite the fact that they are more than 42 days apart.

at least fortnightly for high-risk parolees. Therefore, when there has not been a DRAOR assessment for more than six weeks it suggests that there has been some disruption to the normal supervision schedule. In these circumstances, an assessment from more than 6 weeks ago seems unlikely to be a good measure of an individual's current risk level. Thus, we decided to remove all assessments following a gap of more than 6 weeks. Similar to men who had a final DRAOR assessment prior to the end of the full follow-up period *and* were not reconvicted of a new offence, these men were considered to be *censored* following the final assessment prior to the gap.

Ninety-four men had at least one gap of greater than 6 weeks between assessments. The average size of these gaps ($M = 19.6$ weeks) suggested that there had been a serious disruption to the supervision schedule and provided further justification for truncating the assessments at the final assessment prior to the gap. In about half of these cases the disruption appeared to be the result of recidivism. The cause of the other disruptions was unclear. All assessments following the first gap of more than 6 weeks ($M = 14.4$ assessments for the 94 men) were removed from the dataset.

Step 8: Truncating the follow-up.

Lloyd (2015) truncated the follow-up period to 52 weeks (i.e., assessments more than 52 weeks after release were removed from the dataset). However, Lloyd noted that significant caution needed to be exercised when interpreting the 52-week period as a single whole. Results obtained from the first 6 months appeared to differ from the results obtained from the second 6 months. This discrepancy seemed to be caused by the inclusion of high-risk men who were not released on parole until the very end of their sentence and thus

were only under supervision for the minimum 6 months. Once these men were no longer included in the analyses (i.e., after 26 weeks), the interpretation of the results appeared to change. Since the sample for the current research only included high-risk men, and almost half of those men were released right at the end of their sentence, the risk of a disjointed interpretation of the findings was greatly increased. Therefore, the decision was made to truncate the follow-up for assessments to 26 weeks.

Truncating to 26 weeks was also necessary for the purposes of the analytical technique used in this research. In the analytical technique chosen, a recidivism event needs to occur during every discrete time period. As is described below, the dataset in this research was broken into weekly time periods. During each of the first 26 weeks, there were at least 5 new recidivists, but from weeks 27 to 52 there were several discrete weeks with no recidivists.

The Final Sample

Taking the eight steps described in the preceding section left us with a dataset comprised of all men with a RoC*RoI of 0.65 who were released on parole in New Zealand between September 2012 and August 2015. To be included in the final sample, the men had to have at least one DRAOR assessment completed either within 30 days prior to release or 28 days after release; the assessment also had to be completed prior to any new offence that was committed after release (and resulted in a conviction). The dataset included all DRAOR assessments completed within 6 months of release for these men, provided those assessments did not occur after a new offence, or more than 6 weeks after the next most recent assessment for the same individual. Further details about the components of the dataset are provided below.

Participants.

The final sample consisted of 966 men with an average age at release of 31 years old ($SD = 8.6$, range = 17-60) and an average RoC*RoI score of 0.76 ($SD = .07$), indicating a 76% likelihood of returning to prison within five years. The majority of the sample was recorded as being Maori (64.5%) or New Zealand European (27.3%), with a small proportion recorded as Pasifika (7.6%), Asian (0.3%), or Other (0.3%).

Seventeen men were serving a life sentence; the other 949 men had been sentenced to an average of 3.4 years in prison ($SD = 2.1$). For those 949 men with a finite end to their parole period, the average length of parole, as defined in this research (i.e., the duration of the release conditions), was approximately 10 months ($M = 292.3$ days, $SD = 164.4$). Within that larger sample were two distinct groups: the 401 men who were released within a week of their sentence end date and thus were subject to release conditions for almost exactly 6 months ($M = 183.1$ days, $SD = 10.2$); and the other 548 men who were released prior to their sentence end date and had a significantly longer parole length ($M = 372.3$ days, $SD = 177.7$).

The most serious offence for which the men had been imprisoned on their current sentence was primarily either non-sexual violence (46.6%) or a property offence (34.0%), with a smaller proportion imprisoned for drug-related offences (8.5%), sexual violence (5.7%), offences against justice (e.g., fraud; 3.8%), and traffic offences (1.4%).

Recidivism.

Over an average follow-up of 656 days ($SD = 266$, range = 19-1064), 670 of the 966 men (69.4% recidivism rate) were reconvicted for an offence

committed after release. However, because the follow-up period for DRAOR assessments was truncated to 26 weeks, the follow-up period for recidivism also needed to be truncated. For the purposes of the predictive analysis, recidivism needed to be lagged from the DRAOR assessments by a week, so the follow-up period for recidivism was truncated to 27 weeks after the release date, 1 week after it was possible for an assessment to have occurred. Removing all offences after 27 weeks reduced the number of recidivists to 479 (49.6%).

In addition, the current research was focused on the prediction of *imminent* recidivism. New offences needed to occur within a sufficiently short period of time following a DRAOR assessment. Consistent with Lloyd (2015), a 6-week cut-off point was used as the definition of imminent recidivism; all new offences that occurred more than 6 weeks after an assessment were excluded. This requirement for proximity between assessment and outcome reduced the number of recidivists to 460 (47.6%).

Thirty-four men committed a new offence within 6 weeks of their final DRAOR assessment but because that offence occurred more than 27 weeks after release (and therefore outside the truncated follow-up period), these men were classified as non-recidivists. For example, an individual whose final DRAOR assessment occurred 24 weeks after release and was then reconvicted for an offence committed 28 weeks after release would be classified as a non-recidivist, whereas if that man had recidivated in the 27th week, he would have been classified as a recidivist. This rule reduced the recidivism rate by 3.5% (a relative reduction of 6.8%).

Of the 460 recidivists, 205 (44.6%) were convicted for a breach of their parole conditions, 113 (24.6%) for a non-sexual violent offence, 64 (13.9%) for a

property-related offence, 35 (7.6%) for offences against justice, 29 (6.3%) for traffic-related offences, 13 (2.8%) for drug-related offending, and 1 (0.2%) for a sexual offence.

DRAOR assessments.

The final sample consisted of 13,714 assessments completed between 3 September 2012 and 31 July 2015 (M per person = 14.2, SD = 8.2, range = 1-67). The baseline assessment occurred on the day of release for 318 men (33%), before release for 58 men (6%; M days before release = 7.1) and after release for 590 men (61%; M days since release = 2.6). The vast majority (91%) of the post-release baseline assessments occurred in the first week after release. The average time between assessments was 7.5 days (SD = 5.0).

Psychometric Properties of Baseline DRAOR Assessments

Previous research has established that the DRAOR subscales are reliably associated with general recidivism outcomes and has provided some support for a factor structure that matches the three subscales into which the items are organised. However, given we were using a large, previously untested DRAOR dataset, it was important to re-establish both factor structure and predictive validity of baseline DRAOR scores. The important psychometric properties related to change over time—intra-individual change across time points and measurement invariance—are examined in the next chapter. All factor structure analyses were carried out using the lavaan package (Rosseel, 2012) in R (R Core Team, 2018) and all predictive validity analyses were completed using SPSS (version 25).

Factor structure.

A confirmatory factor analysis (CFA) was carried out to examine how well

the data were explained by the original three-subscale structure. Consistent with Yesberg and Polaschek (2015), fit indices for the uncorrelated model indicated poor fit ($\chi^2(152) = 1470.799, p < .001$; CFI = .748; TLI = .717; RMSEA = .095). The fit for the correlated model was better ($\chi^2(149) = 862.522, p < .001$; CFI = .864; TLI = .844; RMSEA = .070), although the Confirmatory Fit Index (CFI) and Tucker-Lewis Index (TLI) were both still below the acceptable threshold of 0.9 (Brown, 2006). When impulse control and problem solving—two theoretically similar, stable risk items—were co-varied (due to modification indices greater than 240), a more acceptable model fit was revealed ($\chi^2(148) = 627.660, p < .001$; CFI = .908; TLI = .894; RMSEA = .058), but this type of model alteration was not desirable given that substantive analyses will use complete subscale scores.

The analyses presented in the previous paragraph treat the DRAOR items as continuous variables, when they are ordinal variables with three categories (scores of 0, 1, or 2). Confirmatory factor analyses using ordinal data (usually from Likert scales) are often conducted as if the data are continuous (Jöreskog & Moustaki, 2001). This practice is usually justified on the basis that the ordinal scale reflects an underlying continuous construct. However, this approach can result in biased parameter estimates (Li, 2016). For ordinal data, diagonal weighted least squares (WLSMV) estimation has been shown to result in more accurate parameter estimates, particularly when the ordinal variable has fewer than five categories (Rhemtulla, Brosseau-Liard, & Savalei, 2012). Therefore, WLSMV estimation is more appropriate than maximum likelihood for the data in our study. When the WLSMV method was used with this dataset (and the DRAOR items were defined as ordinal), most model fit measures improved (correlated model: $\chi^2(149) = 932.522, p < .001$; CFI = .953; TLI = .946; RMSEA = .074).

The CFA results provide some support for use of the existing subscale structure; however, a principal component analysis (PCA) was also run to test the ways in which these data may be inconsistent with the original three-factor DRAOR structure.¹⁰ This analysis allowed further comparison with Yesberg and Polaschek's (2015) DRAOR validation study using a similar sample group. The Kaiser-Meyer-Olkin (KMO) measure (KMO = .89) and Bartlett's test of sphericity ($\chi^2(171) = 5366.34, p < .001$) were both acceptable, indicating the dataset was appropriate for conducting PCA. Four components had an eigenvalue greater than 1. Inspection of the scree plot also indicated a four-component solution. However, parallel analysis supported a three-component solution. The three- and four-factor solutions explained 45.4% and 51.4% of the variance respectively.

The oblique rotated solutions of both the three- and four-component solution provided interpretations that were consistent with the original subscale structure of the DRAOR. In both solutions, the protective factors and the stable factors (except for the attachment item, which loaded more strongly onto the acute items) loaded relatively cleanly on separate components. In the four-component solution, consistent with Yesberg and Polaschek (2015), the acute items loaded onto two separate components, one comprised of just anger/hostility and negative mood, and one with the other five acute items. In

¹⁰ The PCA presented here treats the DRAOR data as continuous, consistent with Yesberg & Polaschek (2015). A PCA using the polychoric correlation matrix, to account for the ordinal nature of DRAOR data, was also run. The parallel analysis suggested a 4-component solution but otherwise the results of that analysis were not meaningfully different from the results presented here (i.e., under 3- or 4-component solutions, the DRAOR items primarily loaded onto components made up of items matching the original subscales).

the three-component solution, the acute factors loaded onto one component, except for the interpersonal relationships item, which did not load above 0.3 on any component.

Overall, both the CFA and PCA results are consistent with previous research and provide support for using the three-subscale structure of the DRAOR. However, as Lloyd (2015) has pointed out, the factor structure of the baseline assessments may be less important than the consistency of the factor structure over time, or *measurement invariance*. Measurement invariance is important because it suggests that changes in scores reflect real change in that individual's functioning or circumstances. If a scale does not display measurement invariance over time, the observed change may instead reflect a change in the way the underlying construct is being measured across time or suggest that different underlying constructs are being measured at different time points (Marsh et al., 2009). Tests of measurement invariance, using longitudinal CFA methods (Liu et al., 2017), are reported in the next chapter, alongside other measures of change over time.

Predictive validity.

To establish predictive validity of DRAOR scores within this dataset, we tested whether subscales scores from the baseline assessment, as defined earlier in this chapter, were reliably associated with general recidivism (defined as any new conviction). We also tested the predictive validity of our static risk measure, the RoC*RoI, to ensure it was demonstrating acceptable predictive validity with this sample. Predictive validity was assessed using Cox regression analyses and Receiver Operating Characteristic (ROC) analyses. Cox regression was chosen rather than logistic regression because some men were not followed up for the

full 6 months ($n = 44$) or even the full 6 weeks ($n = 7$). Using Cox regression allowed us to account for these variations in follow-up length.¹¹ The ROC analyses provide an Area Under the Curve (AUC) statistic, which can be interpreted as the likelihood of a randomly chosen recidivist having a higher score on that subscale than a randomly chosen non-recidivist. This measure is commonly used in risk assessment research and allows easy comparison with previous research.

DRAOR subscale scores should be reliably associated with both short- and long-term recidivism. Most of the previous research using DRAOR data has examined long-term recidivism outcomes, so it was important to establish whether this dataset was comparable to datasets used in previous DRAOR research. A 27-week follow-up period was used. This follow-up length matches the follow-up used for the predictive models throughout this thesis, and therefore provides evidence of how effectively the first DRAOR assessment could predict recidivism over this period in the event no reassessment occurred. The recidivism rate over this period was 49.6% (479 recidivists). This analysis also provides a direct comparison with (Yesberg & Polaschek, 2015), who used a fixed 6-month follow-up in their DRAOR validation study.

It was arguably more important to establish short-term predictive validity. The focus in this thesis is whether reassessment improves short-term, or *imminent* recidivism, defined as a gap of 6 or fewer weeks between a DRAOR assessment and a new offence. Thus, to see whether DRAOR scores can predict

¹¹ Results from logistic regression analyses run with only men who were followed up for the full 6 weeks or 6 months were not substantially different from the Cox regression analyses.

recidivism over a period of that length, we examined whether DRAOR subscales significantly predicted recidivism within 6 weeks (42 days) of release. The recidivism rate over this period was 16.4% (158 recidivists).

The results of these analyses are presented in Table 3.1. The RoC*RoI and three DRAOR subscale scores significantly predicted general recidivism over both follow-up periods. Predictive accuracy was moderately low for each of the three subscales, and in the moderate range for the RoC*RoI (Rice & Harris, 2005). There was some indication of stronger prediction for the shorter follow-up period, particularly for the RoC*RoI, but the 95% confidence intervals overlapped in all cases. Further Cox regressions (not shown) established that all three subscales independently demonstrated incremental predictive validity over the RoC*RoI, for both follow-up lengths.

These results are consistent with findings from previous DRAOR research. Most importantly, because of the similarities in sample and follow-up length, the results are very similar to Yesberg and Polaschek's (2015) findings in their DRAOR validation study. These findings establish the long- and short-term predictive validity of the DRAOR subscale scores in this dataset. We also note that the moderate level of predictive accuracy of the baseline DRAOR scores provides considerable room for reassessment to enhance predictive accuracy.

Table 3.1

*Area Under the Curve (AUC) Statistics and Hazard Ratios from Cox Regression Analyses for Baseline Measures of Static (RoC*RoI) and Dynamic (DRAOR) Risk Predicting General Recidivism Over Follow-up Periods of 6 Weeks and 6 Months.*

	AUC [95% CI]		Hazard ratio [95% CI]	
	6 weeks	6 months	6 weeks	6 months
RoC*RoI	.66 [.61, .70]	.60 [.57, .64]	850.9 [108.4, 6679.6]	64.8 [19.3, 218.0]
Stable	.57 [.52, .62]	.59 [.55, .62]	1.12 [1.05, 1.21]	1.12 [1.08, 1.17]
Acute	.59 [.54, .63]	.57 [.54, .61]	1.11 [1.04, 1.19]	1.08 [1.04, 1.13]
Protect	.59 [.54, .64]	.57 [.54, .61]	.88 [.82, .94]	.91 [.87, .94]

Analytical Technique

Earlier in this chapter, we described the series of decisions made to move from the raw data to a dataset of all DRAOR assessments completed for each man in the sample between his release date (with a few exceptions where the baseline score was drawn from before release) and either (a) a new offence, (b) the final assessment prior to a gap of more than 6 weeks without any assessments, or (c) the end of the 26-week follow-up period. This dataset could be used to answer several different research questions about the predictive validity of the DRAOR for imminent recidivism or about how DRAOR scores change over time during the first 6 months after release from prison. For example, the previous section showed how the dataset could be used to examine the predictive validity of the baseline assessment scores over different follow-up periods. However, for the research questions we wanted to examine in this thesis, the dataset needed to be restructured.

We chose to use discrete-time hazard models to address the research questions in this thesis. Lloyd (2015) recommended using this modelling technique when longitudinal reassessment data with meaningful censoring is being used to answer a research question focused on predictive accuracy. The decision to use the same modelling technique as Lloyd also ensured easier comparison with his findings. Before describing how the dataset was restructured into a format suitable for discrete-time hazard models, it is useful to provide a little more detail about discrete-time hazard models.

The key feature of the discrete-time hazard model is that the follow-up period is divided into a series of discrete time periods of equal length (Singer & Willett, 2003). During each time period, the dataset must include every

individual's score for every predictor variable, and information about whether the event of interest was observed to occur for that individual in the next discrete period (i.e., time 1 data needs to include the outcomes for time 2). Once the event has occurred, those individuals are excluded from the model. Subsequently, the model allows examination of the impact of variability in the predictor variables on both the occurrence and timing of the outcome. The model is therefore well-suited for use with longitudinal data where the outcome may occur at any time during the follow up period or may not ever occur.

Discrete-time hazard models can be used to compare a time-invariant model to a time-varying model that incorporates reassessment scores in some way. Both the time-invariant and time-varying models require a value for the predictor variable in every discrete time period prior to recidivism or censoring. To compare the models, the same score (e.g., the baseline DRAOR assessment score) is used as the predictor in every discrete time period for the time-invariant model, whereas the time-varying model includes any updates in the value of the predictor as they occur (e.g., as DRAOR scores change through reassessment). In this way, the analyses provide a test of whether an outcome is more accurately predicted by an initial value of a predictor, or by a model that frequently updates the value of the predictor.

The discrete-time hazard model is preferable to other statistical approaches that can potentially be used to analyse longitudinal data. This model is very similar to a Cox regression model using time-varying predictors. A Cox model can provide greater flexibility than a discrete-time hazard model, as it is capable of analysing data obtained through more irregular assessment or data collection. Where, as here, data were obtained on a more consistent, regular

schedule, the simplicity of the discrete-time hazard model is preferable.

However, when repeated assessments are conducted on a regular, discrete-time basis, the Cox regression model is essentially identical to the discrete-time model and will return identical parameter estimates. In fact, after structuring the data for a discrete-time hazard model (see next section), analyses were carried out using both techniques and coefficient estimates were essentially identical between the two approaches.

One important point to note about both discrete-time hazard models and Cox regression models is that neither is designed to *describe* change over time. The aim is to compare how well different models predict event occurrence. Several other statistical approaches such as multilevel modelling are available to researchers who wish to examine change over time (Singer & Willett, 2003). Those techniques are not utilised in this thesis. The analyses we conducted to describe change aimed simply to establish that DRAOR scores change sufficiently over time to justify testing different prediction models using a discrete-time hazard model. It is possible however, to incorporate change scores into a discrete-time hazard (or Cox) model. This approach—incorporating change scores into a predictive model— is discussed in more detail in Chapter 6. The details of how change is explicitly incorporated into the models are provided in that chapter.

The same effect size measures used by Lloyd (2015) will be used in this research to compare different discrete-time hazard (and Cox) models. These two statistics—Xu and O’Quigley’s (1999) R^2 and Heagerty and Zheng’s (2005) c -index—are roughly equivalent to the more commonly used R^2 and AUC statistics. The range of possible values are the same (both can range from 0 to 1), with

higher scores indicating more accurate prediction. For the *c*-index, a score of greater than 0.5 indicates prediction at levels above chance. The main difference with the more familiar statistics is that both Xu and O'Quigley's R^2 and Heagerty and Zheng's *c*-index are specifically designed to account for an outcome that can occur at any point in time. This difference changes interpretation slightly. Xu and O'Quigley's R^2 statistic can be interpreted as the amount of variance in the predictor variable (or *variables* in a multivariate model) which can be explained by the ordering of the event occurrence (e.g., does the order in which parolees recidivated match the order we would expect based on their DRAOR scores). The *c*-index can be interpreted as the probability that a randomly selected recidivist would have a higher DRAOR score than a randomly selected non-recidivist, but only if both of the randomly selected individuals have survived in the dataset up to that point in time.

Structuring the Data for Analysis

The DRAOR assessments in the final sample were not carried out according to a discrete time schedule. An assessment could take place on any day (including weekends), and, on rare occasions, even twice in the same day. For the same parolee, assessments could take place with variable frequency that may change over the course of parole. Assessments were therefore irregular but there was still sufficient consistency to allocate them into a series of discrete time periods. This allocation required further structuring of the data.

The implications of defining a discrete unit of time.

In this research, we chose to use 1 week as the discrete unit of time. Choosing a single week as the discrete unit ensured consistency with Lloyd (2015), and best reflects what happens in practice, where supervision officers

see high-risk parolees once a week on average and need to be thinking about whether the parolee will reoffend before the next supervision meeting. There were three major implications for how the dataset needed to be structured that followed from defining the discrete units of time as a single week. The three implications are described in detail below.

Weeks with assessments and recidivism events.

Both DRAOR assessments and recidivism events needed to be defined as occurring in a discrete week. For example, an assessment or offence that occurred between day 0 (day of release) and 6 would be classified as occurring in week 1, between day 7 and 13 as occurring in week 2, and so on. This reclassification led to a problem with ensuring prospective prediction when assessments and recidivism occurred in the same week. In the discrete-time hazard model, an assessment that occurred in week 2 cannot be used to predict a recidivism event that also occurred in week 2, even if the assessment was conducted on a different day earlier that week.

One method for resolving this issue is to carry an assessment backwards into the previous week. For example, if a man was assessed on day 10 and committed a new offence on day 12, those events would have both happened in week 2. In the discrete-time hazard model, that assessment cannot be used for prediction, because it is not prospective. The assessment could be treated as if it occurred in week 1, thus ensuring the most up-to-date information is used for prediction. The problem with this approach is that assessments being compared for recidivists and non-recidivists are not occurring in the same time period. The information for the recidivists becomes more proximal (i.e., closer in time to the outcome) than the information for the non-recidivists, which could unfairly

inflate predictive accuracy. For this reason, all DRAOR assessments occurring in the same week as a new offence ($n = 165$), except when it occurred in the first week, were removed from the dataset. This move increased the length of time between assessment and the first recidivism event from 6.7 days to 9.4 days ($SD = 5.9$).

We made one exception to this rule of excluding assessments that occurred in the same week as recidivism events. There were 19 men who were both assessed and committed a new offence in the first week after release. Of these 19 men, 15 also had a baseline assessment completed in the first week (the other 4 had a baseline score from prior to release). We decided to retain these men in the sample rather than exclude them. As described below, the majority of all baseline scores for both recidivists and non-recidivists were imputed backwards, therefore substantially reducing the risk of any bias from the use of assessments at different time periods.

Weeks with multiple assessments.

The second implication of defining the discrete units of time as weeks was that a decision needed to be made about which assessment should represent each week, if there was more than one available to choose from. In this dataset, the number of assessments completed in a single week ranged from zero to six. Lloyd (2015), drawing on the guidelines of Singer and Willett (2003), tested two methods for determining how a single week should be represented: (1) taking the final assessment from the week, or (2) taking the average of all assessments during the week. The rationale behind these two strategies is that supervision officers may rely either on the single most recent assessment to determine risk since that assessment contains the most updated information, or they may think

that all assessments made during that week should be used to better account for the full pattern of behaviour exhibited during the week. In Lloyd's research, model fit testing revealed the former method to be more suitable for that dataset. Below, the same two methods are empirically tested with the current dataset, in combination with the different imputation methods (see

Table 3.2).

Weeks with no assessments.

The third implication of defining the discrete unit of time as weeks was that a score needed to be imputed for every week where a parolee was not actually assessed on the DRAOR. There were three situations in which a score from an actual assessment needed to be carried forward and one in which a score needed to be carried backwards to fill these missing weeks.

Retrospective imputation.

Scores needed to be carried backwards when the baseline assessment did not occur prior to the day of release. In this situation, imputation was completed by carrying the first actual assessment backwards, to fill earlier missing weeks (e.g., when the first actual assessment occurred during week 4, the exact same baseline assessment score was also recorded for weeks 3, 2, 1 and 0 for that participant). A score needed to be imputed all the way back to "week 0" because week 0 assessments predict outcomes during week 1. Therefore, week 0 scores also represent the baseline scores.

Using a later assessment to represent an earlier week means that prospective prediction is not occurring during those weeks for those participants (e.g., a score from an assessment in week 3 is imputed back to week 1 and then used to predict outcomes during week 2). However, the scores imputed

backwards for these initial weeks were only used to predict an *absence* of offending; no DRAOR assessments that were completed *after* a new offence were included in the dataset. In other words, the dataset excluded any scores where a supervision officer completed a DRAOR assessment with knowledge that the parolee had already committed a new offence for which he would subsequently be convicted. Therefore, the use of backwards-imputed scores is conceptually justified. Additionally, 95% of the sample had been assessed by the end of week 1, so the amount of backwards imputation that occurred during the initial weeks was very small. For week 0, although only 6% of scores represent actual assessments (these are the 58 men whose baseline assessments were taken from before release), recall that 33% of participants were assessed on the day of release, ensuring that week 1 analyses do include a substantial amount of truly prospective prediction.

Prospective imputation.

The most common situation where scores need to be imputed forwards was where there was a gap of at least 1 but no more than 6 weeks between two assessments, likely the result of a parolee reporting to a supervision officer less frequently than once a week. For the purposes of the discrete-time hazard models, a DRAOR score needed to be determined to represent those missing assessments between the two actual assessments.

There were two other situations that required a DRAOR score to be imputed forwards. First, if there was no actual assessment in the week immediately prior to an offence, a score would need to be imputed for every week leading up to the offence (up to a maximum of 6 weeks). Second, at the end of the follow-up, scores were imputed through to week 26 when the following

conditions were met: (a) the final actual assessment took place in weeks 21-25, (b) there was another actual assessment within 6 weeks of the previous assessment but it occurred in weeks 27-32 post-release and thus outside the truncated follow-up period, and (c) no offence took place in between those two assessments. Although this step technically makes use of information outside the follow-up period (i.e., whether there was an assessment in weeks 27-32), this approach is justified on the basis that we have evidence that these men were not reconvicted within the follow-up period. Also, the imputation does not use the actual score from the post-follow-up assessment, just the fact that it exists, thus ensuring that prediction remains prospective.

The following two examples illustrate the three different situations where an earlier score needed to be imputed forwards. Imagine Parolee A was assessed on the day he was released (week 1; also imputes backwards to represent week 0), on day 10 (week 2), on day 25 (week 4) and then offended on day 30 (week 6). For this man, a DRAOR score would need to be imputed for week 3 and week 5. In contrast, imagine Parolee B does not offend at any point following his release. He is assessed on the day of release, every week up until week 24, and then not again until week 27. For this man, a DRAOR score would need to be imputed for weeks 25 and 26, even though the assessment from week 27 was removed from the dataset.

Imputation strategies for carrying forward assessments.

This research used the same three imputation strategies as Lloyd (2015). Where a score needed to be imputed, the imputation was done by each of the following: (a) taking the single most recent DRAOR assessment score; (b) taking an average of the two most recent DRAOR assessment scores; (c) taking an

average of the four most recent DRAOR assessment scores. The underlying logic for these strategies is that a community supervision officer, in making a judgement about the likelihood that a parolee will offend in the next week, may rely on: (a) how the parolee presented at the last completed assessment, (b) how the parolee has presented over the last 2 weeks, or (c) how the parolee has presented over the last month.

These three different imputation strategies were then combined with the two methods for representing multiple assessments in a week: either taking the final assessment or the average of all assessments. Consequently, there were six different possible strategies by which missing weeks could be filled. These six strategies were tested for each of the three DRAOR subscales. Akaike Information Criterion (AIC) values from Cox regression with time-varying predictor models were used to determine which strategy was the best fit for these data, with a lower AIC suggesting better fit.

The results of the model fit testing are presented in

Table 3.2. For all three subscales, the best method of imputation was to carry forward a score from only the previous week, as opposed to an average from multiple previous weeks. However, for the stable and protective subscales, the best model fit was obtained by taking the average of all assessments in a week when there was more than one assessment, and then imputing the average of the single most recent week. In contrast, the best model fit for the acute subscale was obtained by taking the last assessment of the week when there was more than one assessment, and then imputing the single most recent assessment.

Table 3.2

Model Fit Indices (AIC) for Discrete-Time Hazard Models Using DRAOR Subscales to Predict Time to Recidivism, Comparing Six Dataset-Building Strategies.

Strategies for imputing missing weeks				
Strategies for choosing from multiple assessments in the same week		Impute the single most recent	Impute the mean of the previous 2 weeks	Impute the mean of the previous 4 weeks
Choose single last assessment	Stable	4,161.751	4,162.830	4,164.941
	Acute	4,154.142	4,157.614	4,160.688
	Protective	4,166.224	4,167.201	4,168.576
Choose mean of all assessments	Stable	4,161.670	4,162.693	4,164.971
	Acute	4,158.087	4,161.127	4,164.237
	Protective	4,166.247	4,167.176	4,168.383

Note. AIC = Akaike Information Criterion. **Bolded** values indicate the model with the best fit.

To make the rest of the analyses in this thesis easier to interpret, we decided to build the final dataset using the same imputation strategy for all three subscales. In general, as

Table 3.2 shows, the differences between the AICs for each of the different strategies was small. The imputation strategy appeared to make the largest difference for the acute subscale, where using the last score in any week—as opposed to the average of that week—impacted on model fit more substantially than for the stable or protective subscales, which both had almost identical model fit scores regardless of the method used for combining multiple assessments from a previous week. An additional reason for choosing this approach was that it was the same imputation strategy used by Lloyd (2015), thus increasing the comparability of the two studies.

In summary, the final dataset used for all analyses was built by taking the last assessment of the week when there were multiple assessments and, when there was a gap of fewer than 6 weeks between assessments, imputing the score from only the previous week.

Final dataset after imputation.

After imputation, the final dataset comprised of 16,652 assessments (M per person = 17.2, SD = 9.3, range = 1-27). With the exception of week 0, which was discussed earlier, the amount of imputation required in each week ranged from 7-35% (see Table 4.2 in the next chapter), with a gradual increase in the amount of imputation required for weeks later in the follow-up. This pattern can most likely be attributed to parolees being required to report for supervision less frequently as they are out on parole (and are not observed to recidivate) for longer periods of time.

A Final Note About Men with Only One Non-Imputed Assessment

One of the strengths of the dataset used in this thesis is that most individuals were frequently reassessed, allowing us to measure change over time and examine how that change was associated with recidivism. We have already highlighted one exception to that feature: the 19 men who had a recidivism event in week 1, and thus only the single assessment that occurred in week 0. With only one assessment, these men had no opportunity to demonstrate change. There was another group that also could not demonstrate change, even though they had more than one assessment included in the dataset: men with more than one assessment included in the final dataset, but only one *non-imputed* assessment.

There were 46 men who had multiple assessments included in the final dataset, only one of which was a non-imputed assessment. Of these 46 men, 27 only had one non-imputed score in the raw data. The other 19 men were assessed more than once (i.e., they had more than one than one assessment in the raw dataset completed prior to recidivism or censoring) but only one non-imputed assessment was included in the final dataset because all of their assessments had occurred in the same week. As outlined earlier, when there were multiple assessments in the same week, only the final assessment from the week was included in the final dataset.

In total, between the 46 men with multiple assessments in the final dataset but only one non-imputed assessment, and the 19 men who had a recidivism event in week 1, there were 65 men, representing 6.7% of the total sample, who only had a single non-imputed assessment included in the final dataset. The majority of these men ($n = 55$) were reconvicted for an offence committed shortly after the first assessment, which explains the absence of any reassessment. It is unclear why the other 10 men were

censored after only a single assessment, although we might speculate that they were placed on remand or recalled to prison for behaviour that did not result in a conviction.

For these 65 men, there was no possibility of observing change in their scores. Change is essential to the research questions in this thesis; for example, without change, reassessment will not enhance prediction. Therefore, the fact that a small proportion of our sample could not change reduced the likelihood that we would find that reassessment enhanced prediction. This issue was particularly relevant for the 46 men who were reconvicted or censored in week 2 or later, because these men had scores imputed up until the week they were censored or reconvicted, giving the impression that no change had occurred over several weeks, when the fact only one non-imputed assessment had occurred meant it was impossible for change to occur in this period.

Despite these issues, we decided to retain these men in the sample for two main reasons. First, the 46 men reconvicted or censored in week 2 or later were not as problematic as they might have appeared. Recall that most ‘week 0 scores’ were scores imputed backwards from week 1, so the vast majority of the sample only had one non-imputed assessment completed in week 0 or week 1. Thus, no change in scores could be observed between week 0 and week 1. Only six of the 46 men were censored or had a recidivism event in *week 3* or later, and of those six, two had multiple assessments in the same week and only ended up with a single ‘real’ assessment because of the data structuring strategy used. Therefore, the problem of a seemingly artificial absence of change among men with a single non-imputed assessment was relevant for a tiny fraction of the final sample (4 out of 966).

More importantly, this research is not merely about examining the impact of reassessment or change. The primary goal of this research was to better understand what happens to men released onto parole from prison who are assessed to be at high

risk of being reconvicted and returning to prison. Among that population, a very important sub-group is men who recidivate within a very short time of being released. Predictive models that include those men are a better representation of the population we are studying than models that exclude, through a process of self-selection, men who are only assessed once prior to recidivism or censoring. The question of whether reassessment enhances prediction would not be relevant if all men had only been assessed once, but that was not the nature of the sample. Instead, research questions related to reassessment and change arose because of the fact that the vast majority of the men *were* assessed multiple times, and there was an intention to assess *all* men multiple times.

Ultimately, we argue that the method we have chosen is the best available for examining the prediction of imminent recidivism among high-risk parolees. The fact that a small proportion of the sample only had a single non-imputed assessment score included in the final dataset was a side effect of the reasoning process used in developing that method. To avoid any doubt though, all predictive models presented in the following chapters were also run after excluding the 65 men with only one non-imputed assessment. The substantive findings and conclusions from those analyses were not significantly different from the findings and conclusions of the analyses that are presented in the following chapters.

Summary

In this chapter, we have described the methodology behind the selection and creation of a dataset that could be used for analyses that address the primary research question. We laid out the background for why this data is collected in applied settings and described the psychometric properties that the measure we are using has demonstrated in previous research. After that background, we presented our reasons

for choosing this particular dataset, and detailed the steps that were taken to move from raw data to a dataset suitable for the research questions and analyses we are exploring in this thesis. We presented some tests of the psychometric properties of an initial dataset and then described how further data were imputed to create the final dataset that we will use for the rest of the analyses in this thesis.

In the next chapter, we examine that final dataset in close detail, exploring the properties of the data and how well the data meets the assumptions of the regression models generally, and the discrete-time hazard models described in this chapter. At the end of the next chapter, we present the first of our primary analyses, testing whether a discrete-time hazard model that incorporates reassessment scores improves the prediction of imminent recidivism compared to a model that uses only scores from the baseline assessment.

Chapter 4: Does Reassessment Enhance Prediction?

In his thesis, Lloyd (2015) proposed a three-step framework for testing the predictive value of reassessment. In this chapter, we replicate the first step of that framework; the second and third steps are replicated in the following chapter. The first step of the framework requires comparing the predictive accuracy of a baseline assessment in a sequence of assessments against the single most proximal assessment at each time point. We use the dataset and the statistical technique—discrete-time hazard models—described in the previous chapter to investigate that question. As we outlined in Chapter 2, if the variables being measured are truly dynamic risk and protective factors, reassessment should enhance prediction. However, there is limited existing empirical evidence to support that proposition.

Before presenting those analyses, we first present the descriptive statistics and tests of the model assumptions. The descriptive statistics we have chosen to present provide readers with an overview of the rates of recidivism for each week of the 26-week follow up period and describe the change over time in DRAOR subscale scores, at both the group (inter-individual) and individual (intra-individual) level. Assessment scores must demonstrate change over time if they are going to improve prediction compared to a baseline assessment score. This chapter only aims to establish whether change is occurring; the measures of change are not explicitly included in the predictive models presented at the end of the chapter. Models that do include change scores are presented in Chapters 6 and 7. The assumptions we test include those that apply to all regression models, and those that apply specifically to discrete-time hazard models and Cox regression with time-varying predictor models.

All descriptive statistics and model assumption tests presented in this chapter were described in detail by Lloyd (2015). The discussion of the rationale and

underpinnings of each analysis has deliberately been kept shorter in this thesis.

However, the information presented in this chapter should be sufficient for readers to gain an understanding of the dataset, and any transformations that were required prior to running the predictive models that are presented at the end of this chapter. Analyses were run using SPSS (version 25) and R (R Core Team, 2018). The longitudinal measurement invariance analyses were run using R and Mplus (version 8.0).

Descriptive Statistics

Life table.

Table 4.1 shows the order in which the outcomes of interest happened for the men in the sample. There were 966 men who had a baseline DRAOR assessment. Only 325 men had a DRAOR assessment in the final week of the 26-week follow-up. The specific nomenclature in the “Week” column indicates observed scores during one week and outcomes during the following week. For example, the row “[5-6]” indicates that 805 men were observed during week 5 (i.e., they have an assessment score in that week). The next two columns indicate that 29 of the 805 men recidivated during week 6, and 9 men were censored (either because they had no more assessments on file or had a gap of more than 6 weeks before their next assessment). The final two columns show the percentage of men who recidivated in that week (the *hazard rate*), and the cumulative proportion of men remaining in the community who have not recidivated (the *survivor function*). There are 27 rows in the table, representing the 27 discrete time periods (baseline and 26 weeks) into which the follow-up was divided.

Some key features of the life table should be noted. First, there was at least one recidivist in every week of the follow up, which is important for the discrete-time hazard models used later in this chapter and throughout the thesis. Second, unlike most life tables, this table includes time 0. Recall from the previous chapter that week 0

Table 4.1

Life Table Showing When High-Risk Parolees Recidivated or Were Censored Over the First 26 Weeks After Release from Prison.

Week	Entered Week	Recidivated ^a	Censored ^b	% Recidivated	% Remaining
<i>N</i>	966	-	-	-	100
[0-1)	966	19	0	2.0	98.0
[1-2)	947	30	9	3.2	94.9
[2-3)	908	23	6	2.5	92.5
[3-4)	879	27	8	3.1	89.7
[4-5)	844	27	12	3.2	86.8
[5-6)	805	29	9	3.6	83.7
[6-7)	767	19	9	2.5	81.6
[7-8)	739	10	6	1.4	80.5
[8-9)	723	26	5	3.6	77.6
[9-10)	692	22	5	3.2	75.1
[10-11)	665	14	8	2.1	73.6
[11-12)	643	21	6	3.3	71.2
[12-13)	616	14	4	2.3	69.5
[13-14)	598	20	4	3.3	67.2
[14-15)	574	15	10	2.6	65.5
[15-16)	549	20	5	3.6	63.1
[16-17)	524	16	5	3.1	61.2
[17-18)	503	15	4	3.0	59.3
[18-19)	484	13	6	2.7	57.7
[19-20)	465	16	4	3.4	55.7
[20-21)	445	14	6	3.1	54.0
[21-22)	425	7	2	1.6	53.1
[22-23)	416	10	4	2.4	51.8
[23-24)	402	14	7	3.5	50.0
[24-25)	381	7	7	1.8	49.1
[25-26)	367	5	37	1.4	48.4
[26-27)	325	7	56	2.2	47.4

^a A parolee was recorded as a recidivist during week [1-2) for example, if they committed a new offence (that they were later convicted of) during week 2

^b A parolee was considered censored during week [2-3) for example, if they had a DRAOR assessment in week 2 but did not have another assessment until at least week 9, and during that time they also did not offend.

Figure 4.1. Hazard rate (proportion of men who recidivated) for each week of the 26-week follow-up.

Figure 4.2. Cumulative proportion of men remaining crime-free in the community across the 26 weeks of the follow-up.

scores represent the baseline assessment. The 19 recidivists in the first row of data are the men who recidivated during week 1. No men were classified as censored in the week [0-1) period, indicating that all 54 men who had a baseline DRAOR from before release and did not recidivate in the first week had at least one more DRAOR in the dataset after release.

The hazard rates and survivor functions from the life table are graphed in Figure 4.1 and Figure 4.2 respectively. Unlike Lloyd (2015), who found a gradual decrease in hazard rate over time, the hazard rate in our sample was relatively consistent over time, although there was a slight increase in hazard rates through the middle period, and a slight decrease in later weeks. The rates were roughly double those observed by Lloyd, which was expected given our sample is characterised by higher static risk. The graph of the survivor function (Figure 4.2) shows a steady decrease in the proportion of offenders surviving in the community without having recidivated. The estimated survival rates were also much lower than those observed by Lloyd, with 67.2% estimated to be remaining crime-free after 3 months, and less than half (47.4%) remaining at the end of the full 6-month follow-up. By comparison, Lloyd's rates for those same time periods were 79.1% and 65.0%, respectively.

Inter-individual change.

In Table 4.2, we present the means and standard deviations for each of the three DRAOR subscales over the 26 weeks of the follow-up. For all three subscales, scores tended to be close to the midpoint (i.e., 7 for the acute subscale, and 6 for the stable and protective subscales) throughout the follow-up period. There was a consistent trend of decreasing scores across time for the acute and stable subscales, and gradually increasing scores for the protective subscale. This trend suggests that parolees in this

Table 4.2

Means, Standard Deviations and Percentage of Non-Imputed Scores for DRAOR Subscales for Each Week of the 26-Week Follow-Up.

Week	<i>n</i>	Stable	Acute	Protective	% of non-imputed scores
		<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>	
0	966	7.40 (2.16)	6.10 (2.25)	5.31 (2.15)	
1	947	7.41 (2.18)	5.94 (2.21)	5.31 (2.14)	93.5
2	908	7.37 (2.29)	5.75 (2.21)	5.38 (2.18)	87.3
3	879	7.28 (2.36)	5.56 (2.24)	5.47 (2.22)	86.2
4	844	7.20 (2.39)	5.41 (2.26)	5.52 (2.29)	85.3
5	805	7.10 (2.46)	5.24 (2.24)	5.62 (2.29)	83.4
6	767	7.02 (2.48)	5.15 (2.25)	5.73 (2.31)	80.3
7	739	6.96 (2.49)	5.05 (2.24)	5.80 (2.34)	79.7
8	723	6.93 (2.50)	4.98 (2.24)	5.89 (2.38)	79.5
9	692	6.84 (2.49)	4.92 (2.25)	5.95 (2.37)	77.2
10	665	6.79 (2.52)	4.82 (2.26)	6.01 (2.41)	74.6
11	643	6.79 (2.52)	4.81 (2.28)	6.05 (2.44)	75.4
12	616	6.69 (2.53)	4.74 (2.26)	6.15 (2.44)	71.8
13	598	6.65 (2.52)	4.71 (2.26)	6.17 (2.44)	73.2
14	574	6.60 (2.57)	4.74 (2.34)	6.22 (2.49)	71.2
15	549	6.52 (2.60)	4.61 (2.31)	6.27 (2.51)	68.3
16	524	6.45 (2.60)	4.57 (2.24)	6.34 (2.50)	66.4
17	503	6.39 (2.62)	4.53 (2.24)	6.35 (2.54)	64.4
18	484	6.30 (2.60)	4.53 (2.27)	6.41 (2.50)	65.3
19	465	6.27 (2.61)	4.47 (2.31)	6.45 (2.54)	64.3
20	445	6.17 (2.58)	4.47 (2.32)	6.50 (2.53)	63.8
21	425	6.11 (2.58)	4.42 (2.34)	6.56 (2.55)	64.9
22	416	6.07 (2.61)	4.32 (2.23)	6.59 (2.53)	61.5
23	402	6.05 (2.63)	4.25 (2.22)	6.62 (2.56)	62.9
24	381	6.04 (2.65)	4.26 (2.29)	6.64 (2.55)	57.5
25	367	5.96 (2.67)	4.22 (2.26)	6.68 (2.56)	65.7
26	325	5.89 (2.61)	4.11 (2.25)	6.81 (2.52)	65.2

sample were rated as less likely to recidivate as they remained in the community for longer periods without receiving any new convictions.

Table 4.2 also includes the proportion of scores in each week that were non-imputed. As we noted in the previous chapter, there was a gradual increase over time in the amount of imputation that occurred. The means and standard deviations thus reflect fewer actual scores during later weeks in the follow-up period.

Intra-individual change.

Table 4.2 demonstrates that mean scores for each of the three DRAOR subscales changed on a week-to-week basis for the whole sample, although only by small increments. It is theoretically possible that this group-level change is the result of weekly changes in the composition of the sample group due to recidivism and censoring. Therefore, before testing whether reassessment improves prediction of imminent recidivism, we also need to consider intra-individual change.

We can assess intra-individual change in several different ways. First, we can calculate the amount of change made by each man between the first and last DRAOR assessments in his sequence.¹² These results ($M_{\text{stable}} = -.43$, $SD = 2.06$; $M_{\text{acute}} = -.90$, $SD = 2.34$; $M_{\text{protect}} = .50$, $SD = 1.99$) show that, on average, scores changed by approximately half a point on the stable and protective subscales, and by almost a full point on the acute subscale. DRAOR items are scored either 0, 1, or 2, so a change in a single item would change the subscale score by at least 1 point. Therefore, the amount of change evident here is very limited, although the high standard deviations indicate that there was considerable variability in change scores. The limited amount of change evident here is relatively consistent with Lloyd's (2015) findings. In his sample, the subscales

¹² These analyses exclude the 19 men who recidivated in week 1.

changed by about one point (slightly more for the acute) over the follow-up, but his follow-up was twice the length of the follow-up used in this research.

Overall, 45.7% of men had the same stable score in their final assessment as they did in their baseline assessment. Similarly, 50.1% of men had the same protective score as they started with, but only 26.0% had the same acute score. These proportions are slightly higher than the proportions observed by Lloyd (2015), but again the shorter follow-up in this research suggests the results are quite consistent.

By focusing on only the baseline and the final assessment to measure change, we risk missing changes that may occur between those assessments. Several men who started and finished with the same score on one of the subscales had some fluctuation in their scores through the follow-up period. When this type of change was accounted for, the proportion of men who demonstrated no change dropped to 42.3% for the stable subscale, 46.0% for the protective subscale, and 15.0% for the acute subscale. These proportions are still relatively low, but suggest there is some fluctuation in scores, particularly for the acute subscale.

We can assess that fluctuation in more detail by looking at the changes in scores across consecutive weeks. In Table 4.3, we present the percentage of scores on the three subscales that changed from the previous week during each week of the follow-up. This table shows that scores for both the stable and protective subscales demonstrated very little change between individual weeks. Between any two weeks, only about 5% of men had any change—positive or negative—in their stable or protective scores. The acute subscale scores changed more frequently (about 15-30% of the time), but it was still far more common for scores not to change across consecutive weeks of the follow-up. When scores did change, for the stable and acute subscales, scores decreased almost twice as often as they increased. For the protective subscale, scores increased almost

Table 4.3

Proportion of Men Showing Weekly Change for DRAOR Subscales for Each Week of the 26-Week Follow-Up.

Week	<i>n</i>	Stable			Acute			Protective		
		Increase	Decrease	No Change	Increase	Decrease	No Change	Increase	Decrease	No Change
0-1	947	3.5%	2.4%	94.1%	4.8%	12.9%	82.4%	2.5%	2.5%	94.9%
1-2	908	6.3%	6.6%	87.1%	9.4%	19.2%	71.5%	6.3%	4.2%	89.5%
2-3	879	2.3%	6.3%	91.5%	6.4%	16.7%	76.9%	5.6%	2.3%	92.2%
3-4	844	3.7%	5.9%	90.4%	8.9%	17.2%	73.9%	5.8%	2.1%	92.1%
4-5	805	2.1%	6.1%	91.8%	7.2%	14.0%	78.8%	5.7%	1.5%	92.8%
5-6	767	2.3%	4.8%	92.8%	8.6%	12.5%	78.9%	5.9%	2.5%	91.7%
6-7	739	2.6%	3.2%	94.2%	8.5%	13.1%	78.3%	3.9%	1.6%	94.5%
7-8	723	2.9%	4.3%	92.8%	10.1%	13.8%	76.1%	6.1%	1.5%	92.4%
8-9	692	1.9%	4.8%	93.4%	8.5%	12.0%	79.5%	4.0%	2.0%	93.9%
9-10	665	2.3%	3.2%	94.6%	7.4%	10.1%	82.6%	3.2%	1.7%	95.2%
10-11	643	1.9%	2.0%	96.1%	10.1%	11.5%	78.4%	2.0%	1.4%	96.6%
11-12	616	1.3%	4.1%	94.6%	7.6%	10.7%	81.7%	3.4%	1.0%	95.6%

12-13	598	1.0%	2.5%	96.5%	9.9%	10.2%	79.9%	1.8%	0.8%	97.3%
13-14	574	1.7%	2.6%	95.6%	10.1%	9.9%	80.0%	3.0%	1.6%	95.5%
14-15	549	1.6%	2.6%	95.8%	7.5%	9.8%	82.7%	2.7%	1.1%	96.2%
15-16	524	1.5%	3.4%	95.0%	8.2%	8.8%	83.0%	2.9%	1.3%	95.8%
16-17	503	0.8%	2.4%	96.8%	7.2%	9.1%	83.7%	2.4%	0.8%	96.8%
17-18	484	0.8%	3.7%	95.5%	7.9%	8.3%	83.9%	2.9%	1.4%	95.7%
18-19	465	1.5%	1.9%	96.6%	6.2%	9.9%	83.9%	2.8%	1.7%	95.5%
19-20	445	0.9%	2.9%	96.2%	7.0%	8.5%	84.5%	4.5%	1.6%	93.9%
20-21	425	1.4%	3.3%	95.3%	7.1%	9.4%	83.5%	4.0%	1.9%	94.1%
21-22	416	1.4%	2.6%	95.9%	7.2%	11.3%	81.5%	2.2%	1.0%	96.9%
22-23	402	1.0%	2.5%	96.5%	5.2%	7.7%	87.1%	1.5%	0.5%	98.0%
23-24	381	1.6%	2.4%	96.1%	5.5%	6.8%	87.7%	2.9%	0.5%	96.6%
24-25	367	1.9%	3.0%	95.1%	8.4%	7.6%	83.9%	3.0%	1.4%	95.6%
25-26	325	0.9%	1.2%	97.8%	6.8%	8.3%	84.9%	3.1%	0.6%	96.3%

twice as often as they decreased. These patterns were expected based on the declining group scores seen in Table 4.2.

Unsurprisingly, the *amount* of change observed on a week-by-week basis was also limited. The mean weekly change across consecutive weeks was very low for all three subscales ($M_{\text{stable}} = -.01, SD = .22$; $M_{\text{acute}} = -.05, SD = .30$; $M_{\text{protect}} = .01, SD = .19$). The mean *absolute* change (i.e., in either direction) across consecutive weeks was much higher for all three subscales ($M_{\text{stable}} = .12, SD = .22$; $M_{\text{acute}} = .33, SD = .36$; $M_{\text{protect}} = .11, SD = .19$), again suggesting there was some fluctuation over time (i.e., the overall pattern of gradual decline included several weeks where scores increased before declining again). Acute subscale scores fluctuated almost three times as much as the other two subscales.

Table 4.3 also shows that changes were more likely to occur during earlier weeks of the follow-up period, particularly for the acute subscale. Since more imputation occurred during later weeks in the follow-up, it is possible that the imputation strategy may partly explain the decreasing amount of change during these weeks. When an imputed score follows a non-imputed score, no change can occur because the same score is carried forward. However, this explanation cannot account for the discrepancy in change between the acute subscale and the other two subscales. Instead, the change patterns may reflect increasing stability as the men survive for longer periods in the community, or that supervision officers became less likely to revise their ratings once they felt they had got to know the parolee.

Intra-class correlations (ICCs) provide further information about intra-individual change. These measures indicate the level of variability within assessment sequences for the same individual compared to the variability of the full sample. One-way random, average measures ICCs were calculated for the three subscales (stable: $ICC_1 = .988$, 95% CI [.981, .993]; acute: $ICC_1 = .970$, 95% CI [.952, .983]; protective: $ICC_1 = .989$, 95% CI

[.983, .994]). For all three subscales, these scores indicate a very high level of consistency of scores across time for the same individual.

Finally, a stability coefficient can be used to show how much intra-individual change affects the ranking of participants across time. The stability coefficient is calculated by correlating all scores in all weeks (for all participants who have scores in both of those weeks) and taking the average of all correlations. For this sample, the stability coefficients were high for all subscales (stable: $r = .85$, acute: $r = .73$; protective: $r = .86$). These values suggest that there was some change in rankings by DRAOR scores across time but, in general, men who started with high DRAOR scores maintained relatively higher scores compared to other men throughout the follow-up.

In summary, intra-individual change was evident in this sample. The frequency and amount of change observed was slightly less than Lloyd (2015) observed with his sample, but these differences are at least partly explained by the shorter follow-up period being used in this research. Conceptual issues related to intra-individual change and the relationship between such change and imminent recidivism is explored in much greater detail in Chapter 6.

Correlations and internal reliability.

Subscale correlations provide information about the relationship between the different measures, and how those relationships changed (or not) over time. Correlations provide important information about the similarities in measures and can be used as an indicator of multicollinearity (another measure of multicollinearity is presented later in this chapter). As a general rule, measures of risk factors are expected to be positively correlated with other measures of risk factors and negatively correlated with measures of protective factors. Correlations between dynamic variables are not expected to change over time, but correlations between dynamic and static variables

Table 4.4

*Internal Reliability (Cronbach's Alpha), Inter-Subscale Correlations, and Correlations with Static Risk (RoC*RoI) for Each Week of the 26-Week Follow-Up.*

Week	<i>n</i>	Pearson's <i>r</i>						Cronbach's α		
		Stable & Acute	Stable & Protective	Acute & Protective	Stable & RoC*RoI	Acute & RoC*RoI	Protective & RoC*RoI	Stable	Acute	Protective
0	966	0.46	-0.58	-0.38	0.11	0.08	-0.13	0.77	0.61	0.85
1	947	0.45	-0.58	-0.37	0.12	0.09	-0.13	0.77	0.60	0.85
2	908	0.42	-0.60	-0.39	0.14	0.10	-0.13	0.77	0.60	0.84
3	879	0.42	-0.60	-0.41	0.14	0.11	-0.13	0.77	0.62	0.83
4	844	0.44	-0.62	-0.44	0.14	0.10	-0.11	0.77	0.62	0.83
5	805	0.43	-0.64	-0.43	0.14	0.07	-0.12	0.77	0.62	0.82
6	767	0.45	-0.64	-0.45	0.15	0.07	-0.13	0.77	0.61	0.82
7	739	0.45	-0.66	-0.48	0.14	0.07	-0.13	0.77	0.60	0.82
8	723	0.47	-0.67	-0.49	0.14	0.10	-0.13	0.77	0.59	0.81
9	692	0.48	-0.66	-0.47	0.12	0.10	-0.14	0.76	0.59	0.81
10	665	0.49	-0.66	-0.48	0.12	0.11	-0.14	0.76	0.59	0.81
11	643	0.50	-0.66	-0.48	0.12	0.09	-0.14	0.76	0.60	0.81

12	616	0.49	-0.67	-0.46	0.10	0.07	-0.12	0.76	0.60	0.81
13	598	0.49	-0.67	-0.47	0.10	0.07	-0.11	0.76	0.60	0.81
14	574	0.49	-0.68	-0.48	0.10	0.07	-0.10	0.77	0.61	0.82
15	549	0.49	-0.69	-0.51	0.09	0.05	-0.10	0.77	0.62	0.82
16	524	0.48	-0.69	-0.49	0.10	0.08	-0.11	0.77	0.59	0.82
17	503	0.50	-0.69	-0.49	0.10	0.07	-0.12	0.78	0.58	0.82
18	484	0.51	-0.69	-0.50	0.12	0.07	-0.13	0.77	0.60	0.82
19	465	0.52	-0.69	-0.50	0.12	0.09	-0.15	0.77	0.61	0.82
20	445	0.50	-0.70	-0.51	0.12	0.08	-0.15	0.77	0.62	0.82
21	425	0.51	-0.71	-0.50	0.11	0.06	-0.14	0.77	0.64	0.82
22	416	0.49	-0.71	-0.50	0.11	0.05	-0.14	0.77	0.61	0.82
23	402	0.50	-0.70	-0.50	0.10	0.04	-0.14	0.78	0.61	0.82
24	381	0.49	-0.68	-0.50	0.10	0.06	-0.15	0.78	0.64	0.82
25	367	0.50	-0.70	-0.50	0.13	0.06	-0.16	0.79	0.62	0.82
26	325	0.48	-0.70	-0.46	0.16	0.07	-0.16	0.78	0.62	0.81

might be expected to decline in strength over time as dynamic scores move further from their baseline values.

In Table 4.4, we present the correlations between DRAOR subscale scores, and between subscale scores and the RoC*RoI for each week of the follow-up. The inter-subscale correlations were moderately strong, with the strongest correlations evident between the stable and protective subscales and were in the expected direction (risk factors correlated positively with risk factors and negatively with protective factors). Correlations tended to increase slightly over time. The correlations between the RoC*RoI and the DRAOR subscales were much weaker than the inter-subscale correlations. The correlations with the acute subscale were particularly weak, and despite the large sample size, were not significant from week 12 onwards (significance not shown). Other than the non-significance, there was no clear trend in the correlations over time; they tended to be consistent strength throughout the follow-up.

The Cronbach's alpha of the individual subscales for each week are also presented in Table 4.4. The results show that the internal consistency for the protective subscale was good throughout the follow-up and acceptable for the stable subscale. The internal consistency for the acute subscale was below the acceptable threshold at all time points. These results are consistent with the levels of internal reliability observed by Chadwick (2014) and Lloyd (2015).

Longitudinal measurement invariance.

Another important condition for meaningful examination of change is longitudinal measurement invariance. In the previous chapter, confirmatory factor analysis (CFA) and principal components analysis (PCA) were used to demonstrate that individual DRAOR item scores from the baseline assessment reflect three latent constructs that largely correspond with the three DRAOR subscales on which those

items are placed. Those findings were consistent with previous research examining the factor structure of the DRAOR (Chadwick, 2014; Hanby, 2013; Yesberg & Polaschek, 2015), and provide support for running analyses using subscale total scores. However, as Lloyd (2015) noted, where a predictive risk assessment tool is being used, and where change analyses are central to the research question, the factor structure at a single time point may be less important than the consistency of measurement over time.

Previous research has found some support for consistency of the DRAOR factor structure at different points in time (Hanby, 2013; Yesberg & Polaschek, 2015). These studies conducted separate CFA analyses using assessments from two and three different time points respectively. Both studies indicated that the factor structure and model fit did not appear to change over time.

We used a similar approach with this dataset, running a parallel analysis and a CFA of the three-subscale structure for all 27 time points. The results are presented in Table 4.5. Kaiser's criterion (i.e., eigenvalues greater than 1) suggested a four-component solution for almost every week, whereas the parallel analysis results suggested a three- or four-component solution during the initial weeks, but a two-component solution for most weeks after that point. Inspection of factor loadings indicated that the two-component solutions were a result of the stable and protective items loading together onto one component, and the acute items loading onto a second component. The CFA results showed more consistency over time than the estimates of the number of components. Measures of model fit were consistently high for all weeks, with gradual improvements in fit over time, particularly for the RMSEA (likely a result of decreasing sample size). These findings, in addition to the internal reliability results in the previous section, suggest the factor structure is relatively consistent over time.

Table 4.5

Number of Components Obtained from Parallel Analysis, and Measures of Fit from Confirmatory Factor Analyses (CFA) Testing the Original Three-Factor Structure of the DRAOR for Each Week of the 26-Week Follow-Up.

Week	<i>n</i>	Kaiser's criterion	Parallel analysis <i>n</i>	CFA measures of fit		
		<i>n</i> components	components	CFI	TLI	RMSEA
0	966	4	4	0.953	0.946	0.074
1	947	4	4	0.957	0.951	0.070
2	908	5	3	0.953	0.946	0.068
3	879	4	3	0.957	0.951	0.063
4	844	4	2	0.957	0.951	0.062
5	805	4	2	0.96	0.954	0.058
6	767	4	2	0.955	0.949	0.060
7	739	4	2	0.959	0.953	0.056
8	723	4	2	0.956	0.950	0.057
9	692	4	2	0.949	0.942	0.059
10	665	4	3	0.953	0.946	0.058
11	643	4	3	0.951	0.943	0.059
12	616	4	2	0.952	0.945	0.059
13	598	4	2	0.954	0.947	0.057
14	574	4	2	0.953	0.946	0.059
15	549	4	2	0.962	0.957	0.053
16	524	4	2	0.958	0.951	0.055
17	503	4	2	0.960	0.954	0.054
18	484	4	2	0.963	0.958	0.051
19	465	4	2	0.966	0.961	0.049
20	445	4	2	0.967	0.962	0.049
21	425	3	2	0.971	0.967	0.046
22	416	3	2	0.976	0.972	0.042
23	402	3	2	0.973	0.969	0.045
24	381	4	2	0.976	0.972	0.042
25	367	4	2	0.979	0.976	0.040
26	325	4	2	0.973	0.969	0.043

Note. Kaiser's criterion = number of components with eigenvalues greater than 1; CFI = Confirmatory Factor Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation

Comparison of multiple individual factor analyses is informative, but longitudinal measurement invariance is most robustly established by fitting models to data from multiple time points simultaneously. Liu et al. (2017) outline how this should be done with ordered-categorical data, using a CFA framework. Exploratory approaches such as exploratory structural equation modelling (ESEM; Asparouhov & Muthén, 2009) can also be used for this purpose, but were not chosen here due to the satisfactory fit of the CFA results with individual time points.

As described by Liu et al. (2017), the CFA approach compares progressively more constrained models of data from at least two time points. An initial *baseline* model tests whether similar factor structure is evident at different time points. If the baseline model demonstrates appropriate fit, that model is then compared to a nested *loading invariance* model, where factor loadings are constrained to be equal across time points. Invariance of loadings, as indicated by no significant change in model fit, suggests that the same underlying constructs are being measured at each time point. The third model is a *threshold invariance* model, where, in addition to factor loadings, item thresholds are constrained to equality across time points. Invariance of thresholds suggests that the same scores on the individual items have the same meaning at different time points (i.e., they are being measured consistently).

Initial attempts to test these models with our dataset were unsuccessful. Inspection of the problem indicated the presence of linear dependence between the same items across time. Linear dependence can result when the correlation between two variables in a model is nearly perfect. In the previous section, we showed that subscale scores demonstrated change over time, but the amount of change was minimal, particularly for the stable and protective subscales. Underlying this limited change at the subscale level was even less change at the item level. The lack of item change over

time meant that the correlations between the same items, even at the most distant time points, were very high.¹³ To resolve this issue, the standardised residual covariances between several of the same items (all from the stable or protective subscales) at different time points were constrained to values of no greater than 1. These constraints did not resolve the issues with linear dependency when three time points were included in the models. Therefore, tests were only run with two time points.

We tested whether the assumption of longitudinal measurement invariance held across the first and middle week (week 13), and the first and last week (week 26) of the follow-up, the maximum period over which change can occur. The results are presented in Table 4.6. For both tests, the change in measures of fit was below the recommended thresholds (Chen, 2007) for both the CFI ($> .01$) and the RMSEA ($> .015$), indicating measurement invariance had been established. However, Liu et al. (2017) have questioned the use of changes in fit indices as a valid way of establishing measurement invariance. Instead, they recommend using robust chi-square difference testing (Asparouhov & Muthén, 2006). In contrast to the model fit results, all robust chi-square difference results were significant, indicating that neither loading nor threshold invariance had been established.

Overall, these results provide mixed support for the presence of longitudinal measurement invariance. Consistent with previous research, CFA results with the individual time points were consistent across all weeks of the follow-up, although there was some fluctuation in the number of components indicated by parallel analysis. The

¹³ One possible explanation for this finding was the method of imputation. By imputing the same score in missing weeks, we may have increased the correlation across time. However, when the longitudinal invariance analyses were run using only non-imputed scores, the same problems arose, suggesting that item scores did not change much over time, even before imputation.

Table 4.6

Tests for Measurement Invariance Between Scores from the Beginning and Middle, and the Beginning and End of the 26-Week Follow-Up Using Confirmatory Factor Analysis.

	<i>RMSEA</i>	Δ <i>RMSEA</i>	<i>CFI</i>	Δ <i>CFI</i>	$\Delta \chi^2(df)$
<u>Week 0 and 13</u>					
Baseline model	.041	-	.967	-	
Loading invariance model ^a	.041	0	.965	.002	
Threshold invariance model ^b	.042	.001	.963	.004	193.5(31)***
<u>Week 0 and 26</u>					
Baseline model	.030	-	.968	-	
Loading invariance model ^a	.030	0	.968	0	
Threshold invariance model ^b	.032	.002	.964	.004	312.1(31)***

^a In the loading invariance model, constraints are placed on the factor loadings across time.

^b In the threshold invariance model, constraints are placed on the factor loadings and the item thresholds.

Note. CFI = Confirmatory Factor Index; RMSEA = Root Mean Square Error of Approximation. Δ CFI of less than .01 or a Δ RMSEA of less than .015 are generally used as the thresholds for establishing measurement invariance. The $\Delta \chi^2$ values are drawn from Asparouhov and Muthén's (2006) DIFFTEST in *Mplus*.

*** $p < .001$

small amount of change in the individual DRAOR items across time presented a serious challenge for more robust tests of invariance. When these tests were run, model fit indices indicated that longitudinal invariance had been established, but chi-square difference testing suggested that invariance may have been violated. Failure to find longitudinal invariance across all methods should be considered a limitation of the dataset, particularly when interpreting analyses that assess the impact of changes in subscale scores.

General Regression Assumptions

In this section, we examine the assumptions of normality, homoscedasticity, linearity, multicollinearity and test for both univariate and multivariate outliers. Although some of these assumptions (e.g., normality) are not essential for Cox regression or discrete-time hazard models, violations of general regression assumptions can impact on the power to detect effects (Tabachnick & Fidel, 2013).

Normality and univariate outliers.

In Table 4.7, we present measures of skew and kurtosis, and counts of the number of univariate outliers for each week of the follow-up. Scores on the acute subscale demonstrated a consistent positive skew throughout the follow-up (i.e., scores were grouped at the lower end of the distribution). Even after applying square-root transformations, the skewness issues remained.¹⁴ The protective subscale initially demonstrated negative skew, but this skew reduced over later weeks, and did not depart significantly from normality for the entire second half of the follow-up period. There did not appear to be any clear departures from normality for the stable subscale.

¹⁴ Models tested using these transformed acute subscale scores resulted in the same substantive conclusions as analyses using the un-transformed variable. All reported models use the un-transformed variable.

Table 4.7

Measures of Normality—Skewness Ratios, Kurtosis Ratios, and Number of Extreme Outliers—for Each Week of the 26-Week Follow-Up.

Week	Stable			Acute			Protective		
	Skew	Kurtosis	<i>n</i> extreme outliers	Skew	Kurtosis	<i>n</i> extreme outliers	Skew	Kurtosis	<i>n</i> extreme outliers
0	2.51	-1.79	2	5.53	1.12	4	-8.97	4.66	0
1	1.39	-1.92	2	5.60	1.68	3	-8.71	4.57	0
2	-0.22	-2.08	0	5.57	1.34	2	-7.15	3.22	0
3	-0.97	-2.55	0	6.12	2.08	4	-5.87	2.56	0
4	-1.63	-2.30	0	5.10	0.88	2	-5.05	1.52	0
5	-1.84	-2.58	0	5.61	1.39	2	-4.69	1.15	0
6	-1.66	-2.71	0	5.75	2.44	3	-4.11	1.04	0
7	-1.38	-3.01	0	5.08	1.58	2	-3.56	0.69	0
8	-1.42	-3.15	0	4.92	1.02	2	-3.68	0.17	0
9	-1.30	-2.90	0	4.11	0.07	1	-3.38	0.02	0
10	-1.46	-3.13	0	4.47	0.50	1	-3.23	-0.49	0
11	-1.10	-3.16	0	4.04	-0.73	1	-3.30	-0.63	0
12	-0.82	-3.24	0	4.33	-0.08	1	-2.66	-0.64	0

13	-0.83	-3.19	0	4.99	0.65	1	-2.70	-0.57	0
14	-0.59	-3.29	0	4.99	0.13	1	-2.67	-0.67	0
15	-0.36	-3.13	0	6.10	1.89	1	-2.62	-0.76	0
16	-0.39	-3.06	0	5.51	2.06	3	-2.80	-0.52	0
17	-0.06	-3.07	0	4.92	1.42	3	-2.58	-1.02	0
18	-0.09	-3.06	0	5.41	1.85	4	-2.80	-0.76	0
19	0.05	-3.12	0	5.44	1.98	1	-2.68	-1.00	0
20	0.24	-2.96	0	4.95	1.49	1	-2.68	-1.12	0
21	-0.09	-3.00	0	5.75	3.13	2	-2.73	-1.11	0
22	-0.18	-3.29	0	4.83	1.79	2	-2.55	-1.17	0
23	-0.45	-3.13	0	4.95	2.09	2	-2.72	-1.16	0
24	-0.43	-3.29	0	5.53	2.70	3	-2.71	-1.13	0
25	-0.34	-3.10	0	4.82	1.92	2	-2.60	-1.17	0
26	0.10	-2.97	0	4.05	0.90	1	-2.63	-0.69	0

Note. **Bolded values** indicate statistically significant departures from normality, defined as values greater than 3.29. Skewness ratios and kurtosis ratios are calculated by dividing the skewness or kurtosis value by its standard error. Extreme outliers were defined as values with absolute z-scores greater than 3.29.

No issues with kurtosis were evident for any of the three subscales.

There were very few extreme univariate outliers (defined as absolute z-scores greater than 3.29) identified in the dataset. For the stable subscale, only 2 outliers were observed in each of the first 2 weeks, with none observed in any other week. There were no extreme outliers observed for the protective subscale in any week. Between one and four outliers were evident in every week for the acute subscale, representing less than one percent of the scores in any single week.

Homoscedasticity and linearity.

Bivariate scatterplots of the three DRAOR subscales suggested that the predictors all shared linear relationships with no obvious departures from homoscedasticity. There was also no evidence of heteroscedasticity in the relationships between the subscales and the outcome.

The Box-Tidwell approach was used to examine whether the predictors shared linear relationships with the logit-transformed outcome. In this approach, interaction terms of the predictors and their natural logs are added to the models. For the stable and protective subscales, these interaction terms were not significant predictors, indicating a linear relationship. However, for the acute subscale, the interaction term was a significant predictor, suggesting a possible non-linear relationship between the outcome and the predictor. To understand this interaction further, we graphed the logit-transformed hazard rates against the acute scores (graph not shown). Visual inspection of this graph showed a clear linear relationship for all values of the acute subscale except 13 and 14 (the two highest values). For these two scores, recidivism rates were lower than expected. Since only ten men had scores of 13 or 14 at any point in the follow-up, we conducted Box-Tidwell tests with these men excluded. However, the interaction term between the acute and log of the acute remained significant. When

a quadratic transformation of the acute variable was added to the models, the interaction was not significant.

The Box-Tidwell results suggested a possible non-linear effect for the acute subscale. To further examine this finding, we ran models that included quadratic transformations of the predictor variables for the DRAOR subscales and the RoC*RoI. These results are presented in Table 4.8. Consistent with the Box-Tidwell test results, the quadratic predictor was significant for the acute subscale but not for the stable and protective subscales. The quadratic predictor was also not significant for the RoC*RoI. The significant, negative coefficient for the quadratic predictor in the acute model is consistent with the earlier observation that the highest values of the acute subscale were associated with hazard rates that were lower than expected.

The models presented at the end of this chapter for the acute subscale were run with the quadratic predictor term included. Adding the quadratic term enhanced the effect sizes obtained, but this finding was consistent for both the baseline and reassessment models, so it did not change the overall conclusions. Model fit measured by AIC slightly favoured the quadratic model, but BIC slightly favoured the linear model. For this reason, none of the models presented at the end of the chapter include a quadratic predictor term.

Multivariate outliers.

Mahalanobis distances were calculated for combinations of the three DRAOR subscales and the RoC*RoI to check for multivariate outliers, which were defined as $p < .001$ on the chi-square distribution (Tabachnick & Fidel, 2013). These scores indicated the presence of 43 multivariate outliers from 13 individual sequences. The outliers occurred in consecutive weeks in every instance where there was more than one multivariate outlier within the same sequence. And, in all but four cases, there was

Table 4.8

Results for Fitting Four Discrete-Time Hazard Models Predicting Criminal Recidivism Including Quadratic Predictor Terms.

Parameters	RoC*RoI	Stable	Acute	Protect
	<i>B</i> (SE)	<i>B</i> (SE)	<i>B</i> (SE)	<i>B</i> (SE)
Constant	-13.88** (4.91)	-4.78*** (.37)	-5.10*** (.27)	-2.99*** (.17)
Week	.0002 (.006)	.004 (.006)	.008 (.006)	.004 (.006)
Score	22.30 (12.55)	.19 (.10)	.39*** (.09)	-.06 (.07)
Quadratic Score	-11.40 (7.97)	-.003 (.007)	-.02** (.01)	-.008 (.006)
Deviance	4162.34	4155.50	4139.36	4158.71
AIC	4170.34	4163.50	4147.36	4166.71
BIC	4201.22	4194.38	4178.24	4197.59
$\Delta \chi^2$ from model with no quadratic term	2.12 (1)	.26 (1)	8.79** (1)	1.51 (1)

Note. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria. **Bolded values** indicate statistically significant results.

* $p < .05$, ** $p < .01$, *** $p < .001$

no change on the three DRAOR subscales between weeks (i.e., it was the same combination of scores creating the outlier for each individual).

Visual inspection of the multivariate outliers, and the assessments before and after those outliers, suggested several possible explanations. In general, the explanations centre on the DRAOR subscales. As we saw in Table 4.4, the correlations between the DRAOR subscales were moderate to strong throughout the follow-up. The correlations show that individuals with high scores on the stable subscale were likely to have high scores on the acute subscale and low scores on the protective subscale. We can thus examine individual combinations of scores to see if they are conceptually consistent with this pattern. In contrast, the correlations between the RoC*RoI and the DRAOR subscales were weak throughout the follow-up, making it much more difficult to interpret whether the RoC*RoI score for an individual was contributing towards that combination of scores being a multivariate outlier. For that reason, the explanations that follow focus on only the three DRAOR subscales.

First, 2 of the 13 men (accounting for 6 of the 43 outliers) were assessed as having extremely high acute risk scores (i.e., the maximum 14 out of 14), very high stable risk, and moderate protective factors during early periods in the follow-up. Conceptually, although the protective scores might be expected to be lower in these cases, having only moderate protective factors is not a clear conceptual inconsistency. The outliers were most likely to be the result of the extreme risk values. Notably, these sequences moved back into the normal range when risk scores came down slightly in later weeks.

Second, for 5 of the 13 men (and 24 outliers), either the stable or protective scores were inconsistent with the other two DRAOR subscales during the initial weeks of the follow-up. For example, protective subscale scores were low in the presence of

low stable and acute scores. In each of these cases, the problematic subscale was adjusted in later weeks, suggesting supervision officers may have taken a few weeks to gain a proper understanding of the relevant variables. These outliers can be considered conceptually inconsistent.

Third, for 6 of the 13 men (and 13 outliers), the outliers appeared to be a result of acute scores that were inconsistent with the stable and protective scores. It is theoretically possible that acute risk may be at least partially independent from the other two DRAOR subscales and could rise or drop suddenly when similar patterns are not observed in the other subscales. Therefore, these cases are not necessarily conceptually inconsistent. One case where a sudden large increase in acute scores occurred immediately prior to a recidivism event, could be viewed as a good example of how the acute subscale is theoretically supposed to work. However, in two other cases, sudden increases in acute risk were followed either by censoring or an immediate return to the lower score. The censoring may correspond with undetected offending, but equally, the latter case may have been caused by manual error in data entry of the DRAOR score. It is thus difficult to classify these cases as either conceptually consistent or inconsistent, but for the purposes of deciding which sequences to exclude, we classified all of these 6 cases as conceptually inconsistent, with the exception of the case where the sudden increase in acute scores immediately preceded recidivism.

Multivariate outliers are not unexpected in large datasets (Tabachnick & Fidel, 2013). However, it is good practice to ensure the outliers are not having a major influence on the results. Thus, the models presented at the end of this chapter were run without the 13 sequences containing multivariate outliers, and then without only the 10 sequences that were deemed conceptually inconsistent. Exclusion of sequences with outliers, whether conceptually consistent or inconsistent, did not affect the overall

results. Therefore, all models presented include the sequences with multivariate outliers.

Multicollinearity.

Although in several weeks the DRAOR subscales were quite strongly correlated with each other, there did not appear to be any problems with multicollinearity between predictors. All tolerance values for the RoC*RoI and the three DRAOR subscales were well above 0.2 in every week of the follow-up.

Assumptions Specific to Cox Regression and Discrete-Time Hazard Models

In this section, we examine three additional assumptions of the data. These first and third assumptions—equivalence of censored cases and proportionality of hazards—are essential to both Cox regression and discrete-time hazard models. The second assumption—choosing a representation of time—is only a requirement for discrete-time hazard models, where a variable representing time needs to be included in the models (Tabachnick & Fidel, 2013).

Equivalence of censored and remaining cases.

In Chapter 3, we outlined the different ways in which an individual might be censored. Most commonly, individuals were censored because of a recidivism event. However, as the life table (Table 4.1) at the beginning of this chapter shows, in each week there were several men who were censored for reasons other than recidivism. For the modelling approach used in this thesis, it is essential to establish that these men were not meaningfully different than those who remained in the dataset at that point of the follow-up. To test this assumption, we chose to use the approach taken by Lloyd (2015): a series of independent t-tests on RoC*RoI and DRAOR subscale scores to compare the censored and remaining cases in every week of the follow-up.

Table 4.9

*Independent Samples t-Tests Comparing the Static Risk (RoC*RoI) and DRAOR Subscale Scores of Censored and Uncensored Non-Recidivists for Each Week of the 26-Week Follow-Up.*

Week	n Censored	RoC*RoI		Stable		Acute		Protective	
		t	p	t	p	t	p	t	p
0-1	0	-	-	-	-	-	-	-	-
1-2	9	-1.12	0.26	-1.16	0.25	-0.74	0.46	0.80	0.42
2-3	6	-1.07	0.29	-1.05	0.29	-0.10	0.92	1.40	0.16
3-4	8	0.96	0.34	-2.41	0.02	-0.59	0.56	0.13	0.90
4-5	12	-0.72	0.47	-0.47	0.64	-2.43	0.02	1.12	0.29
5-6	9	-0.38	0.70	-1.68	0.09	-0.22	0.84	2.05	0.04
6-7	9	-1.62	0.11	-1.89	0.06	-1.90	0.06	1.12	0.26
7-8	6	-0.81	0.42	-1.53	0.13	-0.87	0.39	1.55	0.12
8-9	5	-0.96	0.34	-0.62	0.54	-1.02	0.31	0.86	0.39
9-10	5	-1.08	0.28	1.58	0.18	-0.33	0.74	-0.01	0.99
10-11	8	-0.21	0.84	1.15	0.25	-0.10	0.92	0.18	0.86
11-12	6	-0.35	0.72	0.40	0.69	-0.88	0.42	-0.23	0.82
12-13	4	-2.70	0.01	-0.47	0.64	-2.69	0.01	1.36	0.18
13-14	4	-2.02	0.04	-3.31	0.04	-0.72	0.48	0.37	0.71
14-15	10	-3.07	<0.01	-1.04	0.30	-3.18	<0.01	0.28	0.95
15-16	5	1.20	0.23	-1.48	0.14	-1.39	0.17	1.36	0.18
16-17	5	-1.00	0.32	0.18	0.85	-1.47	0.14	-0.92	0.39
17-18	4	1.74	0.08	-2.85	0.06	-0.20	0.84	1.10	0.27
18-19	6	0.60	0.55	-1.30	0.19	-0.35	0.73	0.91	0.36
19-20	4	-1.35	0.18	-0.79	0.43	-0.89	0.38	-0.61	0.54
20-21	6	0.07	0.95	-2.09	0.04	-0.74	0.46	1.32	0.19
21-22	2	1.10	0.27	-0.50	0.62	-1.01	0.31	0.61	0.54

22-23	4	-0.97	0.33	1.01	0.31	-0.77	0.50	-0.12	0.91
23-24	7	1.81	0.07	-0.23	0.82	1.83	0.07	-0.98	0.33
24-25	7	-0.57	0.57	-0.19	0.07	-0.93	0.35	-0.22	0.83
25-26	37	-0.05	0.96	-0.86	0.39	-2.03	0.04	1.76	0.08
26-27	56	-0.84	0.40	-2.60	0.01	-1.65	0.10	3.96	<0.01

Note. The notation 1-2 for “Week”, indicates these are the scores obtained during assessment in week 1 for cases that were censored (or not) in week 2. Positive t statistics indicate the mean for uncensored cases was higher than the mean for censored cases; negative t statistics indicate the mean for censored cases was higher than the mean for uncensored cases. **Bolded values** indicate statistically significant ($p < .05$) results.

The results of these t-tests are presented in Table 4.9. The table indicates that there was a consistent trend for censored cases to have higher static, stable, and acute risk, and lower protective scores. However, most of these differences were not significant. The only point where there was any evidence of the scores being consistently higher was a 3-week period in the middle of the follow-up. Overall, the assumption of equivalence between censored and remaining cases appears valid for these data.

Choosing a representation of time.

A discrete-time hazard model must include a representation of time in the model based on the association between the outcome and time (Singer & Willett, 2003). In Lloyd's (2015) data, he found that a quadratic representation of time provided the most accurate estimates of the hazard rate over time, and thus he included a 'week-squared' variable in each of his models. We used the same method to examine the optimal representation of time with our dataset.

The results of the comparison tests are presented in Table 4.10. The table shows that none of the representations of time resulted in a significantly worse fit than a general model where all 27 time points were entered as separate variables. Of the different representations, we see that none of the stepwise progressions to linear, quadratic, or cubic models significantly improved the fit from a constant-only model. The AIC and BIC scores suggest the constant-only model provides the best approach to modelling the effect of time on hazard in our dataset, with some evidence that the quadratic model may provide a better fit than the linear model.

Figure 4.3 provides a visual representation of these findings. This graph replicates Figure 4.1, except the hazard rate has been logit-transformed, and linear and

Table 4.10

Comparison of Different Representations of Time for a Discrete-Time Hazard Model.

Representation of time	n parameters	Deviance	Δ General Model		Δ Previous Model		AIC	BIC
			χ^2 (df)	p	χ^2 (df)	p		
Constant	1	4209.11	29.27 (26)	.299			4211.11	4218.83
Linear	2	4208.76	28.92 (25)	.267	0.35 (1)	.557	4212.76	4228.20
Quadratic	3	4205.77	25.92 (24)	.357	3.00 (1)	.083	4211.77	4234.93
Cubic	4	4205.38	25.54 (23)	.323	0.39 (1)	.533	4213.38	4244.26
General	27	4179.84					4233.84	4442.29

Note. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

Figure 4.3. Logit-transformed hazard rate (proportion of men who recidivated) for each week of the 26-week follow-up, with superimposed linear and quadratic lines of best fit.

quadratic lines of best fit are now included. As we noted earlier, there was only a slight decrease in hazard rates over time, an observation confirmed by the gradual decline in the linear line of best fit. We also noted that the hazard rate appeared to increase slightly during the middle weeks and then decrease during later weeks, an observation supported by the quadratic line of best fit. Table 4.10 suggests that neither of these trends was significant and the change in hazard rate over time was not significantly different from zero. However, there was an indication that the quadratic representation may provide more accurate estimation than the linear model.

Although the constant-only model demonstrated slightly better fit than the linear representation of time, the discrete-time hazard models presented in this thesis include the linear representation of time alongside the constant. This choice reflects the fact that the difference in model fit between the constant-only and linear representation was not statistically significant; thus, either were statistically plausible. The main reason for choosing to include the linear representation of time was that it ensured all models could be interpreted as the predictor effect after controlling for time, making the model interpretation more similar to the traditional Cox regression survival analysis model. This decision also ensured easier comparison with Lloyd's results (2015), which included a (quadratic) representation of time. Note that models that included a quadratic representation were also tested in the current thesis, with no change in the substantive results observed.

Proportionality of hazards.

One of the central assumptions of any proportional hazard model—including all forms of Cox regression *and* discrete-time hazard models—is that hazard rates remain proportionate over time (Singer & Willett, 2003). Each unit increase in a predictor variable should result in a similar change in the hazard rate at each point in time. The

base hazard rate may change over time, but the impact of the predictor variable should be consistent over time. This assumption can be tested in a few different ways.

A simple method for testing proportionality is to graph hazard functions across time for different values of each predictor (Singer & Willett, 2003). Consistent with Lloyd (2015), in Figure 4.4, Figure 4.5, Figure 4.6, and Figure 4.7, we present graphs for the RoC*RoI and the three DRAOR subscales. For the RoC*RoI graph, the two lines represent logit-transformed hazard rates for values above and below the median; for the DRAOR subscales, the two lines represent logit-transformed hazard rates for scores on those subscales that were one standard deviation above and below the mean.

In all four graphs, higher scores on risk and lower scores on protective factors were associated with higher hazard rates throughout the follow-up. The individual data points are highly variable though, so it is clearer if interpretation focuses on the linear lines of best fit. On the stable and acute graphs (Figure 4.5 and Figure 4.6) the two lines remain very close to equidistant throughout, indicating proportionality of hazards. For the RoC*RoI (Figure 4.4), the two lines converge as the follow-up progresses, to the point they almost cross. For the protective subscale (Figure 4.7), the two lines diverge as the follow-up progresses, although to a slightly lesser extent than the RoC*RoI. The non-parallel nature of the two graphs suggests that those two variables are not consistent predictors of recidivism throughout the follow-up and could indicate a violation of the assumption of proportional hazards.

One limitation of this approach to testing for proportionality is that it puts equal weight on the hazard rate for each week. Hazard rates in later weeks are calculated using much lower sample sizes, and anomalous results in those weeks can distort the shape of these graphs. For example, the presence of only one or two recidivists in the “low-protective” line during some later weeks appeared to strongly influence the

results. Thus, while these graphs and the patterns observed are informative, more rigorous tests are needed to establish the extent of any breach of proportionality.

A more rigorous way of testing for the proportionality of hazards is to calculate Schoenfeld residuals (Schoenfeld, 1982) for the individual predictors and graph these across time. Schoenfeld residuals represent the difference between the observed value of a predictor and the expected value of that predictor at the time recidivism occurs. Proportionality is present if the residuals are not significantly correlated with time (Singer & Willett, 2003). To test this, we calculated the Spearman rank correlations between Schoenfeld residuals of the Cox regression models and time. The correlations were not significant for the stable ($\rho = -0.04, \chi^2 = 0.64, p = 0.43$), acute ($\rho = 0.01, \chi^2 = 0.03, p = 0.87$), and protective subscales ($\rho = 0.04, \chi^2 = 0.65, p = 0.42$), suggesting no issues with proportionality were present. However, a significant negative correlation ($\rho = -0.15, \chi^2 = 9.34, p < 0.01$) was found for the RoC*RoI, suggesting a violation of proportionality.

A final method for testing proportionality is to include interaction terms between predictors and time in the predictive models. Discrete-time hazard models that included these interaction terms are presented in Table 4.11.¹⁵ The table shows that, for the three DRAOR subscales, including interaction terms did not significantly improve any of the models, and none of the interaction terms were significant within those models. However, the interaction between the RoC*RoI and time was significant, and including this term improved the model fit.

¹⁵ Interactions between the predictors and quadratic time were also examined, but these terms were all non-significant and did not significantly improve model fit.

Figure 4.4. Logit-transformed hazard rates for each week of the 26-week follow-up for RoC*RoI scores above and below the median, with superimposed linear lines of best fit.

Figure 4.5. Logit-transformed hazard rates for each week of the 26-week follow-up for scores approximately 1 SD above and below the mean for the DRAOR stable risk subscale, with superimposed linear lines of best fit.

Figure 4.6. Logit-transformed hazard rates for each week of the 26-week follow-up for scores approximately 1 SD above and below the mean for the DRAOR acute risk subscale, with superimposed linear lines of best fit.

Figure 4.7. Logit-transformed hazard rates for each week of the 26-week follow-up for scores approximately 1 SD above and below the mean for the DRAOR protective subscale, with superimposed linear lines of best fit.

Table 4.11

Results for Fitting Four Discrete-Time Hazard Models Predicting Criminal Recidivism Using Score by Time Interaction Terms

Parameters	RoC*RoI	Stable	Acute	Protect
	<i>B</i> (SE)	<i>B</i> (SE)	<i>B</i> (SE)	<i>B</i> (SE)
Constant	-8.95*** (.85)	-4.74*** (.29)	-4.44*** (.22)	-2.74*** (.18)
Week	.20 (.07)	.01 (.02)	.004 (.02)	-.007 (.02)
Score	7.02*** (1.07)	.16*** (.04)	.15*** (.03)	-.16*** (.03)
Score x Week	-.26** (.09)	-.001 (.003)	.001 (.003)	.002 (.002)
Deviance	4155.13	4155.47	4148.11	4159.58
AIC	4163.13	4163.47	4156.11	4167.58
BIC	4194.02	4194.35	4186.99	4198.46
$\Delta \chi^2$ from model with no interaction term	9.33** (1)	.28 (1)	.04 (1)	.65 (1)

Note. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria. **Bolded values** indicate statistically significant results.

* $p < .05$, ** $p < .01$, *** $p < .001$

Overall, there was consistent evidence from three different methods that the assumption of proportionality of hazards was not violated for the DRAOR subscales in this dataset. There was also consistent evidence that the RoC*RoI scores violated this assumption; all three methods indicated that the RoC*RoI was a weaker predictor of recidivism during later weeks of the follow-up. This finding is difficult to explain: the impact of static risk factors on recidivism is expected to be consistent across time and Lloyd (2015) did not observe a similar interaction in his dataset. Further replication is necessary before any attempt should be made to interpret this finding as meaningful. The issue is of minor importance to the substantial analyses in this thesis though, as the primary research questions in this thesis concern dynamic factors, rather than static factors. However, since we were interested in comparing the predictive accuracy of the RoC*RoI to the DRAOR subscales, and testing models where the RoC*RoI was included alongside the dynamic predictors, a “RoC*RoI by time” interaction term was added to all models that included the RoC*RoI.

Baseline vs. Reassessment Models

In this section, we present the first set of analyses in this thesis that directly test the primary research question. First, we present a series of univariate models. These analyses compare models for each of the three DRAOR subscales that predicted imminent recidivism using the baseline assessment score at each time point against models that used the single most proximal assessment score at each point (i.e., the score obtained through reassessment). Second, we present multivariate models, where all three DRAOR subscales were entered in models alongside the RoC*RoI. These analyses also compare the predictive accuracy of the baseline score against the proximal score. For all models, we present overall measures of accuracy and weekly accuracy measures,

to get a complete picture of the extent to which reassessment impacts on the prediction of imminent recidivism using these different predictor variables.

All models were checked for the presence of highly influential cases. DFBeta values were well below the recommended cut-off value of 2 (Ryan et al., 2015) for all predictors in all models, indicating that no cases were having a disproportionate influence on parameter estimation.

Univariate baseline and reassessment models.

In Table 4.12, Table 4.13, and Table 4.14, we present the results of three discrete-time hazard models that used the DRAOR subscales to predict imminent recidivism. In all three tables, Model A includes only the initial, baseline assessment score at every time point, whereas Model B includes the most proximal score at each time point obtained through reassessment. Model C includes both baseline and proximal scores at each time point.

The results presented in Table 4.12 suggest that reassessment of the stable subscale improved the prediction of imminent recidivism. Individually, baseline stable scores and proximal stable scores were both significant predictors and had similar odds ratios. However, the AIC and BIC values indicated that the reassessment model (Model B) had a better fit than the baseline model (Model A). The two measures of effect size also favoured Model B. The differences in Xu and O'Quigley's (1999) R^2 measure indicated that the reassessment model explained twice as much variance in the predictor variable as the baseline model. Note, however, that the overall amount of variance explained remained low: 12% vs. 7%. Heagerty and Zheng's (2005) concordance index (c -index) showed that the proximal score provided a more accurate ranking of the men in the study than the baseline score, although again both

Table 4.12

Results from Three Discrete-Time Hazard Models Predicting Time to Recidivism Using Scores from the DRAOR Stable Risk Subscale.

Parameters	Model A		Model B		Model C	
	<i>B (SE)</i>	OR [95% CI]	<i>B (SE)</i>	OR [95% CI]	<i>B (SE)</i>	OR [95% CI]
Constant	-4.44*** (.19)	.01 [.01, .02]	-4.62*** (.18)	.01 [.01, .01]	-4.62*** (.20)	.01 [.01, .02]
Week	-.001 (.006)	1.00 [.99, 1.01]	.004 (.006)	1.00 [.99, 1.02]	.004 (.006)	1.00 [.99, 1.02]
Stable (Baseline Score)	.12*** (.02)	1.13 [1.08, 1.18]			-.001 (.03)	1.00 [.94, 1.07]
Stable (Proximal Score)			.14*** (.02)	1.15 [1.11, 1.20]	.14*** (.03)	1.15 [1.09, 1.22]
R^2 (XO)		.07		.12		.13
c-index		.57		.60		.60
Deviance		4179.01		4155.75		4155.75
AIC		4185.01		4161.75		4163.75
BIC		4208.17		4184.91		4194.63
$\Delta \chi^2$ Model C vs. Model A						23.26*** (1)

Note. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria. **Bolded values** indicate statistically significant results. Model A predicts time to recidivism using only baseline scores at every time point, Model B predicts time to recidivism using only the most proximal score at every time point, and Model C predicts time to recidivism using the baseline and the most proximal score at every time point.

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 4.13

Results from Three Discrete Time Hazard Models Predicting Time to Recidivism Using Scores from the DRAOR Acute Risk Subscale.

Parameters	Model A		Model B		Model C	
	<i>B (SE)</i>	OR [95% CI]	<i>B (SE)</i>	OR [95% CI]	<i>B (SE)</i>	OR [95% CI]
Constant	-4.07*** (.15)	.02 [.01, .02]	-4.47*** (.15)	.01 [.01, .02]	-4.41*** (.17)	.01 [.01, .02]
Week	-.003 (.006)	1.00 [.99, 1.01]	.007 (.006)	1.01 [.99, 1.02]	.007 (.006)	1.01 [1.00, 1.02]
Acute (Baseline Score)	.09*** (.02)	1.09 [1.05, 1.13]			-.02 (.03)	.98 [.93, 1.03]
Acute (Proximal Score)			.16*** (.02)	1.17 [1.12, 1.21]	.17*** (.03)	1.18 [1.13, 1.24]
R ² (XO)		.03		.11		.13
c-index		.55		.60		.60
Deviance		4190.85		4148.14		4147.50
AIC		4196.85		4154.14		4155.50
BIC		4220.01		4177.30		4186.38
$\Delta \chi^2$ Model C vs. Model A						43.35*** (1)

Note. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria. **Bolded values** indicate statistically significant results. Model A predicts time to recidivism using only baseline scores at every time point, Model B predicts time to recidivism using only the most proximal score at every time point, and Model C predicts time to recidivism using the baseline and the most proximal score at every time point.

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 4.14

Results from Three Discrete Time Hazard Models Predicting Time to Recidivism Using Scores from the DRAOR Protective Factor Subscale.

Parameters	Model A		Model B		Model C	
	<i>B (SE)</i>	OR [95% CI]	<i>B (SE)</i>	OR [95% CI]	<i>B (SE)</i>	OR [95% CI]
Constant	-3.00*** (.13)	.05 [.04, .06]	-2.85*** (.12)	.06 [.05, .07]	-2.89*** (.13)	.01 [.01, .02]
Week	-.002 (.006)	1.00 [.99, 1.01]	.004 (.006)	1.00 [.99, 1.02]	.004 (.006)	1.00 [.99, 1.02]
Protect (Baseline Score)	-.10*** (.02)	.90 [.87, .94]			.03 (.03)	1.03 [.96, 1.10]
Protect (Proximal Score)			-.13*** (.02)	.87 [.84, .91]	-.15*** (.03)	.86 [.81, .91]
R ² (XO)		.06		.12		.12
c-index		.56		.59		.59
Deviance		4186.34		4160.22		4159.48
AIC		4192.34		4166.22		4167.48
BIC		4215.50		4189.38		4198.36
$\Delta \chi^2$ Model C vs. Model A						26.86*** (1)

Note. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria. **Bolded values** indicate statistically significant results. Model A predicts time to recidivism using only baseline scores at every time point, Model B predicts time to recidivism using only the most proximal score at every time point, and Model C predicts time to recidivism using the baseline and the most proximal score at every time point.

* $p < .05$, ** $p < .01$, *** $p < .001$

concordance indices indicated relatively small effects. Finally, the model comparisons at the bottom of Table 4.12 show that the proximal score added incremental predictive validity to the baseline score. When both measures were included in the same model, only the proximal score was a significant predictor.

The pattern of results for the acute and protective subscales was very similar. Table 4.13 and Table 4.14 show that the baseline and proximal scores were significant predictors in their individual models, but only the proximal score remained significant in the combined models. Model fit was better and effect sizes were larger for the proximal score for both subscales. The acute subscale in particular appeared to demonstrate a substantial difference, with the proximal score explaining almost four times as much variance as the baseline score. The proximal score also showed significant incremental predictive validity over the baseline score for both subscales. These results suggest that, as with the stable subscale, reassessment of the acute and protective DRAOR subscales improved the prediction of imminent recidivism.

To further compare the baseline and proximal scores for the three DRAOR subscales, we graphed the time dependent AUC values across the 26-week follow-up. The results are presented in Figure 4.8, Figure 4.9, and Figure 4.10. In each graph, the AUCs for the two models are presented alone in the top half, and then repeated in the bottom half, but this time with bootstrapped 95% confidence intervals added to the graphs. From these graphs, we can see at which points in the follow-up the predictive accuracy of the two models was significantly different.

The results for the stable subscale are presented in Figure 4.8. The AUC values for both models initially start around 0.65, before dropping off quite sharply, rising again during the middle weeks, dropping off again, and then rising slightly at the end of the follow-up. The most striking feature of the graph is the parallelism of the two lines

Figure 4.8. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for proximal and baseline scores on the DRAOR stable subscale. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid.

Figure 4.9. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for reassessment and baseline scores on the DRAOR acute subscale. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid.

Figure 4.10. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for reassessment and baseline scores on the DRAOR protective subscale. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid.

for most of the follow-up period. The proximal score has slightly better predictive accuracy than the baseline score in almost every week except at the very beginning and end. However, we can see in the bottom half of Figure 4.8 that the 95% confidence intervals overlapped at all times, indicating that none of these differences were statistically significant.

The time-dependent AUC values for the acute subscale, plotted in Figure 4.9, showed a slightly different pattern. The values started much lower than the values for the stable subscale—around 0.55—and rose during the initial weeks instead of falling. The pattern of parallel lines was present, but the pattern was not as consistent as was observed for the stable subscale. Most notably, when we look at the bottom half of the figure, we see that for 4 weeks during the middle of the follow-up (weeks 11-14), the confidence intervals of the two models did not overlap. There was also considerable separation in the two lines during the final weeks, but these differences were not significant, partly because the confidence intervals get much wider towards the end of the follow-up due to the decreased sample size.

Finally, the time-dependent AUC values for the protective subscale models are presented in Figure 4.10. The AUC values for the proximal score were relatively consistent over time, with the exception of a small peak in the early to middle weeks. Similar to the stable subscale, the AUC values for the baseline model were slightly lower than the values for the proximal score except for a few weeks at the start and end of the follow-up. And as with the stable subscale, the confidence intervals indicated that the differences in AUC values were not significant at any point.

Multivariate baseline and reassessment models.

In practice, the DRAOR subscales are not used or considered in isolation. It is strongly recommended that supervision officers and other decision-makers consider all

three subscales, in addition to the RoC*RoI, to get a more comprehensive picture of an individual's current risk level. Therefore, in addition to the univariate prediction models presented in the previous section, it is also important to evaluate how accurately a combination of these measures predict recidivism. The question we are seeking to answer remains the same though: does reassessment improve the prediction of imminent recidivism?

In Table 4.15, we present the results of three discrete-time hazard models predicting time to recidivism. The three models presented in this table are slightly different from the models presented in the univariate analyses. Model A includes only the RoC*RoI scores. In Lloyd's (2015) dataset, the strongest predictor of short-term recidivism was the RoC*RoI. As a single predictor, Lloyd found that the RoC*RoI explained 44% of the variance in the predictor and the individual DRAOR subscales explained between 12% and 15% of the variance. In our dataset, the RoC*RoI was a significant predictor, but the effect sizes indicated that the RoC*RoI was a much weaker predictor than it was in Lloyd's dataset. This result was not unexpected, given the truncated range of RoC*RoI scores in our sample. Less expected was the finding that the univariate RoC*RoI model was only slightly more accurate at predicting recidivism than the univariate reassessment models for the individual DRAOR subscales, presented in the previous section; in Lloyd's dataset there was a much more substantial difference between the RoC*RoI and the individual DRAOR subscales, even using proximal scores.

Model B and Model C in Table 4.15 represent the two multivariate prediction models. Model B includes the RoC*RoI scores and the baseline scores for all three DRAOR subscales; Model C includes the proximal scores for the DRAOR subscales instead of the baseline scores. Although the RoC*RoI remained a significant predictor in both models, inclusion of the subscales added incremental predictive validity over the

model that included only the RoC*RoI. The effect sizes and model fit statistics also supported this finding. Both the baseline and reassessment models had lower AIC values and larger effect sizes than the RoC*RoI-only model (although the BIC for the baseline multivariate model was higher). These results suggest that including DRAOR subscales improved the prediction of imminent recidivism; in other words, dynamic variables added incremental predictive validity over static variables. It is also interesting to note that, of the three DRAOR subscales, baseline stable scores were the only significant predictor in the baseline model, while both stable and the acute proximal scores were individually significant predictors in the reassessment model.

The finding that models including DRAOR subscale scores and the RoC*RoI are better predictors than solely RoC*RoI scores is further supported by Figure 4.11 and Figure 4.12. These two graphs compare the time-dependent AUC values across time for the RoC*RoI-only model against the multivariate baseline and reassessment models. Similar patterns are evident in the two graphs. In both graphs, the AUC values were very similar for the first 8 or 9 weeks of the follow-up, which then separated and ran largely in parallel for the rest of the follow-up. The AUC values for the multivariate baseline model were only slightly higher than the RoC*RoI, and the confidence intervals showed that none of the differences were statistically significant. In contrast, the difference between the AUC values for the multivariate reassessment model and the RoC*RoI model were much larger, with a period of 5 weeks in the middle of the follow-up (from week 10 to 14) where the confidence intervals did not overlap.

Comparison of the multivariate models suggests that the proximal subscale scores were a better predictor of imminent recidivism than the baseline scores. As Table 4.15 shows, the model fit statistics were lower, and the effect sizes were larger for the proximal model than the baseline model. Also, when the proximal scores were

Table 4.15

*Results from Three Discrete Time Hazard Models Predicting Time to Recidivism Using Scores from the RoC*RoI and the DRAOR Subscales.*

Parameters	Model A		Model B		Model C	
	<i>B (SE)</i>	OR [95% CI]	<i>B (SE)</i>	OR [95% CI]	<i>B (SE)</i>	OR [95% CI]
Constant	-8.95*** (.85)	.0001 [.00003, .001]	-9.16*** (.91)	.001 [.00002, .001]	-9.23*** (.90)	.0001 [.00002, .001]
Week	.20** (.07)	1.22 [1.07, 1.39]	.19** (.07)	1.22 [1.07, 1.38]	.21** (.07)	1.24 [1.09, 1.41]
RoC*RoI	7.02*** (1.07)	1119.1 [136.5, 9171.5]	6.55*** (1.08)	697.55 [84.0, 5793.0]	6.30*** (1.08)	542.88 [65.8, 4481.1]
RoC*RoI x Week	-.26** (.09)	.77 [.65, .91]	-.25** (.09)	.78 [.66, .92]	-.26** (.09)	.77 [.65, .91]
Stable (Baseline)			.07* (.03)	1.07 [1.01, 1.13]		
Acute (Baseline)			.04 (.02)	1.04 [.99, 1.09]		
Protect (Baseline)			-.03 (.03)	.97 [.92, 1.02]		
Stable (Proximal)					.06* (.03)	1.06 [1.00, 1.12]
Acute (Proximal)					.10*** (.02)	1.10 [1.05, 1.15]
Protect (Proximal)					-.04 (.03)	.96 [.91, 1.01]
R ² (XO)		.18		.23		.27
c-index		.59		.61		.64

Deviance	4155.13	4127.67	4088.53
AIC	4163.13	4141.67	4102.53
BIC	4194.02	4195.71	4156.57
$\Delta \chi^2$ Model B vs. Model A		27.47*** (3)	
$\Delta \chi^2$ Model C vs. Model A			66.60*** (3)

Note. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria. **Bolded values** indicate statistically significant results. Model A predicts time to recidivism using only static RoC*RoI scores at every time point, Model B predicts time to recidivism using RoC*RoI scores and baseline DRAOR scores at every time point, and Model C predicts time to recidivism using RoC*RoI scores and proximal DRAOR scores at every time point.

* $p < .05$, ** $p < .01$, *** $p < .001$

Figure 4.11. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for RoC*RoI scores and a model including RoC*RoI scores and baseline scores on the three DRAOR subscales. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid.

Figure 4.12. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for RoC*RoI scores and a model including RoC*RoI scores and proximal scores on the three DRAOR subscales. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid.

Figure 4.13. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for multivariate models including RoC*RoI scores and either baseline or proximal scores on the three DRAOR subscales. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid.

added to the baseline model (not shown), significant incremental predictive validity was observed. Of all the models presented so far, the multivariate reassessment model demonstrated the most accurate prediction of imminent recidivism. However, note that when the time-dependent AUC values for the two multivariate models were graphed in Figure 4.13, the reassessment model had higher AUC values, but the confidence intervals overlapped in every week of the follow-up.

Discussion

Summary of findings.

The primary aim of this chapter was to complete a replication of the first step of Lloyd's (2015) three-step framework for examining whether reassessment enhances the prediction of imminent recidivism. First, we set out to describe the dataset and highlight similarities and differences with the dataset Lloyd used for his analyses. A life table showed that there were a sufficient number of recidivism events in each week to justify using a discrete-time hazard model with weeks as our discrete time period. We established that both inter- and intra-individual change was occurring to a small but significant extent, sufficient to justify testing whether reassessment might enhance prediction. The amount of change was greatest for the DRAOR acute subscale, but all subscales showed less change than they had in Lloyd's sample, indicating a smaller impact of reassessment on predictive accuracy might be expected. Various measures indicated that the subscales were being measured consistently across time. However, robust tests of longitudinal invariance did highlight some violations of invariance, suggesting a possible limitation of this dataset.

Second, we tested several assumptions of regression, and several assumptions relevant specifically to Cox regression and discrete-time hazard models. The data were largely consistent with these assumptions. When any violations of these assumptions

were identified, alternative models were run that either excluded problematic cases or used transformations of the problematic predictor variables. There was no indication that these alternative models resulted in meaningful changes to the substantive findings or conclusions, so these alternative models were not used in the final analyses.

Finally, we used Lloyd's (2015) approach for testing the extent to which reassessment enhances the prediction of imminent recidivism in an attempt to replicate his study. For each of the three DRAOR subscales, we tested univariate discrete-time hazard models that included either the baseline assessment score at every time point, or the most proximal score at that time point (i.e., the score obtained through reassessment). The results indicated that reassessment did improve the prediction of imminent recidivism, although the effects were small. The effect was strongest for the acute subscale, as demonstrated by increased effect sizes for the overall models, and significant differences in time-dependent AUC values during several individual weeks of the follow-up.

We also tested multivariate models, where a measure of static risk, the RoC*RoI, was simultaneously entered into models with scores from all three DRAOR subscales. Consistent with the univariate models, using the most proximal DRAOR scores resulted in more accurate prediction of imminent recidivism than using the baseline scores. The models including RoC*RoI scores and DRAOR scores were more accurate predictors than a model that included only the RoC*RoI scores. The most substantial difference was observed between the accuracy of the multivariate proximal assessment score model and the RoC*RoI-only model, with much larger overall effect sizes and significantly higher time-dependent AUC values in several weeks during the middle of the follow-up.

Conclusions.

The broader implications of the results in this chapter are considered in detail in the final chapter of this thesis. However, a few important points are worth noting here. First, these results were broadly consistent with the findings of similar previous studies, including Babchishin (2013), Brown et al. (2009), Greiner et al. (2015), Hanson et al. (2007), Howard and Dixon (2013), Jones et al. (2010) and Lloyd (2015). Our results provide further evidence that reassessment of dynamic risk and protective factors improves the prediction of imminent recidivism. Most importantly, these results are a successful replication of the findings in Lloyd's research using the framework that he proposed. His research aimed to develop a framework for testing the essential criteria of dynamic risk and protective factors. Our results suggest that when this framework was applied to a different dataset, findings were similar.

Second, our results also replicate Lloyd's (2015) finding that reassessment has the largest effect on predictive accuracy for acute risk factors. In contrast with previous research, which had largely failed to find evidence of a conceptual distinction between stable and acute risk factors (Brown et al., 2009; Hanson et al., 2007; Jones et al., 2010), Lloyd found that reassessment improved the prediction of imminent recidivism for the acute subscale more than for the other two DRAOR subscales. On that basis, we predicted that we would find a similar distinction in our results, and our results were consistent with that hypothesis. Reassessment enhanced prediction for all three subscales to a small extent, but the increase in overall effect sizes was largest for the acute subscale. The acute subscale was also the only one of the three DRAOR subscales to demonstrate significant differences in time-dependent AUC values during individual weeks of the follow-up. The broader implications of this finding for the concepts of acute and stable predictors are discussed in the final chapter of this thesis; for now, it is

sufficient to note that clear differences were observed between the subscales on the analyses conducted in this chapter.

Third, the predictive accuracy of our models was lower than the accuracy observed by Lloyd (2015). Both the baseline and reassessment univariate DRAOR subscale models in Lloyd's study explained close to 20% of the variance in the predictor (compared with 10% in our study), and predictive accuracy suggested moderate effects (concordance indices of 0.63-0.66) as opposed to the small effects that we found (0.55-0.60). For the multivariate models, the differences were more substantial, with Lloyd's models explaining over 50% of the variance and demonstrating strong predictive accuracy. Our models explained closer to 25% of the variance in the predictors and demonstrated moderate predictive accuracy. These discrepancies, particularly for the multivariate models, were most likely due to the use of a sample covering a more restricted range of static risk scores.

Interestingly though, the *differences* in effect sizes observed in our study between baseline and proximal models was very similar to the differences found in Lloyd's (2015) study. Reassessment of the dynamic variables appeared to have about the same effect in our high-risk sample as it did in Lloyd's sample, which included individuals across all static risk groups. The exception to this point was the time-dependent AUC values. We only found significant differences in these values for the acute subscale, and only for a few weeks in the middle of the follow-up. Lloyd found significant differences for all three subscales, and across a much greater number of weeks. However, this finding may have been due to the fact that Lloyd's sample was almost three times as large as the sample used in this thesis.

Next steps.

Overall, the finding of small effect sizes is consistent with much of the previous research in this area. Lloyd's (2015) study remains the strongest empirical evidence of reassessment of dynamic risk and protective factors enhancing the prediction of imminent recidivism. The smaller effects seen with our dataset can potentially be explained by our choice of sample, but it is also possible that the analyses in this chapter tested the predictor variables in a way that did not optimise their predictive validity. Put simply, predicting imminent recidivism using the single most proximal score may not represent the most accurate method of prediction. That possibility is tested in the three chapters that follow, starting with completion of the replication of Lloyd's research. In this chapter, we have tested the first step in his framework. In the next chapter, we critically examine and test the second and third steps in the framework, which focus on the idea of measurement aggregation.

Chapter 5: Does Aggregation Enhance Prediction?

In the previous chapter, we tested the first step in Lloyd's (2015) three-step framework for examining whether reassessment of dynamic risk and protective factors improves the prediction of imminent recidivism. The first step involved comparing baseline scores from the assessment closest to release to the scores from the most proximal assessment at each time point of the follow-up. Consistent with Lloyd's findings, we found evidence that the single most proximal score was a better predictor of imminent recidivism than a baseline assessment score, thus replicating the overall conclusion that reassessment enhances prediction. We also replicated the finding that reassessment of acute risk factors improved prediction more than for stable risk and protective factors.

The second and third steps in Lloyd's framework focus on the principle of measurement aggregation. In brief, this principle states that aggregation across measurement occasions can reduce measurement error. Reductions in measurement error should result in more accurate estimates of the 'true' value of the variables being measured, and more accurate estimation should lead to more accurate prediction of outcomes associated with those variables. However, it is unclear whether that argument can be applied to risk assessment and to the concepts of dynamic risk and protective factors. Aggregation may also have a differential impact for different types of dynamic variables, specifically acute and stable variables.

In this chapter, we discuss the principle of aggregation in more detail and examine whether aggregating reassessments of dynamic risk and protective factors may improve the estimation of those variables and enhance the prediction of imminent recidivism. There is limited theoretical and empirical literature in this area, but the few studies that have previously explored these issues are discussed. Following that

discussion, we present analyses that represent a test of the second and third steps in Lloyd's framework. The results of those analyses are set out and the implications of those findings are discussed.

Theoretical Background

When attempting to predict future behaviour, it is well-established in personality psychology that multiple ratings of previous behaviour are required (Epstein, 1979). A single instance of behaviour is considered too specific to the situation and too prone to measurement error to be used by itself to predict future behaviour (Epstein, 1983). Instead, it is strongly recommended that behavioural observations for the same individual should be aggregated across time, situations, observers, and measures (Epstein, 1980, 1986). These ideas have been highly influential in many areas of psychology, including risk assessment. The aggregation of multiple risk factors (as opposed to relying on single risk factors) is already a core principle of risk assessment (Mann et al., 2010).

Chu, Thomas, Daffern, and Ogloff (2013) demonstrated the benefits of aggregation in risk assessment with a sample of forensic psychiatric inpatients. They examined the predictive validity of the mean score, peak score, and single most recent score from daily behaviour ratings on a dynamic risk assessment instrument over the course of a week. The measure used in that study—the Dynamic Appraisal of Situational Aggression (DASA; Ogloff & Daffern, 2006)—required assessors to rate whether a small set of dynamic risk factors were present or absent in the previous 24 hours. Over a 6-month follow-up, mean scores (and peak scores) were found to be more strongly predictive of interpersonal violence, verbal threats, and inpatient aggression than the single most recent score, suggesting aggregation had enhanced prediction.

This type of research suggests that aggregation can provide a more accurate measure of an underlying variable that influences behaviour in different situations. In theory, aggregation reduces the measurement error associated with a single behaviour rating. However, the principle of aggregation may not apply in the same way if the aim is to explicitly measure the underlying variable. For example, if the aim is to measure the trait of impulsivity, one approach would be to record the presence or absence of impulsive behaviour every day for a week, as Chu et al. (2013) did using the DASA. Averaging those behaviour ratings should provide a more accurate measure of the underlying trait of impulsivity than the rating from any single day, even the most recent day. However, if assessors were instead asked to complete a week of daily ratings of the *trait* of impulsivity, assessors would need to consider all previous instances of impulsive behaviour. Use of this approach would mean that each rating would be an implicit average of the presence or absence of previous impulsive behaviours over that time period, with each new rating able to consider more recent behaviour.

Dynamic risk assessment tools primarily recommend the latter approach. To measure dynamic risk and protective factors, most risk instruments suggest looking for a pattern of behaviour, rather than the presence or absence of a single instance of a behaviour during a specified time period. For example, in the DRAOR manual, when rating items, assessors are encouraged to look for “evidence that is consistent and forms a pattern over time and place” (Serin et al., 2012, p. 4). Consequently, each individual assessment of a variable such as impulsivity is likely to be functionally similar to an average of a series of ratings of the presence or absence of impulsive behaviours.

Even if an assessment of a dynamic variable is already an (implicit) aggregate measure, aggregation of several reassessments could still improve estimation of the variable. Each assessment will have measurement error associated with the

measurement occasion: this error variance could be attributed to a range of factors including the mood of the assessor, time constraints on completion of the assessment, or the offender failing to report relevant information or reporting false information. Therefore, aggregation across measurement occasions should help to reduce measurement error and provide a more accurate measure of the relevant variables, and more accurate prediction of outcomes associated with those variables.

A more substantial problem for aggregation of dynamic risk reassessments arises when the underlying variable being measured can change over time. Aggregation across measurement occasions will be most effective at reducing measurement error when the variable being measured does not change between assessments. When the underlying variable does not change, any change in scores across assessments can be attributed to measurement error. When this is the case, as more assessments are completed, the average should increasingly approach the “true” value of that underlying variable. However, when the underlying variable can change over time, variation across measurement occasions can no longer be attributed solely to measurement error. Most likely, some of the variation will be due to measurement error, but some will reflect true change in the underlying variable. Thus, aggregating a series of assessments of a dynamic variable should not be viewed as an approximation of the true value of that variable. Instead, the average should be viewed as a reflection of the mean level of that variable over the specified period of time, an entirely different construct.

Both stable and acute dynamic variables are theoretically able to change over time. As mentioned previously, stable variables have been conceptualised as traits (Beech & Ward, 2004) or propensities (Mann et al., 2010; Thornton, 2015) for future offending that are expected to change relatively slowly (Hanson & Harris, 2000). Thus, averaging stable variables across multiple reassessments could improve their

estimation if the assessments occur more frequently than the period over which the variables are expected to show change (i.e., over a period where no change occurs in the underlying variable). For example, if change is expected to occur over several months, averaging scores from weekly assessments could result in more accurate estimation of the true value of those variables. The problem with this argument is that although stable variables are expected to change gradually, they are not expected to change in discrete time intervals (e.g., once every 3 months). In fact, very little is known about the process of change of stable variables such as antisocial attitudes or identity (Serin et al., 2013). Even an average of only a few assessments may include some amount of true change. Theory would therefore suggest that, based on the existing conceptualisation of stable variables (both risk and protective), it remains possible but unlikely that aggregation may improve the estimation of stable variables.

Aggregation is even less likely to improve the estimation of acute variables. Recall that acute variables have been conceptualised as states (Beech & Ward, 2004), whether internal mental states (e.g., mood) or external situational states (e.g., employment status), that are much more dynamic than stable variables (Hanson & Harris, 2000); Hanson et al. (2007) suggested change may occur over a few weeks, days, or even hours. Under this conceptualisation, even daily assessments might be capturing true change in acute variables. If true change is occurring between assessments, averaging across assessments is very unlikely to improve estimation of the true value of those variables, nor improve the ability to predict imminent recidivism using those variables.

A final issue with the argument that aggregation may improve prediction is worthy of mention. Lloyd (2015) speculated that assessors may implicitly average previous assessments when completing a reassessment. If this is true, the most

proximal score may not accurately represent the current true value of the variable being measured, as it may too heavily weight previous ratings. This issue would appear to be more relevant for acute variables, where current states are rated, as opposed to the trait-based stable variables, which are unlikely to be affected by accounting for past information (and for which raters are actively encouraged to factor in past behaviour). For example, a rating on current employment status that includes past employment history would make the current rating less accurate and would be likely to decrease the predictive accuracy of that rating. We might speculate that if this process was occurring, it would reduce the accuracy of the single most proximal score, making it more likely that aggregation will improve prediction. However, this is speculation. Research examining the process assessors follow when completing a reassessment is needed. In the absence of such research, this issue is acknowledged as a limitation of the current research. Further possible implications of this limitation are discussed in more detail in the final chapter of this thesis.

Empirical Research

There is limited existing empirical research testing whether aggregation of dynamic risk and protective factors enhances predictive accuracy. One study that examined the predictive value of an average of dynamic risk factors was the prospective, multi-wave study of a large sample of sex offenders in Canada by Hanson et al. (2007), discussed briefly in Chapter 2. There, it was noted that they had found little to no evidence that a single assessment of either stable or acute variables conducted closer in time to recidivism was more accurate than a single, earlier assessment. That study also examined the predictive validity of average ratings of the acute variables across measurement occasions. They compared the single most proximal rating (occurring within 45 days of recidivism) of each acute variable, to the average of all

ratings done within 45 days, 90 days, or 6 months of recidivism. They found that predictive validity gradually improved as averages from longer periods of time were used to predict recidivism. These results suggested that aggregation may improve the prediction of imminent recidivism compared to using the single most proximal assessment score.

In contrast, Lloyd (2015) found consistent evidence that averaging across assessments did not improve prediction compared to the single most proximal assessment score, for acute risk factors, and stable risk and protective factors. He compared the single most proximal assessment score to (a) a rolling mean of all existing assessment scores (step two in his framework), and (b) a moving rolling mean of between two and eight of the most proximal assessment scores (step three). There was no evidence that the rolling mean scores and the moving rolling mean scores were more accurate predictors of imminent recidivism than the single most proximal score. For the moving rolling mean scores, for all three subscales, model fit gradually improved as fewer assessments were included in the averages. In other words, the single most proximal score was a slightly better predictor than the average of the two most proximal scores, which was a better predictor than the average of the three most proximal scores, and so on. The only exception was the protective factors: compared to the single most proximal score, slightly better model fit and predictive accuracy was obtained when the two most proximal assessments were averaged compared to just the single most proximal assessment.

There are several possible explanations for the contrasting findings of Lloyd (2015) and Hanson et al. (2007). Differences in the statistical approaches used in the two studies are likely to explain some of the difference. The strongest explanation is that the acute variables in Lloyd's study showed more change between assessments

than the acute variables in Hanson et al.'s study.¹⁶ Even the stable risk and protective factors in Lloyd's study demonstrated greater change than the acute variables in Hanson et al.'s study. The "acute" variables in Hanson et al.'s study behaved more like stable or even static risk factors. This limited amount of change may explain why aggregation improved prediction. When the acute variables demonstrated more change, as happened in Lloyd's study, the more expected outcome was observed: the single most proximal score predicted better than the average scores. The two studies also used different sample groups (general offending versus sexual offending). It is possible this factor may have affected the results, or been a moderating factor (i.e., sex offenders may live more stable lives than general offenders, thus explaining the reduced amount of change observed in the Hanson et al. study), but that idea requires examination.

Evidence presented earlier in this thesis also supports the idea that aggregation of dynamic risk assessments may not improve prediction. In Chapter 3, we tested different methods for imputing scores in weeks where more than one assessment had been completed. For the acute subscale, using the score from the end of the week (i.e., the more proximal score) provided better model fit than averaging. For the stable and protective subscales, there was evidence that averaging scores from within the same

¹⁶ This difference cannot be explained by the different acute variables used because several of the acute variables in the two studies were the same (this is not coincidental; the development of the DRAOR was strongly influenced by the Acute-2000). Access to victims, social support, hostility, and substance abuse are all included on both the Acute-2000 (Hanson et al., 2007) and the DRAOR (Serin, 2007). While some items (e.g., social support) are conceptually distinct despite having similar names, the others are conceptually consistent. Thus, the observed differences in the amount those variables changed over time seems more likely to be the result of measurement and assessment practices (and possibly the sample group) as opposed to conceptual differences between the variables used in the two studies. This point also reinforces the importance of evaluating descriptive statistics of any variables classified as "dynamic", to ensure they are in fact behaving dynamically before drawing any conclusions about those variables' relationship with recidivism.

week slightly improved model fit compared to using only the most proximal assessment. These findings support the idea that averaging stable variables may improve prediction when multiple assessments are conducted within a short period of time, whereas averaging acute variables across assessments does not affect prediction. It should be noted that the differences in model fit obtained from the different methods were very small for all three subscales.

In summary, application of the principles of measurement theory to the concepts of dynamic risk and protective factors suggests aggregation across measurement occasions is unlikely to improve the prediction of imminent recidivism compared to the single most proximal assessment of those variables. According to their conceptualisation, acute variables should be very unlikely to benefit from aggregation. It is possible that stable variables may see some benefit from aggregation, but even this seems unlikely. Empirical evidence testing these ideas is very limited. The two studies (Hanson et al., 2007; Lloyd, 2015) that have conducted the most rigorous testing thus far found markedly different results. Although there are several plausible explanations for these differences, it is clear more research is needed on the application of the principle of aggregation to dynamic risk assessment.

The Current Study

We aimed to replicate Lloyd's approach for testing whether aggregation across multiple reassessments improves the prediction of imminent recidivism using our dataset. Given the limited empirical evidence for this question, any hypotheses were made tentatively. Consistent with Lloyd's findings and the theoretical background set out above, we expected that aggregation would not improve prediction for any of the three DRAOR subscales. Both a rolling mean score and moving rolling mean scores of between two and eight of the most proximal assessments were not expected to be

stronger predictors of recidivism than the single most proximal score. We thought it was possible that the stable and protective subscales may benefit from aggregation across a few weeks. Thus, for these two subscales, we thought the rolling mean scores would not improve prediction, but the moving rolling means may produce some improvement.

Preparation for Analyses

For these analyses, the same dataset described in detail in Chapter 3 was used and the dataset was structured in the same way. To test the value of aggregation, eight additional predictor variables were created for each DRAOR subscale: a *rolling mean* and seven *moving rolling means*, using between two and eight weeks of assessment scores. Rolling means were calculated for each week by taking the mean of every score from the same individual up to and including the current week. For example, the rolling mean in week 2 was the mean of scores in week 2, week 1, and week 0 (baseline). Moving rolling means were calculated for each week by taking the mean of the specified number of the most proximal assessments. For example, the ‘3-week rolling mean’ in week 4 was the mean of scores in week 4, week 3, and week 2; in week 5, it was the mean of scores from week 5, week 4, and week 3. The same method was used for all three DRAOR subscales.

In Table 5.1, we present a fictional example to illustrate how the different mean scores were calculated. In the example, observed scores on the acute subscale decrease gradually over the course of the follow-up. This pattern is broadly representative of what would be expected from a typical individual given the results presented in the previous chapter showing that DRAOR scores on the two risk subscales tended to gradually decline over time, and scores on the protective subscale tended to gradually increase. Table 5.1 shows the impact of that pattern of observed scores on the mean

score variables we calculated for the analyses in this chapter. Essentially, when observed scores decrease over time, mean scores will also decrease, but the decrease will occur more slowly than the decrease in observed scores. In other words, the mean scores will lag behind the observed scores. The higher the number of assessments included in the mean score, the greater the lag (i.e., the more slowly the decrease will occur). Thus, we can see in Table 5.1 that in every week, the rolling mean has the highest value, and the moving rolling means have gradually lower values as fewer weeks are included in their calculation.

Table 5.1 also highlights an issue with the mean score analyses. During initial weeks, there was an insufficient number of scores to calculate the moving rolling means. For example, technically it was not possible to calculate an 8-week mean score until after eight scores were available. One solution to this issue was to exclude all men who were censored or recidivated before a sufficient number of assessments had been completed. For example, for analyses using 8-week mean scores, we could exclude all participants who were censored or recidivated before week 8 of the follow-up. This approach would ensure that all men had the minimum number of assessment scores to calculate the correct predictor variable. An alternative solution was to use rolling mean scores during the initial weeks when insufficient assessments had been completed. For example, for 8-week mean scores, a rolling mean could be calculated for the first 7 weeks until the minimum eight assessments had been completed. Using this approach would mean that '8-week mean scores' are technically a combination of 8-week means and rolling means. Lloyd (2015) used the former approach to deal with this issue, excluding all participants with fewer than 8 weeks of assessment; it was unclear how the issue was addressed by Hanson et al. (2007).

Table 5.1

Fictional Example of a Single Individual's Score on the DRAOR Acute Subscale Across the First 13 Weeks of the Follow-up, and the Rolling Mean, and Moving Rolling Mean Scores Calculated from those Observed Scores.

Week	Observed score	2-week mean	3-week mean	4-week mean	5-week mean	6-week mean	7-week mean	8-week mean	Rolling mean
0	12								12.0
1	12	12.0							12.0
2	11	11.5	11.7						11.7
3	9	10.0	10.7	11.0					11.0
4	10	9.5	10.0	10.5	10.8				10.8
5	9	9.5	9.3	9.8	10.2	10.5			10.5
6	9	9.0	9.3	9.3	9.6	10.0	10.3		10.3
7	9	9.0	9.0	9.3	9.2	9.5	9.9	10.1	10.1
8	8	8.5	8.7	8.8	9.0	9.0	9.3	9.6	9.9
9	8	8.0	8.3	8.5	8.6	8.8	8.9	9.1	9.7
10	6	7.0	7.3	7.8	8.0	8.2	8.4	8.5	9.4
11	6	6.0	6.7	7.0	7.4	7.7	7.9	8.1	9.1
12	6	6.0	6.0	6.5	6.8	7.2	7.4	7.6	8.8
13	4	5.0	5.3	5.5	6.0	6.3	6.7	7.0	8.5

Note. Blank cells represent the weeks where an insufficient number of assessments had been completed so the models using the moving rolling means instead used the rolling mean scores as the predictor variable.

We wanted to simultaneously compare the rolling means and moving rolling means, rather than comparing each variable separately against the single most proximal score. For this comparison to be a valid test of the best model, each model needed to be tested on the same sample group. If we chose to use the exclusion rule for the moving rolling means, this would have meant excluding men who were censored or recidivated before week 8 from all analyses. In our dataset, 227 men (23.5% of the total sample) were censored or recidivated before week 8; removing these men would have seen 37.8% of all recidivists removed and 36.7% of all assessments. Even though our sample would remain large without these men, we decided the best approach was not to exclude them from the analyses. The fact that failing to examine the predictive validity of a measure during the initial weeks, which are known to be a very challenging period for most men coming out of prison with high recidivism rates, would limit the ecological validity of our results had a strong influence on our decision. Therefore, for the moving rolling mean scores, during the initial weeks, rolling average scores were used until a sufficient number of scores was available.¹⁷

Results

Univariate discrete-time hazard models were run for the eight mean score variables for all three DRAOR subscales. In every model, the predictor variable was statistically significant. Therefore, the focus was on comparing the model fit, measured using the Akaike Information Criterion (AIC), and two measures of predictive accuracy introduced previously: Xu and O'Quigley's (1999) R^2 measure, and Heagerty and Zheng's (2005) concordance index (c -index). The results on all three of these measures

¹⁷ The analyses were also run excluding men who were censored or recidivated before week 8 and excluding men who were censored or recidivated before week 4 (here, we tested only 2-, 3-, and 4-week mean scores). The substantive findings and conclusions did not differ from those presented below.

for all the different models are presented in Table 5.2, alongside the results obtained using only the single most proximal assessment reported in the previous chapter.

Table 5.2 shows that, for the stable subscale, the evidence suggested that aggregation did not improve the prediction of imminent recidivism. The best model fit and the largest effect sizes were obtained from the model using only the single most proximal score. The poorest model fit and smallest effect sizes were obtained from the model using the rolling mean score. These results formed part of a general pattern suggesting mean scores across fewer weeks were stronger predictors of recidivism than mean scores across a higher number of weeks. However, there was very little difference between the models; the rolling mean score explained roughly 2% less variance than the single most proximal score, with the moving rolling mean scores falling in between.

The pattern of results was very similar for the acute subscale. There was a gradual decline in predictive accuracy as more weeks were included in the mean scores, the single most proximal score was the strongest predictor, and the rolling mean was the weakest predictor. However, the differences between models were more substantial for the acute subscale, with the single most proximal score explaining double the amount of variance (about 6% more) than the rolling mean score.

Results for the protective subscale were less consistent across the three measures used. The best model fit was obtained using the 2-week mean scores, the largest R^2 was obtained using the 4-week mean score, and the largest c -index was obtained using the single most proximal score. Similar to the stable subscale though, the differences between the models were minimal. Consistent with the other two subscales,

Table 5.2

Model Fit (AIC) for Discrete-Time Hazard Models and Effect Size for Cox Regression with Time-Varying Predictor Models for Rolling Mean and Moving Rolling Mean Scores for the Three DRAOR Subscales.

Predictor variable	Stable			Acute			Protective		
	AIC	R^2 (XO)	c-index	AIC	R^2 (XO)	c-index	AIC	R^2 (XO)	c-index
Most proximal score	4161.75	.12	.60	4154.14	.11	.60	4166.22	.12	.59
2-week mean	4162.36	.12	.60	4157.73	.11	.60	4166.20	.12	.59
3-week mean	4163.85	.12	.60	4163.71	.09	.60	4167.09	.12	.59
4-week mean	4166.05	.12	.59	4168.25	.08	.59	4168.54	.12	.59
5-week mean	4167.44	.12	.59	4170.03	.07	.59	4169.38	.12	.59
6-week mean	4168.11	.12	.59	4172.14	.07	.59	4170.27	.12	.58
7-week mean	4168.82	.12	.59	4173.84	.06	.59	4171.49	.12	.58
8-week mean	4169.47	.11	.59	4174.28	.06	.59	4172.48	.12	.58
Rolling mean	4173.74	.11	.59	4178.79	.05	.58	4178.37	.11	.58

Note. AIC = Akaike Information Criteria. **Bolded values** indicate the model with the best fit, or largest effect size.

there was also an overall pattern suggesting predictive accuracy improved as fewer scores were included in the means.

Discussion

In this chapter, we have looked at the impact of aggregating reassessments of dynamic risk and protective factors. The review of the existing theoretical and empirical literature suggested there may not be a strong rationale for expecting that aggregation would enhance the prediction of imminent recidivism beyond the level of accuracy obtained by using only the single most proximal assessment score. The only area where a small benefit might have been expected was for stable variables; acute variables were not expected to benefit from aggregation of reassessment scores, and instead it was hypothesised that aggregation would decrease predictive accuracy. However, we also noted that there is a substantial dearth of research in this area and indicated that the question required further investigation.

We investigated the issue of aggregation by looking at the extent to which mean scores on the three subscales of the DRAOR improved the prediction of imminent recidivism. Several different mean scores were calculated, and these scores were then used in univariate discrete-time hazard models to assess their predictive accuracy. Measures of predictive accuracy were then compared to the predictive accuracy obtained from models that used the single most proximal assessment score, the results of which were presented in full detail in the previous chapter. This approach represented a test of the second and third steps in the framework proposed by Lloyd (2015) for examining whether reassessment improves the prediction of imminent recidivism, and a further replication of his research testing that framework.

Our results largely supported our hypotheses. Consistent evidence was found indicating aggregation of dynamic risk and protective factors did not improve the

prediction of imminent recidivism. Mean scores, calculated using all existing previous assessments or only a few of the most proximal assessments, were not better predictors of recidivism than the single most proximal score. For the acute subscale in particular, there was evidence that the most proximal score was a better predictor than any mean score, and predictive accuracy appeared to decrease as more weeks were included in the mean scores. For the stable and protective subscales, the single most proximal score was a better predictor than the rolling mean score. The difference between the most proximal score and the moving rolling mean scores for these two subscales was not as substantial, but there was still evidence of a general pattern favouring proximity over aggregation.

The results are consistent with Lloyd's (2015) results and stand in contrast to the results of Hanson et al. (2007). Hanson et al.'s results suggested averaging dynamic risk scores over longer periods of time may enhance prediction compared to using only the most proximal score. Like Lloyd, we found no evidence to support that idea. Instead, our results suggested aggregation does not improve prediction and may instead have a detrimental effect on predictive accuracy, particularly for a scale of acute variables. It is important to note that the two studies that have found divergent results from Hanson et al. have both used DRAOR data from a sample of New Zealand parolees. It remains possible that the findings of Lloyd's study and of the current study may be specific to either this risk assessment tool or sample group or both. It is strongly recommended that further research be conducted with different sample groups and different risk assessment tools to establish the generalisability of these findings.

The most likely explanation for our results is that *true* change in the predictor variables occurred between reassessments. Measurement theory suggests aggregation will most improve estimation when estimates differ across measurement occasions but

the underlying variable being measured does not change during that time. If the underlying variable does change, aggregation is unlikely to improve estimation and may even worsen it. In our study, aggregation had almost no effect for the stable risk and protective variables. We saw in the previous chapter that although intra-individual change did occur over the course of the follow-up, the amount of change occurring was limited. In contrast, substantially more intra-individual change was observed on the acute subscale. When the acute subscale was aggregated across reassessments, predictive accuracy decreased, more so as more reassessments were included in the aggregate measures. This finding is consistent with an explanation that true change was occurring in the acute variables, ensuring that aggregate measures which included previous assessments provided poorer estimates of those variables than the most proximal assessment on its own.

Theoretically, these results are consistent with existing conceptualisations of dynamic risk and protective factors, and of acute and stable dynamic variables. Taking an average of several measurements should not improve the estimation of the current value of that variable if acute variables are internal states or current external circumstantial factors (Beech & Ward, 2004). Similarly, for stable variables, although some benefit might have been expected from aggregation of propensity or trait-type variables (Mann et al., 2010; Thornton, 2015), the dynamic nature of those variables still made it unlikely that aggregation would be beneficial. Our results were consistent with these expectations and thus provide support for existing conceptualisations of these variables, as well as for the theorised distinction between acute and stable predictor variables.

Practically, these findings suggest that supervision officers should be focused most closely on the most recent assessment. The results presented in this chapter,

alongside Lloyd's (2015) results, provide consistent evidence that professionals working with the DRAOR should not be calculating average scores for individual offenders and using those averages to guide their supervision and management *ahead* of using the most proximal score. However, the differences in predictive accuracy between the most recent assessment and averages of the two or three most recent assessments, particularly for the stable risk and protective factors, were small. Therefore, the evidence would also support a recommendation that aggregation might be beneficial in a situation where the single most proximal assessment is thought to be unreliable for some reason.

In summary, these results provide support for the argument laid out at the start of this chapter that aggregation will not improve the estimation of dynamic risk and protective factors. Several different aggregate measures were tested, and none were shown to be better predictors of imminent recidivism than the single most proximal assessment. This chapter has focused on whether aggregation provides a *better estimate* of dynamic variables than a single, proximal assessment, and that question has been answered in the negative. What this chapter has not tested is whether these aggregate measures have incremental predictive value over the most proximal assessment score. Mean scores may not be better predictors than the most proximal score but they may add incremental predictive validity. In addition to mean scores, the amount of change and variability in a series of multiple reassessments may also add incremental predictive validity to the most proximal assessment score. In the following two chapters, we explore the rationale behind these ideas and complete a series of analyses to test the extent to which mean scores and other measures of intra-individual change, particularly change scores, when used alongside the most proximal assessment, enhance the prediction of imminent recidivism.

Chapter 6: Intra-Individual Change and Imminent Recidivism

In Chapter 1, we argued that, in addition to the clear practical benefits, testing whether reassessment of dynamic risk and protective factors enhances the prediction of imminent recidivism represents a robust method for testing the theoretical concepts of dynamic risk and protective factors. Variables must satisfy three criteria—an association with the outcome, change over time, and an association between change and outcome—to be accurately classified as dynamic risk or protective factors (Brown et al., 2009; Caudy et al., 2013; Jones et al., 2010). A finding that reassessment enhances prediction, as observed in Chapter 4, suggests variables meet all three of these criteria. In particular, this method establishes the third criterion, indicating that intra-individual change in these variables is associated with imminent recidivism.

The relationship between intra-individual change in dynamic variables and recidivism has more commonly been investigated using a different method. The majority of the research in this area has examined the association between change during some form of correctional rehabilitation or treatment, and recidivism following that treatment. This type of research aims to provide empirical support for the proposition that change in dynamic risk and protective factors is the mechanism through which treatment reduces recidivism. This proposition is central to the RNR model of effective correctional intervention (Bonta & Andrews, 2016). However, the idea that positive intra-individual change (i.e., decreasing risk and increasing protective factors) should reduce the likelihood of *an individual* reoffending is not the same as the idea that *individuals* who make positive change should be less likely to reoffend than other *individuals* who fail to make positive change or even make negative change. These two ideas need to be carefully distinguished as they have different practical and theoretical implications.

In this chapter, we explore the distinction between these ideas and lay out a method that we believe can better examine these ideas. The first step in that method is to establish the extent to which intra-individual change is associated with imminent recidivism. As mentioned, the results from Chapter 4 can be interpreted as providing evidence for this relationship. In other words, the incremental validity of the most proximal score over the baseline score will be equivalent to the incremental validity of raw change scores over the baseline scores. We demonstrate that point again, empirically, later in this chapter. However, the analyses in Chapter 4 only represent one approach for examining the relationship between intra-individual change and imminent recidivism. In this chapter, we look at several other change scores in addition to the total change between the baseline and most proximal assessment. Specifically, we explore the extent to which change scores *over different time periods* are associated with imminent recidivism. The rationale for testing change score across different time periods is explored in the introduction section.

The second step of our proposed method is to investigate the extent to which intra-individual change impacts not only on an individual's likelihood of recidivism but also on that individual's likelihood of recidivism relative to other individuals. Statistically, this step involves testing whether intra-individual change scores demonstrate incremental validity over the single most proximal assessment score. The aim is to see if intra-individual change is *uniquely* associated with recidivism. Alternatively, this aim can be viewed as a question of whether change matters (i.e., is the fact of change relevant per se, or is it only relevant because it allows calculation of proximal risk). In this respect, the aim is similar to the concept of clinically significant change (Jacobson & Truax, 1991), which also seeks to determine whether change really matters or whether change is insufficient and not meaningfully related to relevant

outcomes. In the introduction section of this chapter, we examine the theoretical rationale for *why* and *how* intra-individual change might be uniquely associated with imminent recidivism and summarise and discuss the limited empirical research relevant to this question that already exists.

Overall, the method we are proposing in this chapter represents an extension of the framework proposed by Lloyd (2015). His framework intended to provide a method for testing the extent to which reassessment of dynamic risk and protective factors enhances the prediction of imminent recidivism. That framework also tests the extent to which intra-individual change is associated with recidivism. Our proposed extension provides steps for further establishing the *most accurate* way of predicting imminent recidivism using reassessments of dynamic risk and protective factors. The extension also tests the extent to which intra-individual change is *uniquely* associated with recidivism. These additions to the framework will have important theoretical and practical implications; the implications of the findings obtained applying this framework to our dataset are discussed at the end of this chapter.

Introduction

Is intra-individual change associated with recidivism?

The concept of dynamic risk factors was originally proposed to provide an intermediate target for correctional rehabilitation (Andrews, Bonta, et al., 1990). Actuarial risk assessment using static risk factors had improved the ability to predict which individuals were most likely to recidivate compared to clinical judgement (Andrews et al., 2006; Hanson & Morton-Bourgon, 2009; Harris & Rice, 2015), but static factors could not be used as treatment targets (Bonta, 1996). In contrast, dynamic risk factors were variables that were associated with recidivism *and* could be changed through appropriate intervention. Consequently, the RNR model argued that

rehabilitation should target these changeable variables in order to reduce the likelihood of recidivism (Andrews, Zinger, et al., 1990).

There is now a large body of research showing that treatment that targets dynamic risk factors is effective at reducing recidivism (Bonta & Andrews, 2016; Smith, Gendreau, & Swartz, 2009). However, there is much less research establishing that the mechanism through which that treatment is effective is the amelioration of relevant risk factors (Serin et al., 2013). There are few high-quality studies that have found an empirical link between change in dynamic risk (or protective) factors and recidivism (Beggs, 2010; Serin et al., 2013). There is evidence that decreases in well-known dynamic risk factors such as antisocial associates, antisocial personality, and antisocial attitudes are associated with lower recidivism (Beggs, 2010; Serin et al., 2013). Overall though, results in this area have been inconsistent, and the literature is plagued by poor-quality studies (Serin et al., 2013).

One explanation for the inconsistent findings is that many studies have tested intra-individual change in isolation. In their reviews of the literature in this area, both Beggs (2010) and Serin et al. (2013) noted that most studies had calculated the association between treatment change and recidivism without controlling for pre-treatment level of risk. The problem with this approach has been neatly illustrated by Baglivio, Wolff, Jackowski, and Greenwald (2017). They provide the example of two individuals, one who demonstrates significant positive change across treatment (i.e., reduced risk) and one who makes no change during treatment. Based on change alone, the first individual might be expected to be less likely to reoffend than the second individual. However, in their example, the first individual started as very high risk, whereas the second offender was very low risk, meaning that despite the differences in treatment change, the second individual was still lower risk at the end of treatment.

This example shows how analyses that do not account for pre-treatment risk are likely to provide an incomplete understanding of the association between intra-individual change and recidivism. Controlling for static risk can partly mitigate this risk, but because static and dynamic risk are not perfectly correlated, pre-treatment dynamic risk is likely to affect the relationship between change in dynamic factors and recidivism. Investigation of the effect of intra-individual change on recidivism needs to control for pre-treatment dynamic risk.

In recent years, several studies have controlled for initial levels of dynamic risk and protective factors. Some of these studies have examined change during treatment, with treatment completed in prison (Beggs & Grace, 2011; Klepfisz, O'Brien, & Daffern, 2014; Lewis et al., 2013; Olver, Lewis, & Wong, 2013; Olver & Wong, 2011), forensic psychiatric hospitals (de Vries Robbé, de Vogel, Douglas, et al., 2015), or the community (Kroner & Yessine, 2013). Other studies have looked at change in dynamic risk factors during residential placement (Baglivio, Wolff, Jackowski, et al., 2017; Baglivio, Wolff, Piquero, DeLisi, & Vaughn, 2018; Baglivio, Wolff, Piquero, Howell, & Greenwald, 2017) or community supervision (T. H. Cohen, Lowenkamp, & VanBenschoten, 2016; T. H. Cohen & VanBenschoten, 2014; Howard & Dixon, 2013; Labrecque, Smith, Lovins, & Latessa, 2014; Vose, Smith, & Cullen, 2013; Wooditch, Tang, & Taxman, 2014). This latter group shows that the association between change and recidivism should not be limited to formal treatment programmes. Change can occur outside treatment or the context of a formal intervention and, since the variables being measured are identical concepts, change outside treatment should be associated with recidivism in the same way as change during treatment. In fact, the ecological validity of studies examining change outside institutional treatment programmes may be higher, as these studies can measure situational or contextual variables such as housing and employment that are

known to be associated with recidivism (Dickson & Polaschek, 2014; Kroner et al., 2013; Scoones et al., 2012), and enable study of how the change process continues in the community after release (Serin et al., 2013).

While the samples and settings have differed, in general these recent studies have followed a common method: they have examined whether intra-individual change in dynamic variables demonstrated incremental predictive validity over pre-treatment or baseline dynamic risk. Results from this growing body of research have been inconsistent. With the exception of Klepfisz et al. (2014), all of the studies cited above found at least some evidence of an association between change scores and recidivism, in the expected direction (i.e., reduced risk associated with lower recidivism). However, several studies found that change on the majority of the dynamic variables assessed was not associated with recidivism (Baglivio, Wolff, Jackowski, et al., 2017; Baglivio et al., 2018; Kroner & Yessine, 2013; Wooditch et al., 2014). Overall, while the field would benefit from further research, there is evidence to suggest that changes in dynamic variables may be associated with recidivism after controlling for initial levels of those variables.

These findings can be interpreted in two ways. They can be taken as evidence that intra-individual change in dynamic risk and protective factors is associated with recidivism. Alternatively, as Howard and Dixon (2013) concluded in their study, the findings can be viewed as evidence that the post-treatment (in rehabilitation studies) or most proximal assessment provides the most accurate indication of the likelihood of recidivism. These interpretations are equivalent because, to take an example from the rehabilitation literature, testing the incremental validity of a change score over a pre-treatment score will produce the same results as testing the incremental validity of a post-treatment score over a pre-treatment score. Statistically, the incremental validity

of both raw change scores (post-treatment minus pre-treatment score), and post-treatment scores over pre-treatment scores, will be identical. The difference is that the former model explicitly incorporates change scores, and the latter does not (Brown et al., 2009).

This convergence means that studies examining whether change is associated with recidivism and studies examining whether reassessment enhances prediction can be treated as equivalent. Both approaches provide evidence of whether intra-individual change is relevant to the prediction of recidivism. However, there is some evidence that the two approaches are viewed differently. In a study by de Vries Robbé, de Vogel, Douglas, et al. (2015), change scores significantly predicted violent recidivism after controlling for pre-treatment scores. The authors concluded that better progress in treatment was associated with lower recidivism (de Vries Robbé, de Vogel, Douglas, et al., 2015). This conclusion was justified on the basis that recidivists and non-recidivists had no significant difference in their pre-treatment dynamic risk and protective factor scores; therefore, change could distinguish recidivists and non-recidivists. However, rather than only presenting the results for change scores, additional analyses were presented testing whether post-treatment scores showed incremental predictive validity over pre-treatment scores. If the incremental validity of change scores and post-treatment scores over pre-treatment scores is equivalent, we would expect that testing the former would present identical results to a test of the latter. Indeed, de Vries Robbé et al. found that post-treatment scores demonstrated significant incremental predictive validity over pre-treatment scores. The duplicated analyses suggest that some unique value was placed on tests involving change scores. These analyses were not perceived as equivalent to analyses using post-treatment scores, when the most likely explanation

for both findings appeared to be that more proximal assessments provide more accurate estimates of the likelihood of recidivism.

Is intra-individual change *uniquely* associated with recidivism?

The approach taken in studies such as the one by de Vries Robbé, de Vogel, Douglas, et al. (2015) appear to suggest that intra-individual change might be *uniquely* associated with recidivism. The idea seems to be that change has an impact on the likelihood of recidivism, separate and additional to the level of dynamic risk after the change occurs. The example provided by Baglivio, Wolff, Jackowski, et al. (2017), highlighted earlier, is useful again here. Imagine two men show the same patterns of change during treatment—one making good progress and one making no progress—but this time they finish treatment with the same overall level of dynamic risk. If intra-individual change is associated with recidivism, as the evidence presented in the previous section would suggest, the man who made positive progress is now less likely to reoffend than when he started treatment, while the likelihood of the other man reoffending remains the same. But with the two men ending treatment with the same dynamic risk scores, the question is whether they currently pose the same likelihood of recidivism as each other. In the original example laid out by Baglivio and colleagues, despite their change, the two men remained at *different* risks of recidivism; in our revised example, the question is whether, despite their change, the two men present the *same* risk of recidivism.

To examine this question, the traditional paradigm for testing change scores needs to be inverted. Rather than testing whether change scores are associated with recidivism after controlling for pre-treatment or baseline risk, research is needed testing whether change scores are associated with recidivism after controlling for *post-treatment* or *proximal* dynamic risk. This approach could identify whether intra-

individual change is associated with recidivism for reasons other than the fact that it allows calculation of post-treatment or proximal dynamic risk.

A review of the existing literature reveals no studies that have utilised this method. However, there is an alternative method for testing this question. As de Vries Robbé, de Vogel, Douglas, et al. (2015) demonstrated, testing the incremental validity of change scores over pre-treatment scores is equivalent to testing the incremental validity of post-treatment scores over pre-treatment scores. The same basic equation—post-treatment score minus pre-treatment score equals change—means that testing the incremental validity of change scores over post-treatment scores will present the same results as testing the incremental validity of pre-treatment scores over post-treatment scores. Expanding the focus beyond the treatment context, rather than asking whether the most proximal score has incremental validity over a baseline score, as we have done previously in this thesis, we can ask whether the baseline score has incremental validity over the most proximal score.

There are also no studies that have intentionally examined this question, but some support can be inferred from studies that asked the opposite question. For example, the primary aim of the study by Lloyd (2015) was to test whether reassessment of dynamic risk and protective factors enhances the prediction of imminent recidivism. One of the methods he used to test that question was to examine the incremental validity of proximal scores over baseline scores, the approach that was taken in Chapter 4 of this thesis to answer the same question. As already discussed, Lloyd found that proximal scores demonstrated significant incremental predictive validity over baseline scores. He also presented the results from models where the two predictor variables were entered into models in the opposite order (i.e., proximal *before* baseline scores). In these models, baseline scores on all three subscales of the DRAOR

demonstrated significant incremental predictive validity over the proximal scores. The results suggested that, after controlling for the most proximal assessment score, higher risk and lower protective subscale scores at baseline were associated with higher recidivism. However, the effect size measures suggested the addition of the baseline scores did not improve the predictive accuracy above the level obtained from using only the most proximal score.

Of the other prospective, multi-wave studies that have tested whether reassessment improves prediction (Babchishin, 2013; Brown et al., 2009; Greiner et al., 2015; Hanson et al., 2007; Howard & Dixon, 2013; Jones et al., 2010; Viljoen et al., 2017), only Babchishin (2013) reported results relevant to this question. In that study, when tested separately, both initial and reassessment scores on the Acute-2007 were significantly associated with several recidivism outcomes. When the two variables were entered into models together, results differed depending on the recidivism outcome used. For sexual recidivism, the initial scores did not demonstrate incremental validity over the reassessment scores. For violent recidivism, the reassessment scores did not demonstrate incremental validity over the initial scores, rendering the question of whether the initial scores would show incremental validity irrelevant. For general recidivism, the initial scores showed incremental validity over the reassessment scores in four of the six measurement waves. Consistent with Lloyd (2015), higher initial scores were associated with higher recidivism when entered alongside reassessment scores.

Mulvey et al. (2016) examined the predictive validity of dynamic risk scores from a baseline assessment and another assessment 6 months later for a large sample of adolescents. Both the baseline and 6-month assessment scores significantly predicted self-reported reoffending and re-arrest over a one-year follow-up period. When the

scores were entered simultaneously, scores from the 6-month assessment were significant predictors of all outcomes. The baseline scores remained significant for self-reported reoffending—thus demonstrating incremental predictive validity—but not for the re-arrest outcome. Once again, higher baseline scores were associated with higher recidivism when entered alongside the 6-month assessment score.

Some of the studies discussed in the previous section in this chapter are also relevant here. T. H. Cohen et al. (2016), T. H. Cohen and VanBenschoten (2014), and Vose et al. (2013) all found evidence that, after controlling for initial dynamic risk, intra-individual change was associated with recidivism. However, they used risk categories as their control measure, rather than a continuous measure of risk. For example, in all three studies individuals considered low risk on an initial assessment and higher risk on a later assessment had higher recidivism rates than individuals who remained in the low risk category throughout. We can re-examine those findings to see whether individuals who finished in the same risk categories had different recidivism rates depending on the category in which they started. When the results were examined in this way, a consistent finding emerged. In none of the studies was there any evidence to suggest that initial risk category had a significant impact on the likelihood of recidivism compared to the final risk category; in other words, recidivism rates were very similar for individuals in the same risk categories at the later assessment, regardless of the category in which those individuals had started. It is important to note though, that this finding was not tested for statistical significance nor was an effect size calculated.

A similar pattern of results was evident in the study by Baglivio, Wolff, Jackowski, et al. (2017). They looked at risk assessment ‘trajectories’ during placement in a juvenile justice facility. Their results supported the idea that intra-individual change in dynamic risk and protective factors is associated with recidivism after

controlling for initial levels of those variables. For example, among the group who started with low protective factors, those who made greater gains during placement had significantly lower rearrest rates. When looked at in reverse, there was some evidence that risk trajectories with the same end point would have different likelihoods of recidivism: individuals on the trajectory with the lower starting point were less likely to be rearrested than those who started with higher protective factor scores. These differences were small though and not tested for statistical significance.

Overall, these studies provide an inconsistent answer to the question of whether intra-individual change is *uniquely* associated with recidivism. The results from Lloyd (2015) suggest there may be a significant unique association, but the size of the effect observed in that study was very small. The rest of the empirical evidence is either mixed (Babchishin, 2013; Baglivio, Wolff, Piquero, et al., 2017; Mulvey et al., 2016) or suggests there is no association (T. H. Cohen et al., 2016; T. H. Cohen & VanBenschoten, 2014; Vose et al., 2013). Interestingly, in all four studies where evidence was found suggesting intra-individual change has incremental predictive validity, higher previous risk and lower protective factor scores were associated with a higher likelihood of recidivism. This pattern of findings suggests that among individuals who have the same risk level at the most proximal assessment, individuals who start with higher risk scores will be more likely to recidivate than individuals who start with lower risk scores. With such limited evidence, little weight can be placed on this pattern of findings, but it warrants further investigation. Potential explanations for this pattern are explored in the next section. Also, the inconsistency in these findings is unsurprising given the markedly different approaches taken in these studies and the fact that none of the studies were intentionally setting out to examine the question of incremental validity of intra-individual change. There is a clear need for more empirical research that deliberately

sets out to examine to what extent intra-individual change is uniquely associated with recidivism and the nature of that relationship.

Why might intra-individual change be uniquely associated with recidivism?

None of the studies discussed in the previous section provided a theoretical rationale for why intra-individual change might show a unique association with recidivism. The closest to a rationale comes from studies that examine dynamic risk ‘trajectories’ (Babchishin, 2013; Baglivio, Wolff, Piquero, et al., 2017). These studies seem to suggest that change in risk can be considered a trajectory, and previous changes in risk will continue in the future along that same trajectory. The implication is that individuals on an improving trajectory are thus less likely to recidivate than individuals who are on a flat or worsening trajectory. But this idea is implicit from the language used, rather than explicit in these studies. Also, neither Babchishin (2013) nor Baglivio, Wolff, Piquero, et al. (2017) controlled for proximal risk when examining the predictive validity of the trajectories, nor discussed the incremental validity of trajectories, so it is difficult to evaluate the validity of this possible rationale.

In the absence of existing theory, hypothetical scenarios can help to illustrate situations where change in risk or prior risk level might have incremental predictive value over proximal risk. For example, imagine two men being supervised in the community who are both unemployed at their most recent assessment, suggesting they pose an equal risk of imminent recidivism. But what if one man has been unemployed since the beginning of supervision, whereas the other man only recently lost his job? It is reasonable to think that the recent job loss may place the second man at higher risk of imminent recidivism, as he may be struggling to cope with the other changes that come with unemployment such as loss of income, loss of identity, and the increase in free time to spend with anti-social peers. The man who has been unemployed for longer may

have already learned to manage these factors, making him less likely to imminently reoffend. In this scenario, it is possible to see how a measure of intra-individual change prior to the most recent assessment could enhance the prediction of future offending beyond the level obtained using only the most recent assessment.

One problem with this hypothetical scenario is that criminal risk assessment does not rely on single variables; rather, aggregation across multiple variables is a central tenet of risk assessment (Mann et al., 2010). In the hypothetical example, we suggested that changes in employment status might enhance the prediction of recidivism beyond what can be obtained using only the current employment status. But a thorough dynamic risk assessment will include a range of other risk (and protective) factors in addition to employment. When the other relevant variables are measured, the relevance of the change scores should disappear. For example, if, as suggested, the recently unemployed man had a more anti-social identity or was spending more time with anti-social peers following his job loss, he would have scored higher on those variables at the most recent assessment than the long-term unemployed man. Thus, the recently unemployed man would be considered more likely to reoffend because of his current risk scores, not his change in score. This example illustrates how differences in intra-individual change on one relevant variable should be captured by differences on other relevant variables at the most recent assessment.

Another problem is that proximal assessments may already, either explicitly or implicitly, account for the change that has occurred. In the previous chapter, we discussed the possibility raised by Lloyd (2015) that reassessments might be implicit averages of current and past behaviour, and thus of prior assessment scores. A similar argument can be made about change. The predictive validity of the most proximal assessment will not be enhanced by change scores if that change is already factored into

the proximal assessment score. For example, if the assessor rates the two men from the previous example differently on the basis of how long they have been unemployed, change scores on that variable would not provide any unique information. Assessors may incorporate this type of additional information because the assessment tool explicitly requires them to (i.e., it calls for weighing of current and past information during scoring), or because of unclear scoring guidelines or measurement error. Determining whether this implicit (or explicit) aggregation is occurring requires research examining the process assessors follow when completing a reassessment. Either way, the possibility of this approach being taken makes it more likely that the proximal assessment will be the best predictor of imminent recidivism and change scores will not enhance prediction.

These problems lead to the conclusion that intra-individual change should not be uniquely associated with imminent recidivism. Intra-individual change should only be relevant to the likelihood of recidivism because it can be used to calculate the most proximal level of the relevant dynamic risk and protective factors. Inter-individual differences in intra-individual change prior to the most recent assessment should be irrelevant to the prediction of recidivism. However, there are several assumptions inherent to this proposition. First, dynamic risk assessments must be completed with sufficient proximity to the outcome to capture all relevant change. Second, risk assessment tools must include all variables associated with recidivism. As we explain in the two sections that follow, neither of these assumptions is likely to be met in practice. And the implication of those assumptions not being met is that a theoretical rationale can be constructed for why intra-individual change may be uniquely associated with imminent recidivism. In the following two sections, we also explore why that relationship might exist and the possible nature of the relationship.

Dynamic risk assessments are not sufficiently proximal to the outcome.

The first issue with the argument that the most proximal assessment will provide the most accurate prediction of recidivism is that dynamic risk assessments do not occur frequently enough to obtain the level of proximity between the assessment and the outcome necessary for optimal prediction (Douglas & Skeem, 2005). Although more frequent reassessment of dynamic variables using tools such as the DRAOR is becoming more common, the frequency required to capture all changes in relevant dynamic variables is likely to be unobtainable. Dynamic variables, particularly acute variables, are hypothesised to change in a matter of weeks, days, or hours (Hanson et al., 2007; Hanson & Harris, 2000), meaning even a weekly assessment schedule will inevitably fail to capture relevant change in the measurements (e.g., a sudden job loss or romantic relationship ending). The frequency of assessment that might be able to capture all relevant changes would constitute a form of surveillance that is unachievable in practice and ethically unacceptable. Similarly, any idea of high frequency self-monitoring or self-report is unrealistic.

One consequence of insufficient proximity is that changes in dynamic risk and protective factors, particularly acute factors, are likely to occur between assessments, and between an assessment and recidivism. If change does occur in these periods, the most proximal assessment score would not represent the true value of the variables when recidivism occurs. In that case, the most proximal assessment would not provide the most accurate prediction of recidivism. Instead, a method of measurement that is able to estimate the unobserved change that will occur prior to recidivism would provide more accurate prediction of recidivism. (By definition, unobserved change will not have been measured, and thus can only be estimated through a proxy.)

What method can be used to estimate unobserved change? The most likely option would appear to be using prior, observed change. Just as past criminal behaviour is a very strong predictor of future criminal behaviour (Bonta & Andrews, 2016), past intra-individual change might be a reliable predictor of future intra-individual change. This appears to be the assumption made in the rehabilitation literature: individuals who have benefitted from treatment, as measured by a reduction in risk, are expected to continue to make similar progress after treatment (i.e., upon release back into the community). By continuing to make progress, these individuals become less likely to reoffend than individuals who did not make such positive progress during treatment. This idea—that previous change will continue following the most recent assessment—appears to be what Babchishin (2013), and Baglivio, Wolff, Piquero, et al. (2017) were suggesting in their research looking at “risk trajectories”.

Prima facie, the retrospective studies by Hanson and Harris (2000) and Zamble and Quinsey (1997) appear to provide some support for this idea of trajectories. Both of those studies found evidence that dynamic risk factors got worse in the month prior to recidivism. However, those studies provided no evidence of whether those changes were preceded by prior change in the same direction. For past, observed change to predict future, unobserved change, the essential component required is *continuity of change*. Unobserved change must continue after the most proximal assessment in the same direction and at a similar rate as prior, observed change. In these retrospective studies, because the change was self-reported, it is unclear at which point prior to the recidivism an assessment might have taken place. Therefore, it is impossible to know to what extent the change that occurred prior to recidivism would have been captured by the most proximal assessment. Only in the situation where an assessment occurred in the middle of the deterioration would prior change be uniquely related to recidivism,

because in that situation there would be prior, observed change that predicted future, unobserved change that significantly impacted on the likelihood of recidivism. These retrospective studies also lacked a comparison group and thus have no information about the changes that occurred prior to the most proximal assessment for men who did *not* reoffend. Thus, while these studies may appear to support the idea that intra-individual change will continue in the same direction after an assessment, a more detailed examination suggests they provide limited evidence to support this idea.

Possibly the strongest evidence for continuity of change was provided by Yesberg (2015), who examined whether change in dynamic risk factors during a psychological treatment programme in prison predicted change on the DRAOR during re-entry into the community. She found that treatment change was significantly associated with change during re-entry on all three DRAOR subscales. The men whose risk reduced the most during treatment were also most likely to show the biggest reduction in risk (and increase in protective factors) during re-entry. Yesberg also found some evidence that change during re-entry mediated the relationship between treatment change and recidivism, but the mediation was weak and inconsistent depending on the DRAOR subscale and outcome used. Overall, this study provides some evidence for the broader idea that there may be continuity in intra-individual change for dynamic risk and protective factors, but further research is needed to examine whether prior change in the community (as opposed to in prison) can predict future change in the community, and to test the association of that change with imminent recidivism.

Continuity of change is not the only way that prior, observed change might predict future, unobserved change. The literature reviewed in the previous section found evidence that, after controlling for proximal dynamic risk, higher prior dynamic

risk was associated with higher recidivism (Babchishin, 2013; Baglivio, Wolff, Piquero, et al., 2017; Lloyd, 2015; Mulvey et al., 2016). Recall that the general pattern observed during re-entry, including in this thesis, is one of decreasing risk and increasing protective factors. Therefore, the evidence from the previous research could be interpreted as showing that, for two individuals with the same level of proximal risk, the individual whose risk has decreased *more* would be more likely to recidivate than the individual whose risk has decreased by a lesser amount. These findings suggest prior intra-individual change is *negatively* associated with future, unobserved change. This effect would appear to be counter-intuitive, and in direct contrast to the idea of continuity of change. It is thus important to remember that most observed effects in this direction have been small, and other studies have found no association between prior risk and recidivism after controlling for proximal risk (T. H. Cohen et al., 2016; T. H. Cohen & VanBenschoten, 2014; Vose et al., 2013). However, the fact that multiple studies have found evidence of this effect is enough to warrant further consideration.

One explanation for these findings could be that risk factors may have a residual impact. Individuals who have previously been higher risk may have been exposed to factors that individuals who have previously been lower risk have not. For example, two men both deemed to have no anti-social associates would be considered to pose the same current risk at the most proximal assessment. But, if one man has previously had anti-social peers (e.g., had been a gang member), and the other man has not, that first man may pose a higher risk by virtue of the fact he would be more likely to come into contact with those anti-social peers again. Therefore, even though he has made what might be deemed a more positive change (e.g., stopped associating with a gang), and thus would have a greater reduction in risk scores, he may still be at greater risk than a similar individual who has not made such a positive change.

Another way of conceptualising this explanation is that current assessments may not cover the full range of expressions of an individual variable. For example, on a 3-point Likert scale ranging from 0-2 (e.g., the DRAOR), both men may be rated as a 0 on antisocial peers because they are not currently associating with any antisocial peers, but the residual impact of previously having peers means they should actually have different scores. In the absence of expanding the range of ratings on the assessment tool, including change alongside the current score could highlight the differences between the men. In this example, after controlling for the proximal rating, that change score would be negatively associated with recidivism because higher prior scores would be associated with a higher likelihood of recidivism.

This argument is speculative and prone to the same problems highlighted earlier in the chapter with basing hypothetical scenarios around a single risk factor. For example, if the former gang member is truly more likely to start associating with the gang again, that should be reflected in other risk and protective factors (e.g., anti-social identity or social support), again rendering the change score irrelevant. Overall, if we accept that unobserved change can occur after the most proximal assessment, it is difficult to construct a coherent explanation for why that unobserved change would be in the opposite direction or opposite magnitude from the prior, observed change. However, if evidence continues to suggest that prior intra-individual change is *negatively* associated with future, unobserved change, more extensive theoretical work will be required to establish an understanding for why that relationship exists.

The nature of the relationship between prior change and imminent recidivism is also likely to be affected by the amount of time between assessments. There are two main issues that need to be considered here. First, the amount of time over which the observed change has occurred is important. Changes over a short time period are more

likely to be temporary compared to changes over longer time periods, which could be expected to be more permanent. If these short-term changes are temporary, a negative relationship between prior change and recidivism might be observed, because the changes may reverse after the most proximal assessment. Second, the follow-up time (or the time between assessments in a design where recidivism can occur at any time) is also likely to be relevant. The longer the follow-up time, the more likely it is that unobserved change will occur, and thus that change will impact on the likelihood of recidivism.

Finally, the relationship between change and recidivism is also likely to be affected by the type of variables being measured. For acute dynamic variables, we highlighted earlier how unobserved change is likely to occur even when the time between assessments is very short. For stable (risk and protective) dynamic variables in particular, unobserved change is most likely to occur when there is a period of several weeks or months after an assessment during which change might occur. If the follow-up period is much shorter than that, it would be more likely that no unobserved change would occur, and therefore the most proximal assessment would still provide the most accurate prediction of recidivism. Empirical research comparing and contrasting these different variables could help to advance the understanding of what, for now, remains a theoretical distinction between stable and acute dynamic factors (Hanson et al., 2007; Hanson & Harris, 2000).

In summary, it seems very likely that dynamic risk assessments are not currently conducted with sufficient frequency to capture intra-individual change in all relevant variables. It is also very unlikely that assessments will be conducted with sufficient frequency in the future. Consequently, it is reasonable to propose that unobserved change in relevant dynamic variables is occurring that is impacting on the likelihood of

recidivism. Prior, observed change may give some indication of the nature of this unobserved change. If this link between prior change and future change does exist, it would provide a reason for why intra-individual change may be uniquely associated with imminent recidivism; however, the direction of that relationship remains unclear. It seems more likely that change will continue in the same direction as prior to the most proximal assessment, so there will be a positive relationship between intra-individual change and recidivism after controlling for proximal risk. But, there is some empirical evidence supporting the opposite effect (i.e., a negative relationship), a finding that is difficult to explain. Furthermore, the existence and direction of the relationship between change and recidivism is also likely to be affected by the time between assessments and time to recidivism, and the types of dynamic variable being tested (e.g., acute versus stable). Clearly, more research exploring these issues is needed.

Risk assessment tools do not include all variables associated with recidivism.

The second major problem with the argument that proximal assessments will provide the most accurate prediction of recidivism is that, currently, prediction is only moderately accurate. Existing research has found that current risk assessment tools demonstrate levels of predictive accuracy that distinguish recidivists from non-recidivists with, at best, moderate accuracy (Bonta & Andrews, 2016; Campbell et al., 2009; Hanson & Morton-Bourgon, 2009). Research suggests a combination of static and dynamic factors currently provides the most accurate prediction of recidivism (Brown et al., 2009; Jones et al., 2010), and, as statistical techniques continue to advance, particularly for the measurement of dynamic variables (Yang, Guo, Olver, Polaschek, & Wong, 2017), it is likely that further improvements in accuracy will be achieved. However, given current levels of predictive accuracy, there is a strong basis on which to

conclude that current risk assessment tools do not include all variables, whether static or dynamic, that are associated with recidivism.

In his thesis, Lloyd (2015) highlighted this issue of unobserved heterogeneity, and cited it as an assumption inherent to Cox and logistic regression that cannot be checked or tested. He proposed two solutions to this problem: 1) use effect size measures to carefully interpret results and acknowledge the amount of *unexplained* variance; and 2) conduct sensitivity analysis to estimate the impact unobserved heterogeneity might be having on the results. His sensitivity analyses showed that the impact of an unobserved variable depends heavily on the correlation between that variable and other predictors in the model, and between the unobserved variable and the outcome. The results indicated that *both* of those relationships needed to be present for the unobserved variable to substantially alter the conclusions; the stronger the relationships, the larger the impact on the conclusions. Lloyd concluded that only further research that explicitly measured these previously unobserved variables could provide more concrete conclusions.

This issue of unobserved heterogeneity is relevant to the relationship between intra-individual change and recidivism because intra-individual change may provide a measure of one or more of these currently unobserved variables. Ward and colleagues (Heffernan & Ward, 2015; Ward, 2016, 2017; Ward & Beech, 2015; Ward & Fortune, 2015) have argued that dynamic risk and protective factors should not be considered synonymous with the causes of criminal behaviour. They postulate that while some variables currently included in risk assessment measures may have a causal role in the behaviour, most dynamic risk and protective factors are not causal mechanisms, suggesting that multiple causal mechanisms remain unmeasured. It is possible that intra-individual change across multiple dynamic variables may reflect one or more of

these currently unobserved variables. For example, a single unobserved variable may cause an individual to have difficulties with interpersonal relationships, employment, and gathering social support. Therefore, a measure of intra-individual change could represent a proxy measure for that underlying variable not provided by the most proximal assessment. For measurement purposes, whether or not the unobserved variable is causal is not important. The important point is that if the unobserved variable is associated with recidivism, the intra-individual change score would enhance the prediction of imminent recidivism beyond the level obtained by the single most proximal assessment.

This rationale is positing that prior, observed intra-individual change may provide a measure of an unobserved variable that affects the likelihood of recidivism. The direction of that relationship remains uncertain. Similar to the discussion around continuity of change, it seems most likely that recidivism will be positively associated with prior change in risk scores (i.e., increases in risk will be associated with higher recidivism) and negatively associated with change in protective factor scores. However, the nature of unobserved variables—inherently unknowable—means we cannot rule out the possibility that the relationship might be negative, which would be consistent with the empirical evidence already reviewed (Babchishin, 2013; Baglivio, Wolff, Piquero, et al., 2017; Lloyd, 2015; Mulvey et al., 2016).

Summary.

This section has examined why intra-individual change in dynamic risk and protective factors might be uniquely associated with imminent recidivism. In the previous section, we suggested that the reason why previous research has found a relationship between change and recidivism, and also why previous research (including in this thesis) has found reassessment may enhance the prediction of recidivism, is that

proximity provides the most accurate estimate of dynamic variables. In this section, we advanced that argument by suggesting that under ideal conditions, the most proximal assessment of dynamic variables will provide the most accurate prediction of recidivism possible, and intra-individual change will not add incremental predictive validity. However, dynamic risk assessments do not occur under ideal conditions, as they are neither sufficiently proximal nor sufficiently comprehensive in their coverage of relevant risk and protective factors. Those two reasons provide separate but related reasons for why intra-individual change may show a significant association with recidivism after controlling for proximal assessment scores.

Conclusion.

Research and theory in the field of intra-individual change in dynamic risk and protective factors are under-developed. Intra-individual change is a central issue in correctional psychology, and there is considerable interest in the area, with sophisticated statistical techniques being developed to examine the change process (Yang et al., 2017). Currently though, little is known about simple questions such as the nature of the relationship between intra-individual change and recidivism. In particular, there is an absence of research looking at the relationship between change occurring in the community and recidivism (Serin et al., 2013).

There is growing evidence that intra-individual change is associated with recidivism, even after controlling for baseline levels of dynamic risk and protective factors. The results from earlier in this thesis have added to that body of evidence. The most likely explanation for those findings would appear to be that the combination of change in risk and baseline risk allows calculation of proximal risk. Although some researchers have drawn that conclusion (Howard & Dixon, 2013), others have implicitly

indicated that the change itself may have some inherent predictive value in addition to the proximal risk level (de Vries Robbé, de Vogel, Douglas, et al., 2015).

There is almost no existing research examining whether intra-individual change is associated with recidivism after controlling for proximal dynamic risk. Conceptually, intra-individual change in dynamic risk and protective factors would not be expected to show a unique association with recidivism because the most proximal level of those variables should be the only relevant factor for the prediction of recidivism. However, practical limitations are likely to prevent that theoretical ideal ever being met. In practice, assessments are not sufficiently proximal to the outcome, and assessment tools do not include all relevant variables. These limitations provide a plausible rationale for why research may in fact find that intra-individual change is uniquely associated with recidivism: observed change may predict future, unobserved change or prior, observed change may reflect an unobserved, underlying variable (or variables) associated with imminent recidivism.

The direction of any relationship between change and recidivism is highly uncertain. A conceptual analysis of dynamic risk and protective factors suggests higher prior risk (and lower protective) scores should be associated with a lower likelihood of recidivism, but the limited existing empirical research has found evidence of a relationship in the opposite direction. This relationship may also differ for acute and stable dynamic variables.

The Current Study

The broad aim of the analyses that follow in this chapter was to empirically test the relationship between intra-individual change and imminent recidivism, using the same dataset that has been used throughout this thesis. The analyses were divided into two steps. The first step looked at the extent to which the different intra-individual

change scores were associated with imminent recidivism. We start by describing in detail how the change scores differed for recidivists and non-recidivists. Following those descriptive statistics, we run separate predictive models for each of the different change scores. As noted in the introduction, when testing change scores, it is important to control for the pre-treatment or baseline risk. Therefore, the baseline subscale score was controlled for in this first step, with the baseline score defined in the way outlined in Chapter 3. The second step looked at the extent to which the change scores were *uniquely* associated with imminent recidivism. For this step, we reran the same models but instead of controlling for the baseline score, we controlled for the single most proximal assessment score in models alongside the different change scores.

The same statistical models used previously in this thesis—discrete-time hazard models and Cox regression with time-varying predictor models—were used for the predictive models in this chapter. The analyses therefore represent an extension of the framework and statistical approach outlined by Lloyd (2015) for testing whether reassessment of dynamic risk and protective factors enhances the prediction of imminent recidivism. The additional two steps we test in this chapter represent a method for establishing the *most accurate* way of predicting imminent recidivism using reassessments of dynamic risk and protective factors.

Defining and calculating intra-individual change.

We highlighted earlier how the relationship between intra-individual change and recidivism may be affected by the amount of time over which the change has been measured. Briefly, intra-individual change over longer time periods might be more permanent and may be more likely to predict future unobserved change than change over shorter time periods, which might be more temporary and thus, more likely to reverse. A reverse would result in a negative relationship between change and

recidivism. We also might expect the relationship between change and recidivism to differ depending on the type of dynamic variables being tested. Sudden changes in acute variables are expected to signal imminent recidivism, whereas changes in stable variables are expected to be more gradual, and less strongly associated with imminent recidivism.

For those reasons, we calculated (and tested) nine separate change scores for each DRAOR subscale, with each score measuring change across a different time period. First, we calculated a rolling, total change measure, defined as the difference between the most proximal assessment and the baseline assessment at each time point. Second, we calculated a series of eight short-term change variables, using the same time periods examined in Chapter 5 when examining aggregation over time (i.e., between 1 and 8 weeks prior to the most proximal assessment). These scores were calculated by subtracting the relevant prior score from the most proximal score. For example, the “4-week change score” was calculated by subtracting the score from 4 weeks ago from the most proximal score. Therefore, positive change scores indicated risk or protective subscale scores had increased, whereas negative change scores indicated they had decreased, and a score of 0 indicated no change. An example of how these scores were calculated is presented in Table 6.1, where we repeat the fictional example from the previous chapter, this time illustrating how change scores were calculated rather than mean scores.

In the previous chapter, we highlighted the issue of calculating a long-term average when an insufficient amount of time had passed. The same issue was relevant again here: to compare different models including change scores, all participants needed to have a minimum number of assessments to ensure each model was tested on the same sample. For example, to compare the incremental validity of 1-week change

Table 6.1

Fictional Example of a Single Individual's Scores on the DRAOR Acute Subscale Across the First 13 Weeks of the Follow-up, and the Change Scores Calculated from those Observed Scores.

Week	Observed score	1-week change	2-week change	3-week change	4-week change	5-week change	6-week change	7-week change	8-week change	Total change
0	12									0
1	12	0								0
2	11	-1	-1							-1
3	9	-2	-3	-3						-3
4	10	1	-1	-2	-2					-2
5	9	-1	0	-2	-3	-3				-3
6	9	0	-1	0	-2	-3	-3			-3
7	9	0	0	-1	0	-2	-3	-3		-3
8	8	-1	-1	-1	-2	-1	-3	-4	-4	-4
9	8	0	-1	-1	-1	-2	-1	-3	-4	-4
10	6	-2	-2	-3	-3	-3	-4	-3	-5	-6
11	6	0	-2	-2	-3	-3	-3	-4	-3	-6
12	6	0	0	-2	-2	-3	-3	-3	-4	-6
13	4	-2	-2	-2	-4	-4	-5	-5	-5	-8

Note. Blank cells represent the weeks where an insufficient number of assessments has been completed so the total change scores were used instead of the short-term change scores.

scores to 4-week change scores with the same sample group, all participants needed to have at least 5 weeks of assessments (i.e., the proximal score plus 4 prior scores). As we noted in the previous chapter, a high proportion of our sample were either censored or recidivated early in the follow-up. Therefore, we had to either exclude all participants who were censored or recidivated before a specified number of assessments had been completed, or use the total change score as the relevant score for that variable until enough assessments had been completed.

For the same reasons provided in the previous chapter, in particular the ability to retain the full sample in all analyses, we chose to use the latter approach.¹⁸ Thus, technically, the change score variables were not always consistent with their name. For example, during the first 4 weeks of the follow-up, the 4-week change scores were actually 1-, 2-, and 3-week change scores because no assessment had occurred 4 weeks prior to the most proximal assessment at that time. This issue was particularly relevant for change scores over longer time periods (e.g., 8-week change scores), where scores incongruent with the name of the variable had to be used more frequently (i.e., for 8 out of 27 weeks), meaning these variables reflect change over shorter time periods than their names might suggest.

One other important point needs to be made about change scores. There is debate in the measurement literature about the use of change scores or, as they are more commonly called, *difference* scores. A difference score will include measurement error from two measurement occasions, leading to the conclusion that “differences between scores tend to be much more unreliable than the scores themselves” (Lord,

¹⁸ Analyses were also run excluding men who were censored or recidivated before week 8 and excluding men who were censored or recidivated before week 4 (here, we tested only up to 4-week change scores). The substantive findings and conclusions did not differ from those presented below.

1956, p. 429). However, it is now accepted that difference scores are more reliable than previously thought (Trafimow, 2015). These critiques are also more commonly applied to the use of difference scores as outcome variables rather than predictor variables (Allison, 1990).

It would have been possible to examine the same research question using prior scores rather than change scores. In the introduction of this chapter, we highlighted the fact that when change scores are calculated by subtracting a baseline score from a proximal score, examining the incremental validity of change scores over baseline scores will produce the same results as examining the incremental validity of proximal scores over baseline scores. Therefore, instead of examining change scores we could have used prior scores (e.g., the baseline score instead of the total change score) and we would have obtained the same results. This equivalency is illustrated in Table 6.2, where we have presented the results for two models using the acute subscale. Model A includes the baseline scores and the proximal scores (replicating Model C in Table 4.13). Model B in Table 6.2 includes the baseline scores and the total change scores. The results show that the incremental validity results, and the parameters for the change scores and proximal scores, were identical in both models. The same equivalency applied to the short-term change variables (e.g., identical results were obtained using scores from 1 week prior as from using 1-week change scores).

We chose to use change scores instead of prior scores. We decided the concerns about the reliability of change scores did not outweigh the fact that using change scores would make for clearer and more consistent interpretation because the incremental validity of change scores over baseline scores was going to be tested first. For

Table 6.2

Results from Two Discrete Time Hazard Models Predicting Time to Recidivism Using Scores from the DRAOR Acute Subscale.

Parameters	Model A		Model B	
	<i>B (SE)</i>	OR [95% CI]	<i>B (SE)</i>	OR [95% CI]
Constant	-4.41*** (.17)	.01 [.01, .02]	-4.41*** (.17)	.01 [.01, .02]
Week	.007 (.006)	1.01 [1.00, 1.02]	.007 (.006)	1.01 [1.00, 1.02]
Baseline	-.02 (.03)	.98 [.93, 1.03]	.15*** (.02)	1.16 [1.11, 1.21]
Proximal	.17*** (.03)	1.18 [1.13, 1.24]		
Total Change			.17*** (.03)	1.18 [1.13, 1.24]
R ² (X0)		.13		.13
c-index		.60		.60
Deviance		4147.50		4147.50
AIC		4155.50		4155.50
BIC		4186.38		4186.38
Δ χ^2 from baseline-only model (df)		43.35*** (1)		43.35*** (1)

Note. **Bolded values** indicate statistically significant results.

* $p < .05$, ** $p < .01$, $p < .001$

completeness, we tested all models using prior scores in place of change scores. When these analyses were run, for all models, confidence intervals for the odds ratios of the proximal scores overlapped, and overall effect sizes for the models were identical regardless of which variable was used.

Results

Was intra-individual change associated with imminent recidivism?

How did change scores differ for recidivists and non-recidivists?

We first examined the total change scores for men classified as either recidivists ($n = 441$)¹⁹ or non-recidivists ($n = 506$). On average, stable and acute subscale scores decreased, and protective subscale scores increased, between the baseline assessment and the most proximal assessment for both recidivists (Stable: $M = -.11$, $SD = 1.67$; Acute: $M = -.58$, $SD = 1.98$; Protective: $M = .18$, $SD = 1.74$) and non-recidivists (Stable: $M = -.71$, $SD = 2.31$; Acute: $M = -1.18$, $SD = 2.58$; Protective: $M = .78$, $SD = 2.15$). The amount of change on all three subscales was significantly lower for recidivists than non-recidivists (all $p < .001$) but effect sizes indicated the magnitude of the differences was small ($d = .26-.31$). The amount of change observed on the acute subscale was substantially higher than for the other two subscales.

These analyses give an indication of the overall trend in scores for recidivists and non-recidivists, but they fail to account for the time over which these changes occurred. For example, non-recidivists had an average gap of 21 weeks between the first and last assessments in their sequences, compared to 9 weeks for recidivists. Therefore, the

¹⁹ The 19 men who recidivated in week 1 and thus had only 1 assessment in their sequence were excluded from these descriptive analyses as they could not demonstrate change. However, these men were retained in the sample for the predictive model analyses to ensure the models could be compared with the models presented in previous chapters.

greater decrease in risk and increase in protective subscale scores observed for non-recidivists may be partly explained by the differences in the time over which they were assessed.

To better account for the differences in the timing of assessments, we examined the average 1-week change scores for recidivists and non-recidivists. These scores were calculated by dividing the total change score for each individual by the number of weeks over which that change occurred and then using those individual averages to calculate a group average. On all three subscales, these scores indicated very little change occurred from week to week for either eventual recidivists (Stable: $M = .01$, $SD = .20$; Acute: $M = -.05$, $SD = .31$; Protective: $M = .01$, $SD = .22$) or non-recidivists (Stable: $M = -.02$, $SD = .23$; Acute: $M = -.05$, $SD = .29$; Protective: $M = .02$, $SD = .15$). For the acute ($p = .90$) and protective subscales ($p = .19$), these change scores were not significantly different for recidivists and non-recidivists; both groups had, on average, a slight decline in risk scores and a slight increase in protective scores across consecutive assessments. For the stable subscale, non-recidivists had a slight decrease whereas recidivists had a very slight increase.²⁰ This difference was significant ($p = .03$) but effect sizes showed the magnitude of the difference was very small ($d = .14$).

For recidivists, it was possible to look at the change observed immediately prior to recidivism over the eight shorter time periods (e.g., between 1 and 8 weeks). These results are presented in Table 6.3. For the stable subscale, a very small *increase* in

²⁰ This result may seem inconsistent with the result in the previous paragraph showing that the average stable subscale total change for recidivists was negative. The discrepancy resulted from the fact that individuals with negative total change scores recidivated later in the follow-up than individuals with comparable positive total change scores, hence the slightly negative average *total* change and the slightly positive average *weekly* change.

Table 6.3

Mean Change Scores for the DRAOR Subscales Over Different Time Periods Immediately Prior to Recidivism.

Weeks Prior to Recidivism	Stable	Acute	Protective
	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>
1	.00 (.32)	.03 (.75)	.02 (.42)
2	.02 (.67)	.10 (1.12)	.02 (.69)
3	.05 (.89)	.10 (1.30)	.00 (.77)
4	.05 (.99)	.03 (1.34)	.01 (.88)
5	.04 (1.05)	.04 (1.42)	.01 (.97)
6	.04 (1.15)	-.01 (1.48)	-.02 (1.11)
7	.05 (1.26)	-.09 (1.56)	-.02 (1.28)
8	.05 (1.33)	-.10 (1.67)	-.03 (1.32)
Total change	-.11 (1.67)	-.58 (1.98)	.18 (1.74)

Note. $n = 441$ men for all values. For men who recidivated in the first 8 weeks of the follow-up, the change from baseline was used in the calculation of the average value. See Table 5.1 for further information.

scores was observed prior to recidivism across all time periods. For the protective subscale, there was no meaningful change in scores over short time periods, but a slight indication that scores decreased over longer periods prior to recidivism (e.g., 6 to 8 weeks prior). For the acute subscale, over short periods (e.g., 2 or 3 weeks prior to recidivism) there was evidence of an increase in scores, whereas over longer periods (e.g., 7 or 8 weeks prior) the more expected decrease in scores was observed.

These results suggest that patterns of change may be different in the period immediately prior to recidivism than at other points, particularly for the acute subscale. In this respect, the results support the idea that sudden changes in acute variables may signal imminent recidivism (Hanson et al., 2007; Hanson & Harris, 2000). However, it is important to view these group-level scores in perspective. For the acute subscale, across the two assessments immediately prior to recidivism (i.e., 1-week change scores), acute subscale scores increased for 53 recidivists (12.0%), decreased for 46 recidivists (10.4%), and did not change for 342 recidivists (77.6%). Even across the 3 weeks prior to recidivism—the period with the highest average increase in Table 6.3—the proportion of recidivists with an increase (22.9%) and a decrease (21.5%) was almost identical. Similar results were observed for the stable and protective subscales (i.e., near identical proportions of recidivists had an increase and a decrease in scores immediately prior to recidivism), except with fewer recidivists showing a change in scores in either direction immediately prior to recidivism.

These results provide an indication of what was happening for recidivists immediately prior to recidivism, but they are limited by the absence of a non-recidivist comparison group. Change occurring prior to the final assessment in the sequence for non-recidivists does not have the same meaning as change occurring up to the final assessment before recidivism because, as the total change scores showed, the timing of

the assessment is a confounding factor. Therefore, without a comparison group, it is impossible to know whether the patterns of change highlighted in Table 6.3 represent a meaningful difference between recidivists and non-recidivists.

One way of addressing that limitation is to examine how frequently different changes preceded recidivism for the full sample. Here, we focus only on change observed between any two consecutive assessments, and categorise changes into increases, decreases, or no change. For the acute subscale, recidivism followed immediately after 4.3% of increases in acute scores (53 recidivism events out of 1178 increases), whereas recidivism immediately followed 2.5% of decreases and 2.7% of weeks where there was no change from the previous week. For the stable subscale, there was a similar discrepancy between increases (2.3%) and decreases (1.2%) but recidivism was most likely to follow a week of no change (3.4%). For the protective subscales, 2.0% of increases, and 2.6% of decreases, and 2.8% of weeks with no change immediately preceded recidivism. These results show that recidivism was more likely when there was an increase in risk or a decrease in protective factors as compared to a change in the opposite direction on those subscales. For the stable and acute subscales, recidivism was nearly twice as likely to occur following an increase than a decrease. Even for these two subscales though, recidivism was very unlikely following an increase in risk, and for the stable subscale, recidivism was more likely to follow a week of no change.

Those percentages increase substantially if we account for less immediate recidivism. We looked at how frequently recidivism occurred within 6 weeks of a change in each of the subscales. For the acute subscale, recidivism followed within 6 weeks of 19.5% of increases and 14.6% of decreases. For the stable and protective subscales, results were almost identical to each other: recidivism followed within 6

weeks of 18.1% of increases, and only 9.8% of decreases on the stable subscale; on the protective subscale, 18.9% of decreases preceded recidivism within 6 weeks, and 9.4% of increases. For all three subscales, a week of no change preceded recidivism in the following 6 weeks 14% of the time. These results show that an increase in stable risk scores or a decrease in protective scores was nearly twice as likely to signal recidivism within 6 weeks as a decrease in stable or an increase in protective scores, and that recidivism was more likely to follow an increase in risk than a week of no change.

In summary, our results showed that group-level comparisons of recidivists and non-recidivists indicated recidivists were likely to demonstrate less of a reduction in risk than non-recidivists over the course of the follow-up. However, when the time over which that change occurred was accounted for, the results indicated that recidivists and non-recidivists had gradually decreasing risk scores and gradually increasing protective factor scores at similar rates. Contrary to these general trends, in the periods immediately prior to recidivism, there was some evidence of increasing risk and decreasing protective factor scores. These differences were very small though, and further analysis suggested that recidivists were as likely to have had a decrease in their risk scores immediately prior to recidivism as they were to have had an increase. Across the full follow-up, there was evidence that increases in risk (and decreases in protective factors) were more likely to precede recidivism, especially when recidivism within 6 weeks of any assessment was considered. Again though, the overall rates of recidivism following any particular pattern of change were very low.

Did change scores predict imminent recidivism?

To examine the significance of the differences between recidivists and non-recidivists just described, we tested a series of discrete-time hazard models. For each of the three DRAOR subscales, we ran nine discrete-time hazard models, one for each

different time period over which change was measured: 1-8 weeks prior to most recent assessment, and rolling, total change from baseline. In each model, we also controlled for the baseline score.²¹ Cox regression with time-varying predictor models were also run to calculate the overall effect sizes for the models, using the same effect size statistics from previous chapters: Xu and O'Quigley's (1999) R^2 measure, and Heagerty and Zheng's (2005) time-dependent AUC or *c*-index.

The results from these models are presented in Table 6.4. Note that each row in the table presents the results for three different models: one for each of the stable, acute, and protective subscales. Only the overall model effect sizes are included in the first row of results because only the baseline score was included as a predictor in that model. In each of the remaining rows, models included both the baseline score and the change score so model effect sizes *and* regression coefficients for the individual change scores are included (regression coefficients for the baseline scores are not shown).

The results show that, with limited exceptions, change scores on all three subscales, across all time periods, were significantly associated with recidivism after controlling for baseline scores. The only variables that were not significant predictors were the 1-week change scores on the stable subscale, and the 1- and 2-week change scores on the protective subscale. Thus, only the very short-term change scores on the two non-acute subscales were not significantly related to recidivism.

²¹ Univariate models where the change scores were the only predictor in the model were also tested. These models examined whether change was significant regardless of baseline. The pattern of results for the univariate models was identical to the pattern of results for the models that controlled for baseline risk (i.e., if change scores were significant after controlling for baseline, without exception they were also significant when we did not control for baseline), so we have chosen not to present those results.

Table 6.4

Parameter Estimates for Discrete-Time Hazard Models, and Effect Sizes for Cox Regression with Time-Varying Predictor Models Using Scores from a Baseline Assessment and Change Scores Across Different Time Periods on the DRAOR Subscales.

Predictor variables	Stable			Acute			Protective		
	<i>B (SE)</i>	<i>R</i> ² (XO)	<i>c</i> -index	<i>B (SE)</i>	<i>R</i> ² (XO)	<i>c</i> -index	<i>B (SE)</i>	<i>R</i> ² (XO)	<i>c</i> -index
Baseline only		.07	.57		.03	.55		.06	.56
Baseline + change									
1-week	.14 (.10)	.07	.57	.15* (.06)	.04	.56	-.09 (.10)	.07	.56
2-week	.17* (.07)	.07	.58	.20*** (.04)	.12	.58	-.12 (.07)	.07	.56
3-week	.22*** (.06)	.08	.58	.21*** (.04)	.13	.59	-.17** (.06)	.07	.57
4-week	.21*** (.05)	.09	.59	.17*** (.04)	.13	.58	-.15** (.05)	.07	.57
5-week	.19*** (.05)	.09	.59	.19*** (.03)	.13	.59	-.16** (.05)	.08	.57
6-week	.19*** (.04)	.11	.59	.18*** (.03)	.13	.59	-.19*** (.04)	.09	.58
7-week	.20*** (.04)	.12	.60	.16*** (.03)	.12	.59	-.19*** (.04)	.10	.58
8-week	.19*** (.04)	.12	.60	.17*** (.03)	.16	.59	-.20*** (.04)	.10	.59
Total change	.14*** (.03)	.13	.60	.17*** (.03)	.13	.60	-.15*** (.03)	.12	.59

Note. Each row represents a separate model, and within each row there are three models represented, one for each DRAOR subscale. **Bolded values** indicate statistically significant results.

* $p < .05$, ** $p < .01$, *** $p < .001$

The regression coefficients showed that, across all time periods, after controlling for baseline score, higher change scores on the risk subscales and lower change scores on the protective subscale were associated with a higher likelihood of recidivism. These relationships need to be carefully interpreted. *Higher* change does not mean *greater* change. Rather, when comparing two individuals, a higher change score suggests either a *more positive* or a *less negative* change score. Therefore, for the risk subscales, it was true to say that the positive regression coefficient indicated that recidivism became more likely the more scores increased. However, it was also true, and probably more consistent with the descriptive statistics, to say that recidivism became more likely the *less* scores *decreased*.

The overall model effect size measures in Table 6.4 suggest that the impact of these change scores on predictive accuracy was small. For the stable and protective subscales, change scores over longer time periods enhanced prediction more than change over short time periods. This finding was consistent with the idea that change scores are relevant to the extent that they allow calculation of proximal risk; change over longer time periods is more likely to give an indication of proximal risk after controlling for baseline. Note however, that even the total change scores added only 6% of explained variance, and the time-dependent AUC measures only increased by .03 on each subscale.

For the acute subscale, the improvements in the time-dependent AUC measure were similar to the other two subscales, with very small gradual increases in predictive accuracy as change was measured over longer time periods. The percentage of variance explained was more consistent across the different change scores than we observed for the other two subscales. Change on the acute subscale substantially increased the proportion of variance explained when compared to the baseline-only model.

Therefore, these results provided mixed evidence that change in the acute subscale was only relevant because it allowed calculation of proximal risk.

Was intra-individual change *uniquely* associated with imminent recidivism?

Although most of the observed effects were weak, the results in the previous section indicated that most change scores were significantly associated with imminent recidivism. Therefore, the first step in our analyses provided sufficient evidence to justify continuing to the second step: testing the extent to which change scores were associated with imminent recidivism after controlling for the single most proximal assessment score, as opposed to the baseline assessment score. To test this question, we again ran nine discrete-time hazard models (and Cox regression with time-varying predictor models) for each of the three DRAOR subscales and calculated overall effect sizes for each of the models.

The results for these models are summarised in Table 6.5. The format used for Table 6.4 is repeated here, where each row in the table presents the results for three different models—one for each of the stable, acute, and protective subscales—and, other than the first row, each model includes both the proximal score and the change score. The results in Table 6.5 show that very few change score variables were significantly associated with imminent recidivism after controlling for proximal scores. For the stable subscale, none of the change scores demonstrated significant incremental predictive validity. For the protective subscale, only the 8-week change scores were significant, with regression coefficients indicating higher change scores were associated with lower recidivism.

There was stronger evidence of incremental validity of the change scores for the acute subscale. The 2-, 3-, 5-, and 6-week change scores were all significant predictors of

imminent recidivism after controlling for proximal score. The positive regression coefficients for all four time periods indicated that higher change scores (i.e., more positive or less negative) were associated with a higher likelihood of imminent recidivism. Thus, among individuals with the same proximal acute subscale score, recidivism was more likely among those with lower prior acute scores. Again, this finding should not be interpreted as evidence that *increases* in acute scores signalled imminent recidivism; more likely, it suggests that the amount that acute scores for recidivists decreased over those time periods was significantly less than the amount by which non-recidivists' scores decreased. Across all three subscales, effect size measures indicated the addition of change scores had little to no impact on predictive accuracy compared to prediction using only the proximal score. For the stable and protective subscales, there was no evidence of a change in effect sizes compared to the effect sizes obtained using only the proximal score. For the acute subscale, there was a small increase in the amount of variance explained, as indicated by Xu and O'Quigley's (1999) R^2 measure, but no change in the time-dependent AUC scores, even for those variables that had demonstrated significant incremental predictive validity.

To further examine the incremental validity of the change scores, we modelled the time dependent AUC values across the 26-week follow-up, using the same approach that was used in Chapter 4. By bootstrapping 95% confidence intervals on the time-dependent AUC values, these graphs can be used to highlight any points in the follow-up where two models significantly differ. We tested this approach for all five change score variables that demonstrated significant incremental predictive validity in the discrete-time hazard models (four from the acute subscale and one from the protective

Table 6.5

Parameter Estimates for Discrete-Time Hazard Models, and Effect Sizes for Cox Regression with Time-Varying Predictor Models Using Scores from the Single Most Proximal Assessment and Change Scores Across Different Time Periods on the DRAOR Subscales.

Predictor variables	Stable			Acute			Protective		
	<i>B (SE)</i>	<i>R</i> ² (XO)	<i>c</i> -index	<i>B (SE)</i>	<i>R</i> ² (XO)	<i>c</i> -index	<i>B (SE)</i>	<i>R</i> ² (XO)	<i>c</i> -index
Proximal only		.12	.60		.11	.60		.12	.59
Proximal + change									
1-week	.01 (.10)	.12	.60	.04 (.06)	.11	.60	.04 (.10)	.12	.59
2-week	.05 (.07)	.13	.60	.10* (.04)	.17	.61	-.003 (.07)	.12	.59
3-week	.09 (.06)	.12	.60	.10** (.04)	.17	.61	-.05 (.06)	.12	.59
4-week	.08 (.05)	.12	.60	.06 (.04)	.16	.61	-.03 (.05)	.12	.59
5-week	.06 (.05)	.12	.60	.08* (.03)	.16	.61	-.04 (.05)	.12	.59
6-week	.07 (.05)	.13	.60	.07* (.03)	.15	.61	-.07 (.05)	.12	.59
7-week	.07 (.04)	.13	.60	.04 (.03)	.14	.61	-.07 (.04)	.12	.59
8-week	.07 (.04)	.13	.60	.05 (.03)	.16	.61	-.09* (.04)	.12	.59
Total change	.001 (.03)	.13	.60	.02 (.03)	.13	.60	-.03 (.03)	.12	.59

Note. Each row represents a separate model, and within each row there are three models represented, one for each DRAOR subscale. **Bolded values** indicate statistically significant results.

* $p < .05$, ** $p < .01$, *** $p < .001$

Figure 6.1. Time-dependent Area Under the Curve (AUC) statistics for each week of the 26-week follow-up for the single most proximal assessment score and for a model including the proximal score and 3-week change scores on the DRAOR acute subscale. The lower graph replicates the upper graph, but with bootstrapped 95% confidence intervals overlaid.

subscale). The time-dependent AUC values obtained using only the proximal score were compared to the time-dependent AUC values obtained by models where the change scores were entered alongside the proximal score.

The results showed that, for all five variables, the confidence intervals overlapped throughout the follow-up. At no point in the follow-up did any of the change scores significantly improve predictive accuracy. As an illustration of that pattern, in Figure 4.8 we present the results for the model where the 3-week change scores on the acute subscale were tested. The results in Table 6.4 had indicated this model had the largest difference in effect size from the reassessment-only model, and thus, this model was most likely to have significant differences during individual weeks of the follow-up.

In Figure 4.8, the results for that model are compared to the AUC values obtained from the model that include only the most recent assessment. The graph shows how similar the AUC values were throughout the follow-up and the large overlap in the confidence intervals for the two models at all points.

Discussion

Summary of results.

In this chapter, the aim was to introduce and test a method for examining the extent to which intra-individual change in dynamic risk and protective factors is uniquely associated with imminent recidivism. The method we have proposed can also be viewed as an additional step in Lloyd's (2015) framework for testing the extent to which reassessment of dynamic risk and protective factors enhances the prediction of imminent recidivism. Previous results in this thesis have found support for the idea that reassessment does significantly enhance prediction. That finding can alternatively be interpreted as evidence that change between the baseline and most proximal assessment was significantly associated with imminent recidivism. This chapter set out

to examine the extent to which those total change scores, plus a series of change scores over shorter time periods, were associated with imminent recidivism, and the extent to which those change scores could enhance the predictive accuracy obtained using the most proximal assessment.

The descriptive statistics highlighted few clear differences between the individuals who would later be reconvicted and those who would not. Overall, DRAOR risk subscale scores decreased over time, and protective subscale scores increased. Although the recidivists showed significantly less total change, comparisons of the average weekly change indicated that scores for both groups changed in a similar way from week to week. There were some differences in the patterns of change over the full follow-up compared to the patterns observed immediately prior to recidivism for the recidivist group. Most notably, the acute subscale scores tended to decrease over the full follow-up period but increase immediately prior to recidivism. Further analyses indicated that these differences were unlikely to be practically meaningful, with recidivism approximately equally likely to follow an increase or decrease in subscale scores; these analyses were also limited by the absence of a comparison group.

When we analysed the predictive validity of these change scores, we found consistent evidence that change leading up to the most proximal assessment was significantly associated with imminent recidivism. Change scores on all three DRAOR subscales were significant predictors of imminent recidivism after controlling for the baseline assessment score. The only exception to this finding was that change on the stable and protective subscales over 1 or 2 weeks prior to the most proximal assessment was not significantly associated with imminent recidivism. We also found consistent evidence that the addition of these change scores enhanced predictive accuracy beyond the level obtained using the baseline score. As expected, after

controlling for baseline scores, change scores on the stable and acute subscales were positively associated with recidivism, while change scores on the protective subscale were negatively associated.

In contrast, we found limited support for the idea that intra-individual change is *uniquely* associated with imminent recidivism. The majority of the change scores we tested did not demonstrate significant incremental predictive validity over the most proximal score and had no impact on predictive accuracy compared to the level of accuracy achieved using only the most proximal score. A small subset of change scores were significant predictors of imminent recidivism after controlling for the most proximal score. Most of these variables were short-term (i.e., between 2 and 6 weeks) change scores on the acute subscale, with change over 8 weeks on the protective subscale also demonstrating significant incremental predictive validity. The relationship between change and recidivism was positive for the acute subscale, and negative for the protective subscale. The protective subscale change score did not appear to enhance predictive accuracy; there was some evidence that the acute change scores had a small impact, but this finding was not consistent across the two effect size measures that we examined.

Comparison with previous research.

The results in this chapter were consistent with Lloyd's (2015) finding that prior or change scores did not enhance the prediction of imminent recidivism compared to only the most proximal score, but in many other respects, our results directly contrasted with Lloyd's findings. We found no evidence of intra-individual change from baseline demonstrating significant incremental predictive validity; Lloyd found baseline scores on all three DRAOR subscales were significant predictors after controlling for the most proximal score. We did find some evidence of significant incremental predictive

validity for change over shorter time periods prior to recidivism, particularly for the acute subscale, which Lloyd did not investigate. Interestingly, we found a positive association between these intra-individual change scores and imminent recidivism for the risk subscales, and a negative association between change and recidivism for the protective subscale. These findings indicated that, after controlling for the most proximal score, lower prior risk and higher prior protective scores were associated with a higher likelihood of recidivism. In contrast, Lloyd found that higher stable and acute subscale scores and lower protective scores at baseline were associated with a higher likelihood of recidivism.

Our results were also inconsistent with the direction of the unique relationship between change and recidivism found in other previous studies. Babchishin (2013), Baglivio, Wolff, Piquero, et al. (2017), and Mulvey et al. (2016) all presented results suggesting that, similar to Lloyd (2015), after controlling for proximal score, individuals with higher prior risk scores and lower prior protective scores had a higher likelihood of recidivism. As mentioned previously, these findings are difficult to explain, as they were in the opposite direction to what would be expected if change occurred continuously following the most proximal assessment. The direction of the relationship that we observed between prior change and imminent recidivism was in a more expected direction. Our results suggest that, among individuals with the same proximal score, those with lower prior risk and higher prior protective subscale scores had a higher likelihood of recidivism than individuals with lower prior risk and higher prior protective factor scores. Framed in terms of change, these findings suggest that the individuals with the highest likelihood of recidivism were those whose risk scores had either increased by a greater amount or, more likely, had decreased by a lesser amount.

This contrast in results is difficult to explain. As detailed in the introduction of this chapter, it is challenging to construct a coherent explanation for the findings of the previous studies, whereas, among other explanations, our results are consistent with the idea of continuity of change (i.e., the idea that change occurring prior to the most proximal assessment would continue in the same direction and magnitude after that assessment). Differences in the risk assessment tool and statistical approach used may explain some of the difference between our results and those of Babchishin (2013), Baglivio, Wolff, Piquero, et al. (2017), and Mulvey et al. (2016), but those factors cannot be used to explain the differences in the results observed by Lloyd (2015) since his study used the same assessment tool and statistical approach. Thus, other factors must be responsible for the differences in results. While some discussion of potential explanations for these differences is warranted, more weight should be placed on the broader finding that intra-individual change did not significantly enhance predictive accuracy. In this respect, our results were consistent with T. H. Cohen et al. (2016), T. H. Cohen and VanBenschoten (2014), Lloyd (2015), and Vose et al. (2013), who all found evidence that change did not enhance predictive accuracy over proximal risk.

Theoretical implications and practical applications.

The results in this chapter indicate that the best explanation for the finding that intra-individual change in dynamic risk and protective factors is associated with imminent recidivism is because change allows calculation of the proximal level of those variables. Change scores were significant predictors after controlling for baseline scores. This finding indicated that among individuals with the same baseline scores, change scores further distinguished who was most likely to be imminently reconvicted. Change scores did not appear to similarly distinguish the individuals most likely to be imminently reconvicted when they were tested alongside the most proximal scores. In

combination, these findings suggest that what matters with respect to the measurement of dynamic risk and protective factors is the finishing point, not the starting point.

In the introduction of this chapter, we suggested two potential mechanisms through which intra-individual change might be uniquely associated with imminent recidivism. The results have provided some support for a trajectory or continuity of change theory of dynamic risk. Contrary to previous research, the relationship we observed between imminent recidivism and the change in dynamic risk was positive; higher change scores, indicating a larger increase or smaller decrease in risk (and a larger decrease or smaller increase in protective factors), were associated with a higher likelihood of recidivism. These findings were consistent with the idea that prior, observed change in dynamic risk and protective factors will continue in the same direction and magnitude after the most proximal assessment. However, the more consistent finding was that change scores did not have an impact on predictive accuracy in addition to the proximal score; that finding suggests limited weight should be placed on the direction of the relationship observed, and further research examining the idea of continuity of change is needed.

The second hypothesised rationale for why intra-individual change might be uniquely associated with imminent recidivism—the idea that change scores might provide a measure of an unobserved variable that is associated with imminent recidivism—is less concerned with the direction of the relationship and more focused on whether change enhances prediction. Therefore, the finding that change scores did not enhance the prediction of imminent recidivism suggests the change score measures that we tested did not provide a measure of unobserved risk-related variables. We should note though, that the low levels of predictive accuracy we observed clearly support the idea that many factors related to the likelihood of recidivism were not

measured by the DRAOR subscales. Our findings *do not* imply that there were *no* unobserved variables; rather, the implication is that the measures of intra-individual change that we tested do not represent a proxy for these unobserved variables.

The results in this chapter provided further evidence of the conceptual distinction between stable and acute dynamic variables that has been highlighted in previous chapters in this thesis. The descriptive statistics found some evidence of a trend of acute subscale scores increasing in the period immediately prior to recidivism, a finding that is consistent with the original conceptualisation of acute risk factors (Hanson et al., 2007; Hanson & Harris, 2000). There was also some evidence of stable subscale scores increasing and protective subscale scores decreasing immediately prior to recidivism, but the trend was not as clear as for the acute subscale, suggesting a distinction between these types of variables. A much clearer difference was observed in the amount of change, with the acute subscale showing more change from week-to-week and across the full follow-up than the other two subscales. The distinction between stable and acute was also evident in the models testing incremental validity after controlling for baseline. Change scores on the stable and protective subscales became more strongly predictive of imminent recidivism when measured over longer time periods, whereas the predictive accuracy of the acute change scores was more consistent across time periods. When the incremental validity of change scores over proximal scores was tested, the acute subscale demonstrated the clearest evidence of a unique association with intra-individual change whereas little to no evidence of a similar relationship was observed for either the stable or protective subscales. Again though, the fact that change scores had such a limited effect on predictive accuracy means these discrepancies should be interpreted with caution.

From a practical perspective, these results provide further evidence for the conclusion drawn from previous results. Specifically, the results suggest that, in a sequence of reassessments of dynamic risk and protective factors using the same assessment tool, the single most proximal assessment is the best indicator of the likelihood of imminent recidivism for a particular individual. Decisions about supervision and intervention should be made based on the most proximal assessment score, alongside any static risk score that is available. The results also suggest practitioners should not be strongly influenced by patterns of change leading up to that most recent assessment. Although there was some indication that changes in acute variables, especially a sudden increase in risk, might signal imminent recidivism, the evidence more strongly suggested that current risk is the most relevant consideration. This evidence would suggest that individuals assessed as being at the same current risk should be supervised in similar way, regardless of their baseline or other prior dynamic risk level.

Limitations and future research.

It is possible that the method we have proposed for testing the extent to which intra-individual change is uniquely associated with recidivism reduced the likelihood of finding a significant relationship or an enhancement in predictive accuracy. Our definition of imminent recidivism as occurring within 6 weeks of an assessment may have been too short for enough unobserved change to occur that would significantly impact on the likelihood of recidivism. In other words, although we argued that reassessments of the DRAOR do not occur frequently enough to obtain perfect prediction, in this case the frequency may have been high enough to prevent a finding that change enhanced prediction. Furthermore, the method we used to structure the dataset and the nature of the data used, meant that the time between the final

assessment and recidivism was even shorter than 6 weeks in most cases, further reducing the amount of time available for unobserved change to occur. A design with a longer time period between assessments, and thus between assessments and recidivism, may have found a stronger association between intra-individual change and recidivism.

The use of an exclusively high static risk community sample remains a limitation of these results, as it was for the other results in this thesis. The restricted range of static risk scores likely reduced the levels of predictive accuracy observed, and also likely reduced the range of change scores, as most of the men started at a similar point. It should be noted though, that the spread of DRAOR scores in our sample was similar to previous DRAOR research (Lloyd, 2015) that did not restrict the sample based on static risk. Further research is needed using a sample with a wider range of static risk scores in order to draw broader conclusions about the relationship between intra-individual change and imminent recidivism.

Similarly, further research is needed in other correctional settings to examine the relationship between intra-individual change and recidivism. The majority of the existing intra-individual change literature has involved investigation of psychological treatment or other intervention programmes taking place in institutions, but little of that research has examined whether treatment change is associated with recidivism after controlling for post-treatment dynamic risk. Similar principles should apply to the dynamic variables measured in that setting, and we might hypothesise that a stronger relationship would be found between intra-individual change and recidivism in that context because of the longer time period between the most proximal assessment and the measurement of recidivism.

Conclusion.

We hope that this chapter provides a framework for the investigation of intra-individual change across a range of different samples and settings. The results observed using the current sample have not provided strong evidence that intra-individual change is uniquely associated with imminent recidivism; in terms of reassessment, the findings suggest the most proximal assessment of dynamic risk and protective factors provides the most accurate estimate of likelihood of recidivism, with some suggestion that acute dynamic risk factors might be different. However, before any substantive conclusions can be drawn, there is a clear need for replication of our findings and further research in this area to resolve the conflicting findings that have been observed in the existing research. These conclusions, some of which we have tentatively outlined in this section, could have substantial implications for both theory and practice in this area. We discuss these implications in more detail in the final chapter of this thesis.

First though, in the next chapter, we use the framework outlined in this chapter to examine the relationship between imminent recidivism and a series of alternative measures of intra-individual change. Change scores (or difference scores) are not the only way of capturing change in an individual's dynamic risk and protective factors over time. Alternative measures such as a mean score or a measure of variability can also provide information about how that individual has changed over time. To what extent those measures are associated with imminent recidivism, and more importantly, as this chapter has argued, to what extent those measures are associated with imminent recidivism after controlling for the single most proximal assessment score, is the focus of the next chapter.

Chapter 7: Alternative Measures of Intra-Individual Change

In the previous chapter, we outlined and tested a revised version of Lloyd's (2015) framework for examining the relationship between intra-individual change in dynamic risk and protective factors and recidivism. In that chapter, intra-individual change referred primarily to change scores, or the difference in score between two specified assessments. Change scores, however, are not the only way of measuring how individuals change over time. A sequence of assessments will have several other statistical properties that provide information about how an individual has changed over time. Those statistical properties become relevant to the extent that they are associated with recidivism. In fact, applying the framework from the previous chapter, the question is not just to what extent those variables are associated with recidivism, but to what extent the variables are associated with recidivism after controlling for the most proximal assessment score.

In this chapter, we explore the relationship between imminent recidivism and two alternative types of intra-individual change variables: mean scores and variability scores. The framework presented in the previous chapter is used to structure the discussion in this chapter. We examine the rationale for why these types of variables might be associated with imminent recidivism and why they might be *uniquely* associated with imminent recidivism. We then test these ideas empirically to examine the extent to which mean scores and variability scores were associated with imminent recidivism in our dataset. These analyses are highly exploratory in nature, as previous research exploring these variables is almost non-existent, particularly using the method we are proposing. Our aim is to highlight some of the additional ways in which a series of reassessment scores can be used to predict recidivism and identify areas that merit further investigation.

Variability Scores

When more than two assessments of the same individual are completed, there will be additional information available that is not incorporated into a change score. For example, two individuals may both have change scores of 0, indicating no change in the dynamic risk and protective factors over the specified time period. However, additional assessments completed between the two assessments used to calculate the change score may suggest the individuals have changed in different ways. One individual's scores may not have changed at all during that period, whereas the other individual's scores may have fluctuated before returning to their original value. These different patterns of change may reflect two individuals with a different likelihood of recidivism despite having identical change scores.

One method for addressing this limitation is to test change scores over different time periods, as we did in the previous chapter. By testing change scores across different time periods, it is more likely that the two individuals described in the previous paragraph will be distinguished on the basis of their change scores (e.g., over a shorter time period, differences in change score will be observed). The problem with this approach is that it relies on finding the time period over which change is most reliably associated with recidivism, which is likely to be difficult. The nature of the relationship between changes in dynamic risk and protective factors and recidivism remains highly uncertain, and there are likely to be considerable individual differences in the time period over which change is most reliably associated with recidivism. That relationship may also differ depending on the nature of the variables (e.g., stable vs. acute dynamic variables). Therefore, testing change scores over different time periods may not be sufficient to address the limitation of relying on two data points.

One alternative measure of intra-individual change that may address this limitation is variability. There is almost no existing research examining variability in a sequence of assessments of dynamic risk and protective factors, which is not surprising given how few studies have used the multi-wave assessment design that would make it possible to examine variability. In the two studies that have explicitly considered it (Richards, 2017; Yesberg, 2015), both of which also used DRAOR scores, variability has been defined in a couple of different ways: (a) the spread of scores within an individual sequence, and (b) the frequency with which scores change, in either direction. In contrast with change scores, both of these measures account for all assessments during the relevant time period, not just the first and last assessments in a sequence. These measures are also distinct from change scores in that they are concerned with the fact that change is occurring rather than the direction of that change.

In the study by Yesberg (2015), the standard deviation of all assessment scores that occurred within 100 days of release was used as the measure of variability. An adapted DRAOR subscale comprising a subset of the acute risk factors was found to demonstrate significantly greater variability than the remaining acute items or the stable and protective subscales. However, no significant relationships were found between the variability scores on any of the subscales and either violent recidivism or reimprisonment outcomes (general recidivism was not tested). Richards (2017) also looked at the standard deviation of all assessment scores that occurred within 100 days of release as a measure of variability. In addition, she calculated a “cumulative change” variable, which was a count of the number of times scores had changed across any two consecutive assessments. Similar to Yesberg, Richards found significantly greater variability in the acute subscale than the stable or protective subscale, but she did not test the relationship between these variability measures and recidivism.

Lloyd (2015) also looked at a measure of variability of DRAOR scores. He calculated the average *absolute* amount of change for recidivists and non-recidivists on consecutive assessments. This variable represented a measure of variability because it was focused on whether change was occurring, not the direction in which the change was occurring. He found that recidivists had greater variability than non-recidivists, despite the fact that average change for non-recidivists had been higher than for recidivists. This finding was interpreted as evidence that non-recidivists' scores were more likely to change in the same direction (i.e., declining risk and increasing protective factor scores), whereas recidivists' scores were more likely to fluctuate in both directions. These variability measures were not included in any of the regression models in his study, so it remains to be seen whether this measure of variability is significantly associated with recidivism.

These previous studies provide limited evidence of the nature of the relationship between variability and recidivism, with only Lloyd (2015) finding any evidence to support the idea that variability is associated with recidivism. Lloyd's results suggested variability might be *positively* associated with recidivism, and the other authors have also argued that greater variability will be associated with a higher likelihood of recidivism. However, none of these studies have laid out a clear rationale for why that relationship may exist. The general idea would appear to be that variability in dynamic risk and protective factor scores is reflective of some level of individual stability, both in internal characteristics and external circumstances. Reintegrating into the community after release from prison is difficult for most individuals, and subsequently many individuals released from prison have very unstable lives (Western, Braga, Davis, & Sirois, 2015). Therefore, the likely rationale for why variability may be associated with

recidivism seems to be clear: individuals released from prison have unstable lives and unstable lives make it more likely that these individuals will reoffend.

More careful examination of this idea suggests the rationale may not be so clear. In the context of re-entry into the community after release from prison, instability generally refers to lives marked by factors such homelessness, unemployment, substance use, and an absence of positive social support (Western et al., 2015). From a measurement perspective, this instability is likely to be reflected in higher risk and lower protective factor scores. In contrast, variability, defined as risk factors increasing *and* decreasing (as opposed to simply moving in one direction), would more likely manifest as individuals moving in and out of employment, or intermittent periods of homelessness between periods spent in a stable living situation. This type of pattern may be observed for some individuals released from prison, and these individuals might be quite likely to recidivate. However, for variability to be statistically associated with recidivism, individuals with higher variability would need to be more likely to recidivate than individuals with low variability. For that to happen, individuals who bounce between being high and low risk would need to be more likely to recidivate than both individuals who have been consistently low risk *and* individuals who have been consistently high risk over the same time period. It is difficult to see why being occasionally high-risk would be associated with a higher likelihood of recidivism than being consistently low risk. This pattern might be possible for some individuals, but on a group level, it seems more likely that variability, in isolation, will be unrelated to the prediction of recidivism.

Constructing a rationale for why variability might be associated with imminent recidivism after controlling for proximal risk is equally difficult. Using the same arguments discussed in the previous chapter, we could postulate that prior, observed

variability will predict future unobserved variability, or that variability may be caused by an unobserved, underlying variable that is associated with recidivism. The problem is that even if prior variability predicts future variability, by definition, variability suggests change in either direction (increasing or decreasing risk). Thus, individuals with higher variability prior to the most proximal assessment may be equally likely to have a reduction in risk as an increase, in which case variability would not be a significant predictor of recidivism. One exception might be where the variation is not equally likely to result in movement in either direction (i.e., scores are more likely to increase). In that situation though, it is more likely that the amount of change (i.e., the change score) will predict the future, unobserved change rather than the variability.

In summary, the idea that variability in intra-individual dynamic risk and protective factor scores might be associated with recidivism is intuitively appealing, but further examination suggests that, in practice, a significant relationship is unlikely to be found. Instability is likely to be associated with recidivism, but it is unclear to what extent measures of statistical variability will represent an operationalisation of instability. Furthermore, the limited existing research that has tested variability has found little to no evidence of a relationship between variability and recidivism. However, only one study to date (Yesberg, 2015), using only a single measure of variability (standard deviation), has explicitly examined the relationship between intra-individual variability and recidivism. There would be value in replicating this finding using multiple measures of variability and testing the extent to which those measures are associated with imminent recidivism, including after controlling for proximal risk.

Mean Scores

Mean scores were the focus of the analyses in Chapter 5. In that chapter, we questioned whether aggregating scores from multiple assessments would result in

more accurate estimation of variables due to a reduction in measurement error, leading to more accurate prediction of imminent recidivism. To test that idea, several mean score variables calculated by aggregating a different number of assessments were examined. Each of the mean scores was a significant predictor of imminent recidivism; however, the results indicated that aggregation did not enhance prediction when separate models were compared to each other: the single most proximal assessment was a more accurate predictor of imminent recidivism than the mean scores. The most likely explanation for that finding was that reassessments were capturing true change in addition to measurement error, so aggregation did not result in more accurate estimation of the variables.

In this chapter, the question is to what extent mean scores might enhance the prediction of imminent recidivism in addition to the most proximal score. Although the question is different, a similar conceptual argument can be made for why the mean scores might demonstrate incremental validity. A single assessment provides an unreliable estimate of a set of variables due to measurement error. Change scores rely on a single, prior assessment as the comparison point for the most proximal assessment. Compared to a single, prior assessment, a mean score incorporates multiple assessments and therefore, may be a more reliable estimate of prior risk. In the previous chapter, the dynamic nature of the variables was presented as the most likely explanation for why aggregation did not enhance prediction. The fact the true value of the variables can change is not as relevant here, where the aim is not to estimate the level of the variables at a single point in time. Rather, the aim is to find the best estimate of prior risk that is most likely to predict future unobserved risk. In other words, in this context, prior risk is not conceptualised as a stable construct. Instead it is accepted that some of the observed change will reflect true change. The mean score will be able to

capture that change to provide a measure of, on average, an individual's risk and protective factors over the relevant time period.

The argument for why mean scores might be uniquely related to imminent recidivism is therefore very similar to the rationale proposed in the previous chapter for why change scores might be uniquely associated with recidivism. The proposition is that two individuals with the same proximal score might be able to be distinguished based on their mean score over a particular time period. Applying the same idea of continuity of change, we might posit that an individual who has been consistently higher risk (as reflected in a higher mean risk score and lower mean protective score) might be expected to be less likely to recidivate than an individual who has been consistently lower risk, because, to have identical proximal scores, the higher risk individual must have made more positive change than the second individual. In other words, prior, observed change will positively predict future unobserved change. The suggested difference between mean scores and change scores is that the mean scores will provide a more reliable measure of change than change scores because they account for multiple prior assessments and are thus likely to be less affected by measurement error.

Of the two studies cited in Chapter 5 that investigated the predictive validity of mean scores (Hanson et al., 2007; Lloyd, 2015), only Lloyd looked at whether those scores demonstrated incremental validity over the proximal scores. He found that the rolling, total mean score demonstrated significant incremental validity for the acute and protective DRAOR subscales, but not the stable subscale. However, even when the mean scores were significant predictors alongside the proximal scores, there was no evidence that they enhanced predictive accuracy. These mixed findings suggest it is possible that mean scores may demonstrate significant incremental validity but may not increase

predictive accuracy. Between Lloyd's findings and the fact that mean scores (across different time periods) have already been shown to be significantly associated with recidivism in our dataset, there is enough evidence to justify further exploratory analysis of whether mean scores can enhance the prediction of imminent recidivism.

The Current Study

The analyses that follow in this chapter examine the extent to which the alternative measures of intra-individual change that have just been discussed—measures of variability and mean scores—were associated with imminent recidivism. The same dataset used throughout this thesis is used for these analyses. For the measures of variability, the two steps proposed in the previous chapter were completed: (1) testing to what extent variability, in isolation, was associated with imminent recidivism, and (2) testing to what extent the measures were uniquely associated with imminent recidivism (i.e., after controlling for the most proximal assessment score). For the mean scores, the first step has already been completed in Chapter 5, where each of the scores was found to be significantly associated with imminent recidivism, so the focus in this chapter was solely on testing incremental validity over the proximal scores. The same statistical models used previously in this thesis—discrete-time hazard models and Cox regression with time-varying predictor models—are used again for these analyses.

Defining and calculating intra-individual change.

Measures of variability.

Drawing on the previous work by Lloyd (2015), Richards (2017), and Yesberg (2015), we chose to test three different measures of variability. The three measures were standard deviation, frequency of change, and cumulative total change. The aim of these variables was to capture the extent to which any change in scores, whether an

increase or a decrease, was occurring, as opposed to change in a particular direction. Because these analyses were exploratory, and to allow for comparison with the change scores and mean scores, all variability measures were calculated across the same nine time periods: a total, rolling score, and between 1 and 8 weeks prior to the most proximal assessment.

In the earlier research (Richards, 2017; Yesberg, 2015), the standard deviation measure was calculated using all assessment scores during the first 100 days after release. But, in those studies, men who were reconvicted during the first 100 days were removed from the sample. In our study, it was important to include the entire sample, so we calculated a standard deviation score for every time point, using all existing scores in an individual's sequence up to that point.

The frequency of change and total cumulative change variables were very similar to each other. The frequency of change variable was the same variable that Richards (2017) had termed "cumulative change". In our study, frequency of change was calculated by summing the number of assessments, including the current assessment, in which observed scores on a subscale had changed (in either direction) from the previous assessment. Similarly, the cumulative total change variable was calculated by summing the absolute amount of change made on the subscale from the previous assessment. Therefore, these two variables would be identical unless a change of greater than 1 point occurred across consecutive assessments.

To highlight how these scores were calculated, in Table 5.1, we present the same fictional example that was presented in the previous two chapters. Unlike the previous two chapters, this time the example includes only the rolling, total scores for the different types of variables; the scores for the shorter time periods are excluded. The

Table 7.1

Fictional Example of a Single Individual's Scores on the DRAOR Acute Subscale Across the First 13 Weeks of the Follow-up, and the Rolling Total Mean, Change Score, and Variability Scores Calculated from those Observed Scores.

Week	Observed score	Rolling mean	Total change	Standard deviation	Change frequency	Cumulative change
0	12	12.0	0	0.0	0	0
1	12	12.0	0	0.0	0	0
2	11	11.7	-1	0.6	1	1
3	9	11.0	-3	1.4	2	3
4	10	10.8	-2	1.3	3	4
5	9	10.5	-3	1.4	4	5
6	9	10.3	-3	1.4	4	5
7	9	10.1	-3	1.4	4	5
8	8	9.9	-4	1.5	5	6
9	8	9.7	-4	1.5	5	6
10	6	9.4	-6	1.8	6	8
11	6	9.1	-6	2.0	6	8
12	6	8.8	-6	2.1	6	8
13	4	8.5	-8	2.4	7	10

table also includes the mean scores and change scores to allow comparison with the three variability measures. In particular, the table highlights the similarities between the change scores and frequency of change and cumulative total change, and illustrates how differences in these variables were most likely to emerge in cases when scores both increased and decreased within the same sequence.

Mean scores.

The mean scores tested in this chapter were the same variables tested in Chapter 5, where we provided detailed information on how the scores were calculated (an illustration of how the rolling, total mean score was calculated is reproduced in Table 5.1). One additional point is worth noting here. Since the “2-week mean scores” are the average of the most proximal assessment and the assessment from the previous week, they are most comparable to the “1-week change scores” because those scores were also calculated using only two assessments. Hence, there is technically no equivalent mean score for the “8-week change scores”, which are calculated using a total of nine assessments. To maintain consistency between chapters, we chose not to include an additional “9-week mean” score in the analyses in this chapter.

Results

Measures of variability.

How did variability scores differ for recidivists and non-recidivists?

For the three variability measures we were testing, one method of describing differences between recidivists and non-recidivists was to look at the overall amount of variability in each individual’s sequence. When this method was used, the three variability measures we calculated indicated non-recidivists had a much greater spread of scores, frequency of changes in scores, and total amount of change on all three subscales, compared to recidivists. However, as with the change scores, comparison

across the full follow-up period is misleading. On average, non-recidivists had more than twice the number of assessments over which to demonstrate change, which also meant they were likely to have a greater spread in their scores.

A fairer comparison was obtained by looking at the mean absolute change observed across consecutive weeks. These scores were calculated by averaging the average absolute weekly change score for each individual. On a weekly basis, recidivists (Stable: $M = .12$, $SD = .22$; Acute: $M = .30$, $SD = .32$; Protective: $M = .10$, $SD = .15$) and non-recidivists (Stable: $M = .11$, $SD = .21$; Acute: $M = .36$, $SD = .40$; Protective: $M = .11$, $SD = .23$) demonstrated very similar fluctuation in scores, with nearly three times as much fluctuation observed for the acute subscale as for the other two subscales. Similar to the analyses in the previous chapter, we also looked at how frequently change (in either direction) or no change immediately preceded recidivism. For the acute subscale, recidivism was slightly more likely to follow weeks of change (3.2% of the time) than weeks of no change (2.7%), but the pattern was reversed for the protective subscale (2.2% vs. 2.8%). On the stable subscale, recidivism was more than twice as likely to immediately follow a week of no change (3.4%) than a week of change (1.6%).

Were variability scores associated with imminent recidivism?

To examine whether the differences just highlighted were meaningful, we tested the extent to which the three variability measures predicted imminent recidivism after controlling for baseline scores. Nine discrete-time hazard models for each different measure on each DRAOR subscale were run. The regression coefficients for each of the variability measures in each model are presented in Table 7.2. These results show that there was inconsistent evidence of an association between variability and imminent recidivism.

Table 7.2

Regression Coefficients of Variability Measures in Discrete-Time Hazard Models Predicting Imminent Recidivism After Controlling for Baseline DRAOR Score.

Predictor variables	Stable			Acute			Protective		
	<i>B (SE)</i>			<i>B (SE)</i>			<i>B (SE)</i>		
	Standard Deviation	Change frequency	Cumulative change	Standard Deviation	Change frequency	Cumulative change	Standard Deviation	Change frequency	Cumulative change
1-week	-.42* (.21)	-.57* (.27)	-.30* (.15)	.04 (.09)	.15 (.12)	.03 (.06)	-.18 (.17)	-.29 (.24)	-.13 (.12)
2-week	-.12 (.13)	-.17 (.16)	-.07 (.08)	.12 (.08)	.13 (.08)	.09* (.04)	-.06 (.13)	-.20 (.16)	-.04 (.08)
3-week	-.06 (.12)	-.15 (.13)	-.04 (.06)	.16* (.08)	.14* (.06)	.08* (.03)	-.12 (.12)	-.21 (.13)	-.06 (.06)
4-week	-.10 (.11)	-.21 (.11)	-.04 (.05)	.11 (.08)	.10 (.05)	.06* (.03)	-.14 (.12)	-.21 (.12)	-.05 (.05)
5-week	-.11 (.11)	-.14 (.10)	-.05 (.05)	.10 (.08)	.11* (.05)	.06* (.03)	-.16 (.11)	-.19 (.07)	-.05 (.05)
6-week	-.07 (.10)	-.11 (.09)	-.02 (.04)	.08 (.07)	.11** (.04)	.06* (.02)	-.15 (.10)	-.19* (.09)	-.05 (.05)
7-week	-.07 (.09)	-.11 (.08)	-.02 (.04)	.08 (.07)	.11** (.04)	.06** (.02)	-.09 (.10)	-.13 (.08)	-.02 (.04)
8-week	-.06 (.09)	-.09 (.08)	-.02 (.04)	.08 (.07)	.11** (.04)	.06** (.02)	-.08 (.09)	-.12 (.08)	-.02 (.04)
Rolling total	-.14 (.08)	-.09 (.06)	-.04 (.03)	-.05 (.08)	.05* (.03)	.03* (.02)	-.13 (.08)	-.07 (.06)	-.02 (.03)

Note. Each individual cell represents a separate model. **Bolded values** indicate statistically significant results. Change frequency = number of weeks any change in score has occurred; Cumulative change = absolute amount of change in scores that has occurred

* $p < .05$, ** $p < .01$, *** $p < .001$

For the stable subscale, all three measures of variability were significant predictors when measured using only the most proximal score and the score from the previous week. There was no evidence of a significant association across any other time period. For the protective subscale, there was also inconsistent evidence of a significant association with imminent recidivism. Only the change frequency variable, when calculated using all assessments from the 6 weeks prior to the most proximal assessment, was a significant predictor. On both these subscales, the variability variables that were significant predictors were negatively associated with recidivism, indicating individuals with greater variability were less likely to be reconvicted.

There was more consistent evidence of a link between variability and recidivism for the acute subscale. Only the standard deviation across the 3 weeks prior to recidivism was a significant, but the change frequency and cumulative change variables across nearly all time periods were weak but significant predictors of imminent recidivism. The relationship between acute variability and recidivism was in the opposite direction than for the other two subscales: a positive regression coefficient indicated greater variability was associated with a higher likelihood of recidivism.

Were variability scores uniquely associated with imminent recidivism?

The previous section showed some evidence of the variability measures demonstrating incremental predictive validity over the baseline scores, particularly for the acute subscale. To examine whether the variability measures would demonstrate incremental validity over proximal scores, we examined the association between variability and imminent recidivism after controlling for the single most proximal score. The results for these models are summarised in Table 7.3. For the stable and protective subscales, none of the three measures of variability demonstrated significant

Table 7.3

Regression Coefficients of Variability Measures in Discrete-Time Hazard Models Predicting Imminent Recidivism After Controlling for Proximal DRAOR Score.

Predictor variables	Stable				Acute			Protective	
	<i>B (SE)</i>				<i>B (SE)</i>			<i>B (SE)</i>	
	Standard Deviation	Change frequency	Cumulative change	SD	Change frequency	Cumulative change	SD	Change frequency	Cumulative change
1-week	-.37 (.21)	-.49 (.27)	-.26 (.15)	.00 (.09)	.12 (.12)	.003 (.06)	-.14 (.17)	-.22 (.24)	-.10 (.12)
2-week	-.06 (.13)	-.10 (.16)	-.03 (.08)	.09 (.08)	.11 (.08)	.07 (.04)	-.02 (.13)	-.13 (.16)	-.02 (.07)
3-week	-.0004 (.12)	-.07 (.13)	-.01 (.06)	.14 (.08)	.12* (.06)	.06* (.03)	-.07 (.12)	-.14 (.13)	-.03 (.06)
4-week	-.03 (.11)	-.13 (.12)	-.01 (.05)	.10 (.08)	.09 (.05)	.05 (.03)	-.09 (.11)	-.14 (.12)	-.02 (.05)
5-week	-.03 (.11)	-.07 (.10)	-.01 (.05)	.10 (.07)	.10* (.05)	.05* (.03)	-.10 (.11)	-.12 (.10)	-.03 (.05)
6-week	.003 (.10)	-.04 (.09)	.01 (.10)	.10 (.07)	.10* (.04)	.05* (.02)	-.09 (.10)	-.12 (.09)	-.02 (.04)
7-week	.01 (.09)	-.04 (.09)	.01 (.04)	.11 (.07)	.10** (.04)	.05* (.02)	-.02 (.09)	-.06 (.08)	.01 (.04)
8-week	.02 (.09)	-.02 (.08)	.01 (.04)	.12 (.07)	.10** (.03)	.05** (.02)	-.01 (.09)	-.05 (.08)	.01 (.04)
Rolling total	-.02 (.08)	-.02 (.06)	.001 (.03)	.08 (.07)	.05 (.03)	.03* (.01)	-.01 (.08)	-.003 (.06)	.01 (.03)

Note. Each individual cell represents a separate model. **Bolded values** indicate statistically significant results. Change frequency = number of weeks any change in score has occurred; Cumulative change = absolute amount of change in scores that has occurred

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 7.4

Effect Sizes for Cox Regression with Time-Varying Predictor Models Using Scores from the Single Most Proximal Assessment and Measures of Variability Across Different Time Periods on the DRAOR Acute Subscale.

Predictor variables	Standard Deviation		Change frequency		Cumulative change	
	R^2 (XO)	c -index	R^2 (XO)	c -index	R^2 (XO)	c -index
Proximal only	.11	.60	.11	.60	.11	.60
Proximal + variability						
1-week	0.12	0.60	0.12	0.60	0.12	0.60
2-week	0.15	0.60	0.14	0.60	0.17	0.60
3-week	0.16	0.60	0.15	0.61	0.16	0.61
4-week	0.14	0.60	0.14	0.60	0.16	0.60
5-week	0.13	0.60	0.15	0.61	0.16	0.61
6-week	0.14	0.60	0.16	0.61	0.17	0.61
7-week	0.13	0.60	0.16	0.61	0.15	0.61
8-week	0.12	0.60	0.16	0.61	0.15	0.61
Rolling total	0.11	0.60	0.13	0.61	0.15	0.61

Note. Each row presents effect sizes for separate models, and within each row effect sizes are presented for the three different models for each of the three measures of variability.

incremental validity over the most proximal score, regardless of the time period over which the measures were calculated. For the acute subscale, both the change frequency and cumulative total change scores showed some evidence of incremental validity, particularly when assessments from between 3 and 8 weeks prior to the most proximal assessment were included in the variability measures.

Overall effect sizes for these models are presented in Table 7.4. These results were very similar to the pattern observed for the change scores: when the variability measure was a significant predictor, there was a small increase in the amount of variance explained but no noticeable increase in the time-dependent AUC value. Subsequently, when the time dependent AUC values and bootstrapped 95% confidence intervals across time were calculated, there was no evidence of a significant difference at any point in the follow-up between the models that included the variability measures and the model that only included the single most proximal score.

Finally, to check it was variability as opposed to the direction of the change that explained these results, we tested whether the variability measures were collinear with the change scores calculated over the same time period. Collinearity diagnostics indicated there were no issues with multicollinearity (i.e., all tolerance values were above 0.2). We also examined whether the variability measures demonstrated incremental validity over both the proximal score and the comparable change score. Results (not shown) indicated that the variability measures demonstrated consistent levels of incremental validity when the change scores were also included in the models. Thus, the results indicated that it was the variability as opposed to the direction of the change that explained why these variables were significantly associated with imminent recidivism after controlling for the most proximal score. For these models where both the variability and the change scores were included, model effect sizes indicated the

combination of the two variables further increased the amount of variance explained (up to as high as 25%) but had no impact on the time-dependent AUC values, which remained no higher than 0.61.

Mean scores.

As noted, the association between mean scores and imminent recidivism on all three DRAOR subscales was established in Chapter 5. Therefore, what remained to be tested was the extent to which the mean scores were significant predictors of imminent recidivism after controlling for the most proximal score.

In Table 7.5, we present a summary of the results from eight models for each of the DRAOR subscales testing the unique association between the mean scores and imminent recidivism. The table shows that mean scores for both the stable and the protective subscale did not demonstrate significant incremental predictive validity. Consistent with this finding, the effect size measures for the models including mean scores showed no noticeable difference from the models than included only the proximal scores. In contrast, evidence of incremental predictive validity was found for the acute subscale. All mean scores between 4 and 8 weeks were found to significantly predict recidivism after controlling for proximal scores. The regression coefficients were all negative, indicating that higher mean acute subscale scores were associated with lower recidivism. Effect sizes indicated that the addition of the mean scores did not increase the time-dependent AUC but did have a noticeable impact on the amount of variance explained. All four models where the mean scores were significant incremental predictors explained 18% of the variance in the predictors, compared with 11% explained by models with only the proximal score.

For each of the four models where mean scores demonstrated incremental predictive validity, we also calculated the time dependent AUC values throughout the

Table 7.5

Parameter Estimates for Discrete-Time Hazard Models, and Effect Sizes for Cox Regression with Time-Varying Predictor Models Using Scores from the Single Most Proximal Assessment and Mean Scores Across Different Time Periods on the DRAOR Subscales.

Predictor variables	Stable			Acute			Protective		
	<i>B</i> (<i>SE</i>)	<i>R</i> ² (XO)	<i>c</i> -index	<i>B</i> (<i>SE</i>)	<i>R</i> ² (XO)	<i>c</i> -index	<i>B</i> (<i>SE</i>)	<i>R</i> ² (XO)	<i>c</i> -index
Proximal only		.12	.60		.11	.60		.12	.59
Proximal + mean									
2-week	-.01 (.20)	.12	.60	-.08 (.12)	.13	.61	-.07 (.20)	.10	.58
3-week	-.07 (.14)	.12	.60	-.13 (.08)	.16	.61	-.01 (.14)	.10	.58
4-week	-.12 (.11)	.12	.60	-.15* (.07)	.18	.61	.04 (.11)	.11	.58
5-week	-.13 (.10)	.12	.60	-.13* (.06)	.18	.61	.05 (.10)	.10	.58
6-week	-.12 (.09)	.12	.60	-.13* (.06)	.18	.61	.05 (.09)	.10	.59
7-week	-.11 (.09)	.12	.60	-.12* (.06)	.18	.61	.07 (.08)	.10	.59
8-week	-.10 (.08)	.12	.60	-.11* (.05)	.18	.61	.08 (.08)	.10	.59
Total mean	-.08 (.06)	.13	.60	-.08 (.05)	.17	.60	.10 (.06)	.10	.59

Note. Each row represents a separate model, and within each row there are three models represented, one for each DRAOR subscale.

Bolded values indicate statistically significant results.

* $p < .05$, ** $p < .01$, *** $p < .001$

follow-up and the bootstrapped 95% confidence intervals. When these results were graphed (not shown), they showed that, as with the change score models, at every point of the follow-up, the confidence intervals of the models that included mean scores overlapped substantially with the model using only the single most proximal score.

One notable feature of these results was the similarity with the results from the previous chapter when we tested the incremental validity of the change scores over the most proximal score. As with the change scores, incremental validity was only evident for the acute subscale, and on that subscale, the individual variables that demonstrated incremental predictive validity were almost identical (i.e., change scores from 2, 3, 5, and 6 weeks prior to the most proximal were significant; mean scores across the 3, 4, 5, 6, and 7 weeks prior to the most proximal were significant). The most likely explanation for this consistency was that the mean scores and the prior scores used to calculate the change scores were collinear. To test this hypothesis, we calculated Pearson's correlation coefficients and tolerance values for the prior score and the corresponding mean score (e.g., the 1-week prior score was tested with the 2-week average score because these variables covered the same time period). All correlation coefficients were above 0.8 (most were well above 0.9) and all tolerance values below 0.2, indicating the presence of a very high degree of multicollinearity for these two variables. Thus, it would appear that the finding here is predominantly a replication of the findings in the previous chapter regarding the incremental validity of change scores over the proximal score.

Discussion

This chapter presented an exploration of two new ways of conceptualising intra-individual change. We argued that, in multi-wave designs, raw change scores may not capture all information about intra-individual change that is relevant to the prediction

of imminent recidivism. We suggested that measures of variability and aggregation of prior scores into mean scores may represent two methods for capturing that additional information in a way that further enhances the prediction of recidivism. However, measures of variability are not consistent with the concept of stability, so we questioned whether variability would show any evidence of an association with recidivism. For the mean scores, it was suggested that aggregation may provide a more reliable estimate of prior risk by which to measure change, and thus the mean scores may enhance prediction more than was observed for the raw change scores. Neither of these areas has been explored using this approach in previous research, so the analyses in this chapter were highly exploratory.

The results for the measures of variability provided mixed support for our hypothesis. For the stable and protective DRAOR subscales there was no evidence of an association between variability and recidivism, including after controlling for the most proximal score. For the acute subscale, both the change frequency and cumulative change variables showed some evidence of an association with imminent recidivism. The results suggested that individuals with more frequent changes in their acute subscale scores were more likely to be reconvicted, even after controlling for proximal risk. This effect remained present even when the corresponding change score was also controlled for, suggesting it was the fluctuation in scores rather than the direction of that fluctuation that was associated with recidivism. Effect size measures indicated that, although these variables demonstrated statistically significant incremental predictive validity, their impact on predictive accuracy was limited.

The finding that some measures of variability were associated with imminent recidivism was unexpected. We had hypothesised that variability would be unrelated to recidivism, save to the extent that variability was collinear with change scores. The

finding that fluctuation on the acute subscale was significantly associated with imminent recidivism, including after controlling for proximal risk and change scores, was inconsistent with this hypothesis. In this dataset, fluctuation of the acute variables was uniquely associated with recidivism. Lloyd (2015) made a similar finding in his study, showing that absolute weekly change in the acute subscale was higher for recidivists than non-recidivists. However, Lloyd did not include that measure in his predictive models, so the results cannot be compared in that respect.

This unexpected finding is difficult to explain. The continuity of change hypothesis would suggest that changes following the most proximal assessment should be equally as likely to be increases as decreases, and thus unrelated to recidivism. A better alternative explanation may be that fluctuation in acute scores is caused by an unobserved, underlying variable that is positively associated with imminent recidivism. We could speculate that variability reflects a variable that is not otherwise captured by the DRAOR, and that variable makes it more likely that individuals will be reconvicted of a new offence. For example, the supervising probation officer is one unobserved variable that could explain these results. Frequent changes in the supervising probation officer would mean frequent changes in the DRAOR assessor, which hypothetically could lead to fluctuation in scores due to subjective differences in item ratings. The changes in probation officer could also affect the relationship between the probation officer and the parolee, which previous research suggest can influence the likelihood of recidivism (Kennealy, Skeem, Manchak, & Eno Loudon, 2012; Polaschek, 2016). Further research is needed to establish whether there is any validity to this suggestion.

The mean scores also demonstrated some evidence of incremental predictive validity. Again, the evidence of incremental validity was only present for the acute subscale; there was no evidence that mean scores on either the stable or protective

subscales were significant predictors of imminent recidivism after controlling for the most proximal score. For the acute subscale, higher mean scores were associated with lower recidivism, which can be interpreted as evidence that those who showed less positive change (e.g., a greater increase or a smaller decrease in scores) were more likely to be reconvicted. Consistent with the variability measures, effect sizes indicated that the variables that demonstrated statistically significant incremental predictive validity had a relatively limited impact on predictive accuracy. Overall, the findings were consistent with Lloyd (2015), who also found some evidence of incremental validity but no evidence that mean scores enhanced predictive accuracy.

Further analysis indicated that the mean scores demonstrated multicollinearity with the prior scores that they were intended to be compared against. This finding explained why the results were essentially identical to the results observed for the change scores in the previous chapter. This issue with multicollinearity suggests the results for the mean scores have few theoretical or practical implications. In the introduction to this chapter, we hypothesised that aggregation may increase the reliability of the estimates of prior values of dynamic risk and protective factors. That hypothesis rests on sufficient change occurring over that period for the mean score to be statistically distinguishable from the prior score; otherwise, there will be no difference in results obtaining using mean scores or prior scores. In our dataset, it appears that sufficient change did not occur, and hence, results did not differ from those observed in the previous chapter. We still believe there is merit in exploring this idea further but only with data that meet these requirements (e.g., sufficient variation in scores such that mean scores are statistically distinguishable from prior scores).

The aim of this chapter was to explore some additional measures of intra-individual change that might address limitations of raw change scores. This work has

represented an effort to further develop the framework proposed by Lloyd (2015) for testing those ideas. The results we have observed in this chapter provide some areas for further research to explore. In particular, we found some evidence that variability in acute dynamic variables might be uniquely associated with imminent recidivism. Overall though, these results do not provide strong enough evidence to support the suggestion that testing these alternative measures of intra-individual change should be incorporated into the revised framework. When examining the predictive accuracy of a series of reassessments of dynamic risk and protective factors, the most proximal assessment should remain the starting point, before then looking at the extent to which the change occurring up to that point enhances prediction. Alternative measures of intra-individual change such as variability or mean scores *may* also enhance prediction, but further research is needed to establish whether that is the case.

Chapter 8: General Discussion

This thesis has centred on the concepts of dynamic risk and protective factors. These concepts essentially refer to changeable variables that are associated with the likelihood of future criminal behaviour. These types of variables are already widely used in criminal justice systems around the world, in particular to guide intervention and management of individuals who are being released from prison. However, these variables remain poorly understood from a theoretical perspective, and their use is supported by a limited base of empirical research, particularly in relation to how they change over time and how change is associated with recidivism. In this thesis, we set out to address some of those gaps in the existing theoretical and empirical literature.

The starting point for this thesis was the framework developed by Lloyd (2015) for testing the extent to which reassessment of dynamic risk and protective factors enhances the prediction of imminent recidivism. The first aim of our thesis was to replicate the results observed by Lloyd when he tested his framework with a large sample of individuals released from prison on parole in New Zealand. In this thesis, we used an almost identical dataset except drawn from a later point in time and including only individuals considered high-risk of being reconvicted and returning to prison based on a static risk measure. The same measure of dynamic risk and protective factors—the DRAOR—was used in both studies; importantly, parolees were frequently *reassessed* on the measure, providing the ability the measure change over time. Overall, the replication was largely successful: results in our thesis were consistent with Lloyd's results, although with slightly smaller effect sizes, likely due to the high-risk sample we used.

In this thesis, we have developed and tested a slightly *revised* framework from the one developed by Lloyd (2015). To recap, the steps in our revised framework are as follows:

- 1) Proximal Score vs. Baseline Score: Compare the predictive accuracy of the single most proximal assessment score in a sequence of repeated assessments to the predictive accuracy of the first assessment score from that sequence.
- 2) Proximal Score vs. Mean Scores: Compare the predictive accuracy of the single most proximal assessment score in a sequence of repeated assessments to the predictive accuracy of an aggregate measure of any combination of prior scores (e.g., all scores, or only a few of the most proximal) from that sequence.
- 3) Proximal Score vs. Proximal Score + Change Scores: Compare the predictive accuracy of the single most proximal assessment score in a sequence of repeated assessments to the predictive accuracy of both the single most proximal score and change scores calculated by subtracting any prior score (e.g., the first assessment or the second-most proximal) from the most proximal score.

The first step remains unchanged from Lloyd's (2015) framework. This step provides a measure of the practical value of reassessment and a test of the essential criteria that a variable must satisfy to be validly classified as a dynamic predictor of recidivism. When we tested this step in Chapter 4, we found evidence that reassessment enhanced the prediction of imminent recidivism. There was consistent evidence, across all three DRAOR subscales, that the single most proximal assessment score was a more accurate predictor of imminent recidivism than the baseline assessment score, although

the observed effect was weaker than in Lloyd's sample. Thus, by successfully replicating Lloyd's results, we provided further evidence that the DRAOR subscales meet the criteria expected of dynamic predictors of recidivism, in particular the criterion that intra-individual change in those subscales must be associated with recidivism.

The second step in the revised framework is an integration of the second and third steps in Lloyd's (2015) framework. We chose to integrate these two steps because, as we argued in Chapter 5, the theoretical rationale for why aggregation of dynamic variables should enhance the prediction of imminent recidivism is unconvincing. In contrast, we believe that aggregation should either fail to improve prediction (when limited change in the variables has occurred) or should worsen prediction (when sufficient change has occurred to make aggregate measures less accurate estimates of the variables). Therefore, as a general rule, all aggregate measures can be subsumed under a single heading and into a single step of our revised framework. When this step was tested using our dataset, results were consistent with the idea that proximity is the most important factor: aggregation of dynamic risk and protective factors did not appear to improve the prediction of imminent recidivism. Again, this finding was consistent across the three DRAOR subscales but the evidence that aggregation *decreased* predictive accuracy was strongest for the acute subscale.

The third step represents the major revision that we have made to Lloyd's (2015) framework. In Chapter 6, we argued that the best theoretical explanation for the association between intra-individual change and imminent recidivism is that change scores provide an indication of current or proximal risk. Thus, when controlling for proximal risk, intra-individual change is unlikely to be associated with imminent recidivism. However, we noted several limitations that provide a rationale for why intra-individual change might be *uniquely* associated with recidivism. When we tested

this step, the results indicated that, with the possible exception of some short-term change scores on the acute subscale, intra-individual change was not uniquely associated with imminent recidivism. These results supported the idea that the relationship between intra-individual change in dynamic variables and recidivism is most likely to be a result of the fact that change scores allow calculation of proximal risk.

In Chapter 7, we explored several alternative measures of intra-individual change—measures of variability and mean scores—but we have not included those measures in the revised framework. We argued that raw change scores are limited by the fact that they do not account for all available information in a sequence of reassessments. In contrast, both mean scores and measures of variability can be calculated using all available assessments over the period of interest and thus might be more reliable measures of change. However, the rationale for why these variables might (a) provide distinct results from raw change scores, or, (b) be uniquely associated with imminent recidivism, is not strong. When we tested these variables, we found that mean scores were largely indistinguishable from change scores, but there was some evidence that measures of variability were uniquely associated with imminent recidivism. Variability continued to demonstrate incremental validity when models also included change scores, suggesting combining the amount and the direction of change may be an interesting avenue for future research. For now, these measures remain exploratory and we do not believe they warrant inclusion in the revised framework until further research is conducted.

In summary, our revised framework draws heavily on Lloyd's (2015) framework, but one step has been altered and one substantial new step has been added. Also, rather than considering the framework as a set of guidelines for testing the extent

to which reassessment enhances prediction, we believe that the revised framework is better thought of as a test of the association between intra-individual change in dynamic variables and recidivism. Reassessment is still an essential component—it is impossible to measure change without at least one reassessment—but we think change is the concept that unifies the full framework. The framework provides a template for testing to what extent change matters and why it matters. The other simplification in the revised framework is that, at each step, models using the proximal score should provide the most accurate prediction. In other words, we predict that change matters, but not in isolation. The major caveat to that hypothesis is that sufficient change must be occurring to distinguish the proximal score from the alternatives at each step. If insufficient change is occurring, it is unlikely that clear evidence will be found showing that the proximal score is better than the comparison measures. Even in these cases though, the proximal score should not be a significantly less accurate predictor of imminent recidivism than the comparison measures.

Theoretical Implications

The concepts of dynamic risk and protective factors.

The proposed framework that we have just laid out, and the results found when testing that framework, have several implications for the concepts of dynamic risk and protective factors.

Firstly, in developing the revised framework, our aim was to provide a clear and simple process for examining whether variables satisfy the three essential criteria of dynamic risk or protective factors of recidivism: an association with recidivism, change over time, and an association between change and recidivism. The first step is particularly important, as it provides a simple test of all three criteria. Consequently, consistent evidence that the proximal score *does not* significantly enhance prediction

over a baseline score would suggest that the variable or variables included in that score should not be considered dynamic risk or protective factors. For the second and third steps, it is more important that the proximal score does not demonstrate *poorer* predictive accuracy than the mean or the combination of change and proximal scores; if it does, that would suggest the concepts of dynamic risk and protective factors may need further refinement.

In general, on all three steps of the framework, our results were consistent with the idea that the DRAOR subscales are dynamic predictors of imminent recidivism. There was evidence that reassessment enhanced prediction, suggesting change was associated with recidivism, and neither aggregating across measurement occasions nor including change scores alongside the most proximal score appeared to significantly improve predictive accuracy. It is important to note that our results *do not* provide evidence that each of the individual variables included in the DRAOR meet the definition of dynamic risk (or protective) factors. All analyses in this thesis have been conducted at the subscale level, not the item level. Further analyses are needed to establish which individual variables meet the requisite criteria to be classified as dynamic risk and protective factors.

Secondly, the fact that changes in scores on these subscales were associated with imminent recidivism suggests these subscales meet Kraemer et al.'s (1997) criteria of *causal risk factors*. As we discussed in Chapter 1, causal risk factors, as defined by Kraemer et al., are not equivalent to causes; however, a link between changes in a set of variables and a change in the likelihood of the outcome provides stronger evidence that those variables are involved in the causal process than a link between a single assessment of those variables and a later outcome. In that respect, our study joins a growing body of research establishing a link between changes in dynamic variables and

recidivism (Babchishin, 2013; Brown et al., 2009; T. H. Cohen et al., 2016; T. H. Cohen & VanBenschoten, 2014; de Vries Robbé, de Vogel, Douglas, et al., 2015; Greiner et al., 2015; Hanson et al., 2007; Howard & Dixon, 2013; Jones et al., 2010; Kroner & Yessine, 2013; Lewis et al., 2013; Lloyd, 2015; Olver et al., 2013; Vose et al., 2013). It remains possible that the relationship between change and recidivism is being driven by unobserved causal variables (i.e., causal mechanisms), as Ward and Beech (2015) suggested may be the case. At the very least, these results provide further evidence that the types of variables included in the DRAOR must be accounted for in any explanation of post-release recidivism, either explicitly or as proxies for the underlying causal mechanisms.

Our revised framework provides a methodology for developing a more comprehensive understanding of the relationship between dynamic variables and criminal behaviour. In particular, we think the additional step in our revised framework could lead to a more nuanced understanding of the role that different variables play in the process leading to recidivism. In Chapter 6, we discussed whether change was important at the intra-individual level *and* the inter-individual level. We can also look at this issue as a question of whether it is only the endpoint of change that matters, or the fact of change occurring that is also important. Applied to a single risk factor, the question can be framed as whether it is only *being* unemployed that matters or whether the timing of *becoming* unemployed is also important. From a theoretical perspective, we concluded that being unemployed should be the only relevant factor; differences in the timing of when two individuals become unemployed should be captured by other risk and protective factors (e.g., different attitudes or social support). Our results largely supported that theoretical argument, with limited evidence found that intra-individual change was relevant to the prediction of recidivism after controlling for the proximal

scores. This finding needs to be replicated before any firmer conclusions are drawn, but for now we hope that this distinction will be carefully considered in any future discussion and investigation of the concepts of dynamic risk and protective factors.

Finally, some of the results in this thesis are relevant to the question of whether protective factors are distinct from risk factors. It was not an aim of this thesis to establish whether the protective subscale was conceptually distinct from the two risk subscales. We also did not conduct any analyses testing the interaction between risk and protective factors, an idea that has been a focus of previous research (de Vries Robbé et al., 2013; MacDonald, 2016; Serin et al., 2016), so we make no comment on that idea. We can state that, in our sample, the stable and protective DRAOR subscales produced consistent results across all the different analyses conducted in this thesis. Specifically, the two subscales demonstrated a similar amount of change and a similar relationship with imminent recidivism, although in opposite directions. These findings appear to support the idea that protective factors may be the reverse of risk factors (Baird, 2009; Harris & Rice, 2015). However, the absence of analyses testing the interaction between risk and protective factors, or analyses testing the risk and protective factor subscales together (to see whether the protective factors represent a unique area of risk), mean that caution should be exercised with this conclusion. Further research exploring the distinction between these concepts is needed.

The distinction between stable and acute variables.

In almost all of the substantial analyses in this thesis, we found discrepancies between the DRAOR acute subscale and the stable and protective subscales. In Chapter 4, reassessment enhanced prediction to a larger extent for the acute subscale than the other two subscales. The acute subscale was the only one to demonstrate a significant improvement in predictive accuracy during several individual weeks of the follow-up. In

Chapter 5, while aggregation had little impact on predictive accuracy for the stable and protective subscales, predictive accuracy progressively worsened as increasingly more scores were included in the aggregate measures for the acute subscale. In Chapter 6, only change scores on the acute subscale demonstrated evidence of significant incremental validity over the proximal score. Similarly, in Chapter 7, it was only measures of variability on the acute subscale that were uniquely associated with imminent recidivism.

This pattern of results appears to support the existence of a conceptual distinction between stable and acute variables, a distinction that has previously lacked any supporting empirical evidence (Mann et al., 2010). Alongside Lloyd (2015), our research stands as one of the few studies to have found clear evidence of a difference between the two concepts. The most likely explanation for the consistent results across several different models is that a greater amount of change was observed on the acute subscale than the other two subscales. More change means a greater difference between the proximal scores and both the baseline and mean scores, and a greater likelihood of change scores adding incremental validity to the proximal score. This difference is consistent with the conceptual distinction between acute and stable variables based on how frequently and rapidly they are expected to change (Hanson et al., 2007; Hanson & Harris, 2000).

There is, however, a major limitation with this explanation. As we have highlighted at several points in the thesis, the process for scoring the DRAOR subscales is not equivalent. Supervision officers are required to reassess the acute subscale after every session, but are only supposed to keep a “watching brief” on the stable and protective items (Serin et al., 2012, p. 4). It is possible that these instructions were at least partly responsible for the observed differences in the frequency and amount of

change on the different subscales, and thus, the differences in the impact of change on predictive accuracy. One way of addressing this limitation would be to complete a similar set of analyses using an assessment regime where theoretically stable and acute variables are scored equivalently. The problem with that approach is that it would place a substantial additional burden on risk assessors. For example, if that approach was taken with the DRAOR, it would mean an increase from 7 to as many as 19 variables that need to be rated after every supervision contact, nearly tripling the amount of work required by probation officers. This additional burden may also affect the integrity of the scoring, which may impact on predictive accuracy. Future research that aims to address this limitation with our findings must account for these practical factors.

Even if the distinction between the subscales was valid and not an artificial product of the scoring procedure, it remains unclear to what extent our results support the existing conceptualisation of acute dynamic risk factors. The results suggested that acute subscale scores changed more rapidly than stable or protective subscale scores, a finding consistent with the existing conceptualisation of acute risk factors (Hanson et al., 2007; Hanson & Harris, 2000). Acute risk factors, or rather *changes* in acute variables, are also supposed to signal imminent recidivism. We found inconsistent evidence supporting this feature of acute risk factors. In our sample, change in the acute subscale had a larger effect on predictive accuracy than for the other two subscales. In isolation though, the relationship between change scores and imminent recidivism was not very strong. For example, recidivism was almost equally as likely to follow a decrease in acute subscale scores as an increase. The results showed that change did matter, but primarily only when enough other acute risk factors were present after the change (i.e., a high overall acute subscale score at the proximal assessment). Therefore, the point highlighted in the previous section about the need to consider change in the

context of the overall level of risk is particularly relevant to the concept of acute risk factors. Our framework and results suggest acute risk factors should not be thought of as factors that signal imminent recidivism; rather, they should be considered variables that signal imminent recidivism *when enough other acute risk factors are present*.

One possible explanation for this finding is that assessments did not occur frequently enough to accurately capture the link between change and recidivism. For example, one of the ways that Beech and Ward (2004) conceptualised acute risk factors was as internal states such as anger, which may only last for a matter of minutes or hours. It is unlikely that the weekly assessments used in our study were sufficient to capture changes of this description, which may explain the weak relationship between change and imminent recidivism observed in our study. For a better test of the concept of acute risk factors, it would be beneficial to have even more frequent assessment scores than were available in this study; however, the practical and ethical limitations that we have highlighted previously are likely to make this challenging to achieve.

Our results challenged the existing conceptualisation of acute risk factors in another way. The idea that these variables *signal* imminent recidivism implicitly suggests that recidivism will follow shortly after an increase in acute risk (e.g., losing a job). Remembering that change should not be considered in isolation, we might more accurately say that a job loss is likely to signal imminent recidivism when it occurs in the context of a difficult living situation, negative mood, and substance abuse, for example. In our study, there was some evidence that acute subscale scores increased immediately prior to recidivism. Short-term change scores were also significantly associated with imminent recidivism. However, the descriptive statistics suggested the best explanation for this finding was that recidivists showed little to no change on the acute subscale compared to non-recidivists who were observed to decrease their acute

scores. This type of pattern has been observed previously (Hanby, 2013; Lloyd, 2015) but not while measuring change over the short time periods used in our study. This finding needs to be replicated with research looking at short time periods prior to recidivism to draw further conclusions about how acute variables might signal imminent recidivism.

Finally, the same caveat highlighted in the previous section regarding individual items applies here. Our results do *not* show that the DRAOR acute items—substance use, anger/hostility, opportunity/access to victims, negative mood, employment, interpersonal relationships, and living situation—should be classified as acute dynamic risk factors. Together, these variables appear to be behaving (statistically) in a way that we might expect of acute factors, but we cannot say with certainty that any single variable should be considered an acute dynamic risk factor. The factor analyses presented in Chapter 3 (using baseline scores) provided some evidence that the individual variables on each of the subscales share a common relationship. With this type of data though, the tests for longitudinal measurement invariance in Chapter 4 were arguably more important; these tests did not clearly establish that the same factor structure was present throughout the follow-up. Overall, although risk assessment should occur at the aggregate level (Kroner et al., 2007), it would be beneficial to know whether any individual variables are driving the predictive validity of the acute subscale or behaving more like stable rather than acute risk factors.

Individual item-level analyses would also allow examination of the interaction between stable (and protective) factors and different acute risk factors. We have already highlighted the point that change in risk cannot be viewed in isolation. For example, change in acute risk factors was primarily relevant to recidivism when it meant that an individual had a high acute subscale score. We did not examine how

changes in acute variables impacted on the likelihood of recidivism after controlling for the proximal level of stable variables, or vice versa. Little is known about the nature of this interaction (Mann et al., 2010; Ward, 2016; Ward & Fortune, 2015), so it represents an area for further theoretical development. That progress relies on first being able to reliably distinguish stable and acute variables based on their changeability and relationship to imminent recidivism, a task that has hopefully been made easier by the framework developed and the evidence presented in this thesis.

Practical Implications

Community supervision.

Our results have several implications for community supervision. First, although we have discussed in detail *why* change impacts on prediction, it is important to emphasise the initial point that change does appear to impact on the prediction of imminent recidivism. The results showed that reassessment of all three DRAOR subscales enhanced the prediction of imminent recidivism. Even change scores across shorter time periods were associated with recidivism after controlling for baseline. The best interpretation of this finding is that when individuals' subscale scores changed over time, the likelihood of those individuals being reconvicted also changed. In this sense, the change we observed appeared to be true change, and not (solely) the product of measurement error. This point should be central to the work of community supervision officers. It suggests that when changes in DRAOR subscale scores occur, those changes should be treated as meaningful and supervision practice should be altered accordingly. Following the risk principle of the RNR model of correctional intervention (Bonta & Andrews, 2016), as individuals become higher risk they should receive more intensive supervision, and as they become lower risk, they should receive less intensive supervision.

The implications of our results for the need principle of the RNR model are less clear. The need principle states that intervention, including community supervision, should target dynamic risk factors as the mechanism that will reduce the likelihood of recidivism (Bonta & Andrews, 2016). The fact that changes in the dynamic variables on the DRAOR were associated with changes in the likelihood of recidivism in our sample provides support for this principle. However, a major limitation of our study is that we did not investigate what efforts were already being made to change the dynamic risk and protective factors. We can speculate that officers were targeting these variables during supervision and that these efforts contributed to the change. However, the change could also have been driven by the parolees themselves, or other individuals supporting the parolee such as case managers, psychologists, friends, or family members. Alternatively, as Lloyd (2015) suggested, the observed changes may reflect the officers gradually establishing the true risk level of the individuals they were supervising, implying that it is only the officer's ratings and not the parolee's true risk level that changed. Further research looking at the practice of completing and responding to DRAOR assessments is needed to tease apart these different possible explanations.

The need principle states which variables should be targeted to produce meaningful change; it does not provide much guidance about how to respond to *change* in those variables. On this point, our results provide some insight. The primary message from this thesis is that supervision officers and community correctional agencies should be responding primarily to the most recent assessment of dynamic risk and protective factors as the most recent assessment provides the most accurate indication of an individual's likelihood of being reconvicted. There was some evidence that change on the acute subscale might be relevant in addition to the most recent assessment, but this

evidence was weak. The better conclusion from our findings is that if an individual's risk has increased recently but they remain moderate risk overall, that individual should continue to be supervised and monitored in the same way as other moderate risk individuals, and not treated as high risk.

This recommendation—that supervision officers should only respond to current risk scores, and not the change in scores—is tentative, as we did not examine how officers responded to the changes observed in our dataset. Lloyd (2015) noted this limitation in his thesis too, hypothesising that supervision officers may respond to their own assessments in a way that reduces the predictive accuracy of those assessments (e.g., in response to an assessment, an officer intervenes in some way that makes recidivism less likely, but that change is not captured by DRAOR scores). This point would seem to be even more relevant to our study than for Lloyd because (a) we used a higher static-risk sample, and (b) we looked at the relationship between recidivism and change in DRAOR scores occurring over shorter time periods. The high-risk sample is relevant because, according to the risk principle, supervision officers should have been providing more attention to the individuals in our sample, making it more likely they would have intervened following a DRAOR assessment. Supervision officers intervening in response to sudden increases in risk could have had a particularly strong effect on the relationship between these short-term change scores and recidivism.

This limitation could be addressed by research integrating the framework developed in this thesis with an examination of supervision practice. This type of research could highlight how supervision officers currently respond to dynamic risk assessments and how that response impacts on the likelihood of recidivism. It could also further advance our understanding of the role of dynamic risk and protective factors in the recidivism process. Initial research looking at supervision practice found

relatively poor particular adherence to the RNR model (Bonta, Rugge, Scott, Bourgon, & Yessine, 2008). As a result, several training programmes for community supervision officers were developed and research suggests these programmes lead to a reduction in recidivism (Chadwick et al., 2015). There is also an increasing body of research investigating the extent to which supervision officers and case managers develop individually-tailored intervention plans based on dynamic risk assessments, and the extent to which these individualised interventions reduce recidivism (Bosker, Witteman, & Hermanns, 2013; Luong & Wormith, 2011; Singh et al., 2014; Vieira, Skilling, & Peterson-Badali, 2009; Vitopoulos, Peterson-Badali, & Skilling, 2012). Integration of the framework developed in this thesis could be a next step for these areas of research, with important implications for how supervision officers should respond to *changes* in dynamic risk and protective factors.

Other areas of correctional practice.

The framework and findings of this thesis are relevant to other areas of correctional practice, in addition to community supervision. Most relevantly, the results have important implications for management and intervention with individuals in prison. As we have highlighted, much of the research on change in dynamic risk and protective factors comes from institutional treatment programmes (Beggs, 2010; Serin et al., 2013). In that context, reassessments are important for tracking treatment progress. Those assessments can take on increased importance if they are used to inform decisions about release from prison. The framework developed in our thesis suggests further caution when using this type of measurement of change to make decisions with such serious implications. Our results suggest that intra-individual change is important but the risk level at the end of treatment should be the most important factor in determining an individual's likelihood of recidivism. As we pointed

out in Chapter 6 though, there are plausible reasons to think that different results might be observed in the institutional treatment context. For example, several variables included in the DRAOR (e.g., living situation) will not be relevant in prison. Also, there is likely to be a longer follow-up time, providing more time for unobserved change to occur following the most proximal assessment, and making it more likely that change will be associated with recidivism after controlling for proximal risk. Therefore, our findings need to be replicated before they can be directly applied in that context.

Our findings also highlight the need for a continuous approach to management and supervision of individuals who are leaving prison and re-entering the community. At several points in this thesis, we have mentioned the high number of men who were reconvicted of an offence committed within a few days or weeks of leaving prison. This finding, while expected on the basis of the high-risk sample and previous research (Nadesu, 2009), highlights the fact that more needs to be done to address the challenges these men face upon leaving prison. An important step in determining the best way to address those challenges is understanding the variables that make recidivism more likely so soon after release. In that respect, we believe the focus in this thesis on *imminent* recidivism is a step in the right direction. Our results replicated Lloyd's (2015) finding that DRAOR subscale scores can reliably distinguish individuals who are most likely to be reconvicted within a very short time (i.e., 6 weeks). Our study extends Lloyd's findings by showing that, even within a sample group who were all classified as high-risk based on their static risk scores, the DRAOR subscales provide a meaningful indication of the likelihood of imminent recidivism. Further research should seek to link dynamic risk assessments completed in prison to assessments occurring shortly after release. Existing research on dynamic risk factors and release planning (Dickson & Polaschek, 2014; Richards, 2017; Scoones et al., 2012), and on the relationship between

change in dynamic factors in prison and change in dynamic factors in the community (Yesberg, 2015), presents a way forward in this area.

General Limitations and Future Research

There are a few final limitations of this research that we should highlight, along with ideas for future research that can address these limitations. First, the majority of the statistical effects observed in this thesis have been small to moderate. For example, although the most proximal score enhanced predictive accuracy, the degree of improvement was small, and overall predictive accuracy using the proximal score of any of the three DRAOR subscales was moderate at best. Predictive accuracy was slightly higher when the static risk measure—the RoC*RoI—was included alongside the DRAOR, but predictive accuracy still remained only moderate (Rice & Harris, 2005). This point suggests caution should be applied to the theoretical and practical implications already discussed in this chapter. While our study adds to the growing body of research highlighting the benefits of frequently reassessing dynamic variables alongside static variables to enhance predictive accuracy (Brown et al., 2009; Jones et al., 2010; Lloyd, 2015), risk assessment tools remain an imperfect method of predicting recidivism. The framework presented in this thesis hopefully provides a method for continuing to improve predictive accuracy.

Second, in this thesis we have only examined a single recidivism outcome: reconvictions for any new offence. Therefore, the extent to which the implications discussed in this chapter will apply to other recidivism outcomes such as violent or sexual recidivism is uncertain. More importantly, of the 47.6% of men in our sample who were categorised as recidivists, nearly half were convicted for a breach of their parole conditions. This fact raises the same issue identified by Lloyd (2015): since supervision officers are responsible for initiating breach proceedings, officers may have

had advance warning that a breach was coming. If this knowledge was reflected in increased DRAOR scores, predictive accuracy may have been unfairly inflated. This issue was even more important to our study because of the focus on short-term change that was not present in Lloyd's research. However, at the same time, this focus on short-term change scores arguably provides the advantage of being able to look closely at whether there was any evidence to suggest that this phenomenon was occurring. Our results showed that short-term change scores were very weakly associated with imminent recidivism, suggesting that it was unlikely that this practice was occurring systematically.

Finally, Lloyd (2015) highlighted several limitations of the DRAOR data, each of which remains relevant to the current thesis. Briefly, these limitations were data nesting within supervision officers, differences in fidelity across supervision officers, and the absence of any inter-rater reliability for DRAOR scores. The limitations related to supervision officers have not been addressed in this thesis. Although not measured, it is very likely that the same supervision officers were supervising multiple men in our sample. Thus, it is possible that differential effects across supervision officers may have been present in our dataset and these differential effects may be partly or wholly the result of differences in fidelity of DRAOR administration between officers. Further research that is able to account for this factor would be beneficial. Similarly, there has still not been any research looking at the inter-rater reliability of DRAOR scores, so that same limitation applies to our dataset. It is arguably more important from a theoretical than a practical perspective that a risk assessment tool demonstrates construct validity. Hence, this limitation applies particularly to the theoretical implications highlighted previously in this chapter. Overall though, while further research examining these limitations would be beneficial, it is more likely that these factors explain the weak

predictive accuracy highlighted earlier in this section as opposed to explaining why the effects were observed at all (i.e., they are more likely to have decreased rather than increased the predictive accuracy and the magnitude of the effects observed).

Summary

In this chapter, we have provided a summary of the framework developed over the course of this thesis and the results obtained when testing that framework. The results of those tests found that the subscales of the DRAOR, which were the primary predictor variables in this thesis, demonstrated a series of statistical properties that are conceptually consistent with dynamic risk and protective factors. The results also highlighted the importance of not considering change in these variables in isolation; change matters but generally only to the extent to which it provides a measure of the current or proximal likelihood of recidivism. This point represents a particular challenge to the concept of acute dynamic risk factors. Our results did support the existence of a conceptual distinction between acute and stable dynamic variables, but it was possible that those findings were the result of the way in which the DRAOR is scored.

Perhaps the major strength of this research is its ecological validity. The dataset was comprised of dynamic risk assessments completed by probation officers in the course of their regular practice. Consequently, there were several practical implications of the findings. Most importantly, the results provided further support for the importance of conducting reassessment of dynamic risk and protective factors, with analyses showing that reassessment enhanced the prediction of recidivism. The results also suggested that proximal risk should be the central factor for individuals and agencies involved in community supervision; while change matters for individuals, it should not be a relevant consideration when comparing two (or more) individuals. The

framework we have developed would be well-suited to further research that also considers how supervision officers respond to changes in dynamic risk and protective factors. And finally, our findings also have potential to impact on correction practice in prison settings; however, further research should seek to establish that our findings do generalise to that context before any recommendations are made in that area.

References

- Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20*, 93-114.
- Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Albany, NY: Lexis Nexis/Anderson.
- Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17*, 19-52.
- Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency, 52*, 7-27.
doi:10.1177/0306624X05282556
- Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology, 28*, 369-404.
doi:10.1111/j.1745-9125.1990.tb01330.x
- Asparouhov, T., & Muthén, B. (2006). Robust chi square difference testing with mean and variance adjusted test statistics. *Mplus Web Notes*. No. 10.
Retrieved from
<http://statmodel2.com/download/webnotes/webnote10.pdf>
- Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 397-438.
doi:10.1080/10705510903008204
- Babchishin, K. M. (2013). *Sex offenders do change on risk-relevant propensities: Evidence from a longitudinal study of the ACUTE-2007*. (Unpublished PhD thesis), Carleton University, Ottawa, ON.

- Baglivio, M. T., Wolff, K. T., Jackowski, K., & Greenwald, M. A. (2017). A multilevel examination of risk/need change scores, community context, and successful reentry of committed juvenile offenders. *Youth Violence and Juvenile Justice*, 15, 38-61. doi:10.1177/1541204015596052
- Baglivio, M. T., Wolff, K. T., Piquero, A. R., DeLisi, M., & Vaughn, M. G. (2018). The effects of changes in dynamic risk on reoffending among serious juvenile offenders returning from residential placement. *Justice Quarterly*, 35, 443-476. doi:10.1080/07418825.2017.1317013
- Baglivio, M. T., Wolff, K. T., Piquero, A. R., Howell, J. C., & Greenwald, M. A. (2017). Risk assessment trajectories of youth during juvenile justice residential placement: Examining risk, promotive, and “buffer” scores. *Criminal Justice and Behavior*, 44, 360-394. doi:10.1177/0093854816668918
- Baird, C. (2009). *A question of evidence: A critique of risk assessment models used in the justice system*. Madison, WI: National Council on Crime and Delinquency. Retrieved from https://www.nccdglobal.org/sites/default/files/publication_pdf/special-report-evidence.pdf
- Bakker, L. W., Riley, D., & O'Malley, J. (1999). *Risk of reconviction: Statistical models predicting four types of re-offending*. Wellington, New Zealand: Department of Corrections
- Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, N.J.: Prentice Hall.
- Beech, A. R., & Ward, T. (2004). The integration of etiology and risk in sexual offenders: A theoretical framework. *Aggression and Violent Behavior*, 10, 31-63. doi:10.1016/j.avb.2003.08.002

- Beggs, S. M. (2010). Within-treatment outcome among sexual offenders: A review. *Aggression and Violent Behavior, 15*, 369-379.
doi:10.1016/j.avb.2010.06.005
- Beggs, S. M., & Grace, R. C. (2010). Assessment of dynamic risk factors: An independent validation study of the Violence Risk Scale: Sexual Offender Version. *Sexual Abuse: A Journal of Research and Treatment, 22*, 234-251.
doi:10.1177/1079063210369014
- Beggs, S. M., & Grace, R. C. (2011). Treatment gain for sexual offenders against children predicts reduced recidivism: A comparative validity study. *Journal of Consulting and Clinical Psychology, 79*, 182-192.
doi:10.1037/a0022900
- Bonta, J. (1996). Risk-needs assessment and treatment. In A. T. Hartland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply*. (pp. 18-32). Thousand Oaks, CA: Sage Publications.
- Bonta, J., & Andrews, D. A. (2016). *The psychology of criminal conduct* (6th ed.). London: Routledge.
- Bonta, J., Bourgon, G., Rugge, T., Scott, T.-L., Yessine, A. K., Gutierrez, L., & Li, J. (2011). An experimental demonstration of training probation officers in evidence-based community supervision. *Criminal Justice and Behavior, 38*, 1127-1148. doi:10.1177/0093854811420678
- Bonta, J., Rugge, T., Scott, T.-L., Bourgon, G., & Yessine, A. K. (2008). Exploring the black box of community supervision. *Journal of Offender Rehabilitation, 47*, 248-270. doi:10.1080/10509670802134085
- Bosker, J., Witteman, C., & Hermanns, J. (2013). Do intervention plans meet criteria for effective practice to reduce recidivism? How probation

- officers forget about social capital and basic needs. *European Journal of Probation*, 5, 65-85. doi:10.1177/206622031300500105
- Brown, S., St. Amand, M., & Zamble, E. (2009). The dynamic prediction of criminal recidivism: A three-wave prospective study. *Law and Human Behavior*, 33, 25-45. doi:10.1007/s10979-008-9139-7
- Burnett, R. (2009). Post-corrections reintegration: Prisoner resettlement and desistance from crime. In J. R. Adler & J. M. Gray (Eds.), *Forensic Psychology: Concepts, debates and practice* (2nd ed., pp. 508-528). Oxon: Willan Publishing.
- Campbell, M. A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders. *Criminal Justice and Behavior*, 36, 567-590. doi:10.1177/0093854809333610
- Caudy, M. S., Durso, J. M., & Taxman, F. S. (2013). How well do dynamic needs predict recidivism? Implications for risk assessment and risk reduction. *Journal of Criminal Justice*, 41, 458-466. doi:10.1016/j.jcrimjus.2013.08.004
- Chadwick, N. (2014). *Validating the Dynamic Risk Assessment for Offender Re-entry (DRAOR) in a sample of US probationers and parolees*. (Unpublished Master's thesis), Carleton University Ottawa, Canada,
- Chadwick, N., Dewolf, A., & Serin, R. (2015). Effectively training community supervision officers: A meta-analytic review of the impact on offender outcome. *Criminal Justice and Behavior*, 42, 977-989. doi:10.1177/0093854815595661

- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 464-504. doi:10.1080/10705510701301834
- Chu, C. M., Thomas, S. D. M., Daffern, M., & Ogloff, J. R. P. (2013). Should clinicians use average or peak scores on a dynamic risk - assessment measure to most accurately predict inpatient aggression? *International Journal of Mental Health Nursing*, 22, 493-499. doi:10.1111/j.1447-0349.2012.00846.x
- Cohen, M. A., Piquero, A. R., & Jennings, W. G. (2010). Studying the costs of crime across offender trajectories. *Criminology & Public Policy*, 9, 279-305. doi:10.1111/j.1745-9133.2010.00627.x
- Cohen, T. H., Lowenkamp, C. T., & VanBenschoten, S. W. (2016). Does change in risk matter? Examining whether changes in offender risk characteristics influence recidivism outcomes. *Criminology & Public Policy*, 15, 263-296. doi:10.1111/1745-9133.12190
- Cohen, T. H., & VanBenschoten, S. W. (2014). Does the risk of recidivism for supervised offenders improve over time: Examining changes in the dynamic risk characteristics for offenders under federal supervision. *Federal Probation*, 78, 41-56.
- de Vogel, V., de Vries Robbé, M., de Ruiter, C., & Bouman, Y. H. (2011). Assessing protective factors in forensic psychiatric practice: Introducing the SAPROF. *International Journal of Forensic Mental Health*, 10, 171-177. doi:10.1080/14999013.2011.600230
- de Vries Robbé, M., de Vogel, V., & de Spa, E. (2011). Protective factors for violence risk in forensic psychiatric patients: A retrospective validation

- study of the SAPROF. *International Journal of Forensic Mental Health*, 10, 178-186. doi:10.1080/14999013.2011.600232
- de Vries Robbé, M., de Vogel, V., & Douglas, K. S. (2013). Risk factors and protective factors: A two-sided dynamic approach to violence risk assessment. *Journal of Forensic Psychiatry & Psychology*, 24, 440-457. doi:10.1080/14789949.2013.818162
- de Vries Robbé, M., de Vogel, V., Douglas, K. S., & Nijman, H. L. (2015). Changes in dynamic risk and protective factors for violence during inpatient forensic psychiatric treatment: Predicting reductions in postdischarge community recidivism. *Law and Human Behavior*, 39, 53-61. doi:10.1037/lhb0000089
- de Vries Robbé, M., de Vogel, V., Koster, K., & Bogaerts, S. (2015). Assessing protective factors for sexually violent offending with the SAPROF. *Sexual Abuse: A Journal of Research and Treatment*, 27, 51-70. doi:10.1177/1079063214550168
- de Vries Robbé, M., de Vogel, V., & Stam, J. (2012). Protective factors for violence risk: The value for clinical practice. *Psychology*, 3, 1259. doi:10.4236/psych.2012.312A187
- de Vries Robbé, M., & Willis, G. M. (2017). Assessment of protective factors in clinical practice. *Aggression and Violent Behavior*, 32, 55-63. doi:10.1016/j.avb.2016.12.006
- Dickson, S. R., & Polaschek, D. L. L. (2014). Planning to avoid risk or planning for a positive life: the relationship between release plan valence and reoffending. *International Journal of Offender Therapy and Comparative Criminology*, 58, 1431-1448. doi:10.1177/0306624X13502631

- Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law*, 11, 347-383. doi:10.1037/1076-8971.11.3.347
- Dowden, C., & Andrews, D. A. (1999a). What works for female offenders: A meta-analytic review. *Crime & Delinquency*, 45, 438-452. doi:10.1177/0011128799045004002
- Dowden, C., & Andrews, D. A. (1999b). *What works in young offender treatment: A meta-analysis*. Paper presented at the Forum on corrections research.
- Dowden, C., & Andrews, D. A. (2000). Effective correctional treatment and violent reoffending: A meta-analysis. *Canadian Journal of Criminology*, 42, 449-467.
- Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, 37, 1097-1126. doi:10.1037/0022-3514.37.7.1097
- Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist*, 35, 790-806. doi:10.1037/0003-066X.35.9.790
- Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, 51, 360-392. doi:10.1111/j.1467-6494.1983.tb00338.x
- Epstein, S. (1986). Does aggregation produce spuriously high estimates of behavior stability? *Journal of Personality and Social Psychology*, 50, 1199-1210. doi:10.1037/0022-3514.50.6.1199
- Farrington, D. P., Coid, J. W., Harnett, L., Jolliffe, D., Soteriou, N., Turner, R., & West, D. J. (2006). Criminal careers up to age 50 and life success up to age

48: New findings from the Cambridge Study in Delinquent Development.

Home Office Research Study, 299.

Farrington, D. P., Jolliffe, D., Hawkins, J. D., Catalano, R. F., Hill, K. G., & Kosterman, R. (2003). Comparing delinquency careers in court records and self-reports. *Criminology, 41*, 933-958. doi:10.1111/j.1745-9125.2003.tb01009.x

Farrington, D. P., Jolliffe, D., Loeber, R., & Homish, D. L. (2007). How many offenses are really committed per Juvenile Court Offender? *Victims & Offenders, 2*, 227-249. doi:10.1080/15564880701403934

Farrington, D. P., Ttofi, M. M., & Piquero, A. R. (2016). Risk, promotive, and protective factors in youth offending: Results from the Cambridge study in delinquent development. *Journal of Criminal Justice, 45*, 63-70. doi:10.1016/j.jcrimjus.2016.02.014

Fazel, S., & Wolf, A. (2015). A systematic review of criminal recidivism rates worldwide: Current difficulties and recommendations for best practice. *PLoS One, 10*, e0130390. doi:10.1371/journal.pone.0130390

Ferguson, A. (2015). *An age-old issue: Evaluating the applicability of adult criminal risk assessment tools for use with youth offenders*. (Unpublished Master's thesis), Victoria University of Wellington, New Zealand.

Flores, A. W., Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2006). Predicting outcome with the Level of Service Inventory-Revised: The importance of implementation integrity. *Journal of Criminal Justice, 34*, 523-529. doi:10.1016/j.jcrimjus.2006.09.007

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to

- improving quality of intervention research implementation. *Clinical Psychology Review*, 31, 79-88. doi:10.1016/j.cpr.2010.09.007
- Gendreau, P., Goggin, C., & Smith, P. (2002). Is the PCL-R really the “unparalleled” measure of offender risk? A lesson in knowledge cumulation. *Criminal Justice and Behavior*, 29, 397-426. doi:10.1177/0093854802029004004
- Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, 34, 575-607. doi:10.1177/0093854896023001010
- Gluckman, P. (2018). *Using evidence to build a better justice system: The challenge of rising prison costs*. Wellington, New Zealand: Office of the Prime Minister’s Chief Science Advisor. Retrieved from <http://apo.org.au/system/files/139026/apo-nid139026-688211.pdf>
- Greiner, L. E., Law, M. A., & Brown, S. L. (2015). Using dynamic factors to predict recidivism among women: A four-wave prospective study. *Criminal Justice and Behavior*, 42, 457-480. doi:10.1177/0093854814553222
- Hanby, L. J. (2013). *A longitudinal study of dynamic risk, protective factors, and criminal recidivism: Change over time and the impact of assessment timing*. (Unpublished PhD thesis), Carleton University, Ottawa, ON.
- Hanson, R. K. (2008). What statistics should we use to report predictive accuracy. *Crime Scene*, 15, 15-17.
- Hanson, R. K., Harris, A. J., Scott, T.-L., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project*. (User Report No. 2007-05). Ottawa, ON: Public Safety Canada

- Hanson, R. K., & Harris, A. J. R. (2000). Where should we intervene? Dynamic predictors of sexual offense recidivism. *Criminal Justice and Behavior*, 27, 6-35. doi:10.1177/0093854800027001002
- Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment*, 21, 1-21. doi:10.1037/a0014421
- Harris, G. T., & Rice, M. E. (2015). Progress in violence risk assessment and communication: Hypothesis versus evidence. *Behavioral Sciences & the Law*, 33, 128-145. doi:10.1002/bsl.2157
- Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61, 92-105. doi:10.1111/j.0006-341X.2005.030814.x
- Heffernan, R., & Ward, T. (2015). The conceptualization of dynamic risk factors in child sex offenders: An agency model. *Aggression and Violent Behavior*, 24, 250-260. doi:10.1016/j.avb.2015.07.001
- Helmus, L. M., Hanson, R. K., Babchishin, K. M., & Mann, R. E. (2013). Attitudes supportive of sexual offending predict recidivism: A meta-analysis. *Trauma, Violence, & Abuse*, 14, 34-53. doi:10.1177/1524838012462244
- Howard, P. D., & Dixon, L. (2013). Identifying change in the likelihood of violent recidivism: Causal dynamic risk factors in the OASys violence predictor. *Law and Human Behavior*, 37, 163-174. doi:10.1037/lhb0000012
- Hudson, S. M., Wales, D. S., Bakker, L., & Ward, T. (2002). Dynamic risk factors: The Kia Marama evaluation. *Sexual Abuse: A Journal of Research and Treatment*, 14, 103-119.

- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.
- James, N. (2015). *Risk and needs assessment in the criminal justice system*. (7-5700). Congressional Research Service. Retrieved from https://www.everycrsreport.com/files/20150724_R44087_0c47cc191ec_c982888fa182c82ef0099a86eca8d.pdf
- Jones, N. J., Brown, S. L., & Zamble, E. (2010). Predicting criminal recidivism in adult male offenders: Researcher versus parole officer assessment of dynamic risk. *Criminal Justice and Behavior, 37*, 860-882. doi:10.1177/0093854810368924
- Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*, 347-387. doi:10.1207/S15327906347-387
- Kennealy, P. J., Skeem, J. L., Manchak, S. M., & Eno Louden, J. (2012). Firm, fair, and caring officer-offender relationships protect against supervision failure. *Law and Human Behavior, 36*, 496-505. doi:10.1037/h0093935
- Klepfisz, G., O'Brien, K., & Daffern, M. (2014). Violent offenders' within-treatment change in anger, criminal attitudes, and violence risk: Associations with violent recidivism. *International Journal of Forensic Mental Health, 13*, 348-362. doi:10.1080/14999013.2014.951107
- Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1997). Coming to terms with the terms of risk. *Archives of General Psychiatry, 54*, 337-343. doi:10.1001/archpsyc.1997.01830160065009

- Kroner, D. G., Gray, A. L., & Goodrich, B. (2013). Integrating risk context into risk assessments: The Risk Context Scale. *Assessment, 20*, 135-149.
doi:10.1177/1073191111429613
- Kroner, D. G., Mills, J. F., Reitzel, L. R., Dow, E., Aufderheide, D. H., & Railey, M. G. (2007). Directions for violence and sexual risk assessment in correctional psychology. *Criminal Justice and Behavior, 34*, 906-918.
doi:10.1177/0093854807301559
- Kroner, D. G., & Yessine, A. K. (2013). Changing risk factors that impact recidivism: In search of mechanisms of change. *Law and Human Behavior, 37*, 321-336. doi:10.1037/lhb0000022
- Labrecque, R. M., Smith, P., Lovins, B. K., & Latessa, E. J. (2014). The importance of reassessment: How changes in the LSI-R risk score can improve the prediction of recidivism. *Journal of Offender Rehabilitation, 53*, 116-128.
doi:10.1080/10509674.2013.868389
- Latessa, E. J., Cullen, F. T., & Gendreau, P. (2002). Beyond correctional quackery: Professionalism and the possibility of effective treatment. *Federal Probation, 66*, 43-49.
- Lewis, K., Olver, M. E., & Wong, S. C. (2013). The Violence Risk Scale: Predictive validity and linking changes in risk with violent recidivism in a sample of high-risk offenders with psychopathic traits. *Assessment, 20*, 150-164.
doi:10.1177/1073191112441242
- Lipsey, M. W., & Cullen, F. T. (2007). The effectiveness of correctional rehabilitation: A review of systematic reviews. *Annual Review of Law and Social Science, 3*, 297-320.
doi:10.1146/annurev.lawsocsci.3.081806.112833

- Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, 22, 486-506. doi:10.1037/met0000075
- Lloyd, C. D. (2015). *Can a dynamic risk instrument make short-term predictions in "real time"? Developing a framework for testing proximal assessment of offender recidivism risk during re-entry*. (Unpublished PhD thesis), Carleton University, Ottawa, ON.
- Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, 16, 421-437. doi:10.1177/001316445601600401
- Lösel, F., & Farrington, D. P. (2012). Direct protective and buffering protective factors in the development of youth violence. *American Journal of Preventive Medicine*, 43, S8-S23. doi:10.1016/j.amepre.2012.04.029
- Luong, D., & Wormith, J. S. (2011). Applying risk/need assessment to probation practice and its impact on the recidivism of young offenders. *Criminal Justice and Behavior*, 38, 1177-1199. doi:10.1177/0093854811421596
- MacDonald, I. M. (2016). *Protective factors in the sexual offending area: Analysis of the concept and a preliminary model*. (Unpublished Master's thesis), Victoria University of Wellington,
- Mann, R. E., Hanson, R. K., & Thornton, D. (2010). Assessing risk for sexual recidivism: Some proposals on the nature of psychologically meaningful risk factors. *Sexual Abuse: A Journal of Research and Treatment*, 22, 191-217. doi:10.1177/1079063210366039
- Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling,

- integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 439-476. doi:10.1080/10705510903008220
- Martinson, R. (1974). What works?-Questions and answers about prison reform. *The Public Interest*, 35, 22-54.
- McGuire, J. (2002). Criminal sanctions versus psychologically-based interventions with offenders: A comparative empirical analysis. *Psychology, Crime & Law*, 8, 183-208. doi:10.1080/10683160208415005
- Monahan, J., & Skeem, J. L. (2014). Risk redux: The resurgence of risk assessment in criminal sanctioning. *Federal Sentencing Reporter*, 26, 158-166. doi:10.1525/fsr.2014.26.3.158
- Mulvey, E. P., Schubert, C. A., Pitzer, L., Hawes, S., Piquero, A., & Cardwell, S. (2016). An examination of change in dynamic risk of offending over time among serious juvenile offenders. *Journal of Criminal Justice*, 45, 48-53. doi:10.1016/j.jcrimjus.2016.02.008
- Nadesu, A. (2007). *Reconviction patterns of released prisoners: A 36-months follow-up analysis*. Wellington, New Zealand: Department of Corrections
- Nadesu, A. (2009). *Reconviction patterns of offenders managed in the community: A 60-months follow-up analysis*. Wellington, New Zealand: Department of Corrections
- Ogloff, J. R., & Daffern, M. (2006). The dynamic appraisal of situational aggression: An instrument to assess risk for imminent aggression in psychiatric inpatients. *Behavioral Sciences & the Law*, 24, 799-813. doi:10.1002/bsl.741

- Olver, M. E., Lewis, K., & Wong, S. C. (2013). Risk reduction treatment of high-risk psychopathic offenders: The relationship of psychopathy and treatment change to violent recidivism. *Personality Disorders: Theory, Research, and Treatment*, 4, 160-167. doi:10.1037/a0029769
- Olver, M. E., & Wong, S. C. (2011). A comparison of static and dynamic assessment of sexual offender risk and need in a treatment context. *Criminal Justice and Behavior*, 38, 113-126. doi:10.1177/0093854810389534
- Pettersen, C., Nunes, K. L., Woods, M., Maimone, S., Hermann, C. A., Looman, J., & Spape, J. (2015). Does change in hostility predict sexual recidivism? *International Journal of Offender Therapy and Comparative Criminology*, 59, 565-587. doi:10.1177/0306624X14547033
- Piquero, A. R., Jennings, W. G., & Farrington, D. P. (2013). The monetary costs of crime to middle adulthood: Findings from the Cambridge Study in Delinquent Development. *Journal of Research in Crime and Delinquency*, 50, 53-74. doi:10.1177/0022427811424505
- Polaschek, D. L. L. (2012). An appraisal of the risk–need–responsivity (RNR) model of offender rehabilitation and its application in correctional treatment. *Legal and Criminological Psychology*, 17, 1-17. doi:10.1111/j.2044-8333.2011.02038.x
- Polaschek, D. L. L. (2016). Do relationships matter? Examining the quality of probation officers' interactions with parolees in preventing recidivism. *Practice—The New Zealand Corrections Journal*, 4(1), 5-9.

- Polaschek, D. L. L. (2017). Protective factors, correctional treatment and desistance. *Aggression and Violent Behavior, 32*, 64-70.
doi:10.1016/j.avb.2016.12.005
- Polaschek, D. L. L., Bell, R. K., Casey, A. R., Dickson, S. R., & Yesberg, J. A. (2014). *The New Zealand parole project* [Unpublished data].
- Polaschek, D. L. L., & Kilgour, T. G. (2013). New Zealand's special treatment units: The development and implementation of intensive treatment for high-risk male prisoners. *Psychology, Crime & Law, 19*, 511-526.
doi:10.1080/1068316X.2013.759004
- R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>.
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*, 354-373. doi:10.1037/a0029315
- Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*, 615-620.
doi:10.1007/s10979-005-6832-7
- Richards, C. M. (2017). *"You can get prepared, but it's still scary walking out the door": Exploring the role of dynamic re-entry factors in release planning for high-risk offenders*. (Unpublished Master's thesis), Victoria University of Wellington, New Zealand.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*, 1-36.

- Ryan, B. L., Koval, J., Corbett, B., Thind, A., Campbell, M. K., & Stewart, M. (2015). *Assessing the impact of potentially influential observations in weighted logistic regression*. The Research Data Centres Information and Technical Bulletin: Statistics Canada. Retrieved from http://www.isr.umich.edu/src/smp/asda/weighted_influence_logistic_2015.pdf
- Sampson, R. J., & Laub, J. H. (2005). A life-course view of the development of crime. *The Annals of the American Academy of Political and Social Science*, 602, 12-45. doi:10.1177/0002716205280075
- Scanlan, J. (2015). *Assessing women's risk of recidivism: An investigation into the predictive validity of the Dynamic Risk Assessment for Offender Re-entry (DRAOR) with matched samples of community-sentenced women and men*. (Unpublished Master's thesis), Victoria University of Wellington, New Zealand.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69, 239-241.
- Scoones, C., Willis, G., & Grace, R. (2012). Beyond static and dynamic risk factors: The incremental validity of release planning for predicting sex offender recidivism. *Journal of Interpersonal Violence*, 27, 222-238. doi:10.1177/0886260511416472
- Serin, R. C. (2007). *The dynamic risk assessment for offender re-entry (DRAOR). Unpublished user manual*.
- Serin, R. C., Chadwick, N., & Lloyd, C. D. (2016). Dynamic risk and protective factors. *Psychology, Crime & Law*, 22, 151-170. doi:10.1080/1068316X.2015.1112013

- Serin, R. C., Lloyd, C. D., Helmus, L., Derkzen, D. M., & Luong, D. (2013). Does intra-individual change predict offender recidivism? Searching for the Holy Grail in assessing offender change. *Aggression and Violent Behavior, 18*, 32-53. doi:10.1016/j.avb.2012.09.002
- Serin, R. C., Mailloux, D., & Wilson, N. (2012). The dynamic risk assessment for offender re-entry (DRAOR), New Zealand adaptation. *Unpublished user manual*.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.
- Singh, J. P., Desmarais, S. L., Sellers, B. G., Hylton, T., Tirotti, M., & Van Dorn, R. A. (2014). From risk assessment to risk management: Matching interventions to adolescent offenders' strengths and vulnerabilities. *Children and Youth Services Review, 47*, 1-9. doi:10.1016/j.childyouth.2013.09.015
- Smeth, A. H. (2013). *Evaluating risk assessments among sex offenders: A comparative analysis of static and dynamic factors*. (Unpublished Master's dissertation), Carleton University Ottawa, Canada.
- Smith, P., Gendreau, P., & Swartz, K. (2009). Validating the principles of effective intervention: A systematic review of the contributions of meta-analysis in the field of corrections. *Victims and Offenders, 4*, 148-169. doi:10.1080/15564880802612581
- Sutherland, E. H. (1947). *Principles of criminology: A sociological theory of criminal behavior*. New York, NY: JB Lippincott.
- Tabachnick, B. G., & Fidel, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson Education.

- Taxman, F. S., & Caudy, M. S. (2015). Risk tells us who, but not what or how. *Criminology & Public Policy*, 14, 71-103. doi:10.1111/1745-9133.12116
- Thornton, D. (2015). Developing a theory of dynamic risk. *Psychology, Crime & Law*, 22, 138-150. doi:10.1080/1068316x.2015.1109092
- Tonry, M. (2014). Legal and ethical issues in the prediction of recidivism. *Federal Sentencing Reporter*, 26, 167-176. doi:10.1525/fsr.2014.26.3.167
- Trafimow, D. (2015). A defense against the alleged unreliability of difference scores. *Cogent Mathematics*, 2, 1064626. doi:10.1080/23311835.2015.1064626
- Ullrich, S., & Coid, J. (2011). Protective factors for violence among released prisoners--effects over time and interactions with static risk. *Journal of Consulting and Clinical Psychology*, 79, 381-390. doi:10.1037/a0023613
- van den Berg, J. W., Smid, W., Schepers, K., Wever, E., van Beek, D., Janssen, E., & Gijs, L. (2017). The predictive properties of dynamic sex offender risk assessment instruments: A meta-analysis. *Psychological Assessment*, 30, 179-191. doi:10.1037/pas0000454
- Vasiljevic, Z., Berglund, M., Öjehagen, A., Höglund, P., & Andersson, C. (2017). Daily assessment of acute dynamic risk in paroled offenders: Prediction, predictive accuracy and intervention effect. *Psychiatry, Psychology and Law*, 24, 715-729. doi:10.1080/13218719.2017.1308219
- Vieira, T. A., Skilling, T. A., & Peterson-Badali, M. (2009). Matching court-ordered services with treatment needs: Predicting treatment success with young offenders. *Criminal Justice and Behavior*, 36, 385-401. doi:10.1177/0093854808331249

- Viljoen, J. L., Gray, A. L., Shaffer, C., Bhanwer, A., Tafreshi, D., & Douglas, K. S. (2017). Does reassessment of risk improve predictions? A framework and examination of the SAVRY and YLS/CMI. *Psychological Assessment, 29*, 1096-1110. doi:10.1037/pas0000402
- Vitopoulos, N. A., Peterson-Badali, M., & Skilling, T. A. (2012). The relationship between matching service to criminogenic need and recidivism in male and female youth: Examining the RNR principles in practice. *Criminal Justice and Behavior, 39*, 1025-1041. doi:10.1177/0093854812442895
- Vose, B. (2016). Risk assessment and reassessment. *Criminology & Public Policy, 15*, 301-308. doi:10.1111/1745-9133.12196
- Vose, B., Smith, P., & Cullen, F. T. (2013). Predictive validity and the impact of change in total LSI-R score on recidivism. *Criminal Justice and Behavior, 40*, 1383-1396. doi:10.1177/0093854813508916
- Wakeling, H., Beech, A. R., & Freemantle, N. (2013). Investigating treatment change and its relationship to recidivism in a sample of 3773 sex offenders in the UK. *Psychology, Crime & Law, 19*, 233-252. doi:10.1080/1068316X.2011.626413
- Ward, T. (2016). Dynamic risk factors: Scientific kinds or predictive constructs. *Psychology, Crime & Law, 22*, 2-16. doi:10.1080/1068316X.2015.1109094
- Ward, T. (2017). Prediction and agency: The role of protective factors in correctional rehabilitation and desistance. *Aggression and Violent Behavior, 32*, 19-28. doi:10.1016/j.avb.2016.11.012
- Ward, T., & Beech, A. R. (2015). Dynamic risk factors: A theoretical dead-end? *Psychology, Crime & Law, 21*, 100-113. doi:10.1080/1068316X.2014.917854

- Ward, T., & Fortune, C.-A. (2015). From dynamic risk factors to causal processes: a methodological framework. *Psychology, Crime & Law*, 22, 190-202.
doi:10.1080/1068316x.2015.1117080
- Welsh, B. C., Loeber, R., Stevens, B. R., Stouthamer-Loeber, M., Cohen, M. A., & Farrington, D. P. (2008). Costs of juvenile crime in urban areas: A longitudinal perspective. *Youth Violence and Juvenile Justice*, 6, 3-27.
doi:10.1177/1541204007308427
- Western, B., Braga, A. A., Davis, J., & Sirois, C. (2015). Stress and hardship after prison. *American Journal of Sociology*, 120, 1512-1547.
doi:10.1086/681301
- Wooditch, A., Tang, L. L., & Taxman, F. S. (2014). Which criminogenic need changes are most important in promoting desistance from crime and substance use? *Criminal Justice and Behavior*, 41, 276-299.
doi:10.1177/0093854813503543
- Xu, R., & O'Quigley, J. (1999). A R 2 type measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics*, 12, 83-107.
doi:10.1080/10485259908832799
- Yang, M., Guo, B., Olver, M. E., Polaschek, D. L. L., & Wong, S. C. (2017). Assessing associations between changes in risk and subsequent reoffending: an introduction to relevant statistical models. *Criminal Justice and Behavior*, 44, 59-84. doi:10.1177/0093854816678648
- Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, 136, 740-767. doi:10.1037/a0020473

- Yesberg, J. A. (2015). *Exploring mechanisms of change in the rehabilitation of high-risk offenders*. (Unpublished PhD thesis), Victoria University of Wellington, New Zealand.
- Yesberg, J. A., & Polaschek, D. L. L. (2015). Assessing dynamic risk and protective factors in the community: Examining the validity of the Dynamic Risk Assessment for Offender Re-entry. *Psychology, Crime & Law*, 21, 80-99.
doi:10.1080/1068316X.2014.935775
- Zamble, E., & Quinsey, V. L. (1997). *The criminal recidivism process*. Cambridge, UK: Cambridge University Press.