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

by

Caleb D. Lloyd

A thesis submitted to
the Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Psychology
Carleton University
Ottawa, Canada

© 2015 Caleb D. Lloyd

Abstract

Correctional psychologists have classified risk factors for criminal recidivism into static (non-changing, non-improving) and dynamic (theoretically changeable) domains. By definition, a dynamic risk factor is a variable that can change across time, and as it changes, offenders' likelihood of recidivism must logically change in the same direction. Thus, a measure that assesses dynamic risk factors repeatedly across time is hypothesized to provide evidence that more proximal re-assessments have greater predictive validity for short-term reoffending outcomes, compared to prior, more distal assessments. This dissertation develops and proposes a framework for testing the proximity hypothesis with longitudinal, multiple re-assessment re-entry data, and argues that Cox regression survival analysis with time-varying predictors is an appropriate statistical model for comparing the relative value of ratings taken later in the follow-up period, compared to baseline ratings taken at the time of release from incarceration. In this dissertation, the proposed three-step framework was applied to a large dataset of paroled offenders in New Zealand who were assessed by their supervision officers on a regular basis ($N = 3694$ offenders represented by $N = 97,185$ assessments on a measure of theoretically dynamic risk factors). Results demonstrated consistent support for the conclusion that re-assessment improves predictive validity, specifically by showing (a) incremental prediction of re-assessments over baseline scores, and (b) incremental prediction of the most proximal assessment compared to averages of earlier scores. Effect size indices (time-dependent area under the curve [AUC] statistics) temper the interpretation of these results by showing that predictive

accuracy was only sometimes enhanced to a statistically significant degree when including re-assessments in the model. Still, the results encourage community supervision correctional agencies to consider employing repeated measurement of offender clients using a dynamic risk instrument. It is also recommended that researchers consider using the proposed framework as a logical approach to verifying the underlying assumption that purportedly dynamic risk factors are both dynamic, and accurate markers of fluctuations in individuals' risk to re-offend.

Table of Contents

Abstract.....	ii
Table of Contents.....	iv
List of Tables.....	ix
List of Figures.....	xv
Introduction.....	1
Chapter 1: Theory, Literature, and Hypothesis.....	2
Dynamic Risk Factors are Important Within Theories of Criminal Behavior.....	5
Comprehensive Theories of Criminal Behavior.....	7
Pro-Crime Cognitions: The Primary Antecedent of Crime.....	11
A Phenomenological Approach to Dynamic Risk.....	13
Dynamic Risk Factors are <i>Theoretically</i> Causes of Criminal Behavior.....	15
Dynamic Risk Factors are Important for Intervention.....	16
Dynamic Risk Factors are the Primary Focus Within	
Offender Management.....	17
Meta-reviews of Intra-Individual Change During Rehabilitation.....	19
Future Directions for Dynamic Risk Research in Rehabilitation.....	22
Dynamic Risk Factors <i>May</i> Be Important for Prediction.....	25
Dissimilarities Between Rehabilitation and	
Risk Assessment Research.....	26
More Community Based Dynamic Risk Research is Necessary.....	29
Is there Evidence that Re-assessments Enhance Prediction?.....	30
Single-occasion measurement strategies.....	30

Segmenting the follow-up.....	32
Comparing post- versus pre-program assessments.....	33
Retrospective matched designs.....	35
Prospective designs with re-assessment data.....	38
Lagged prediction models.....	39
Survival analysis.....	40
Averaging vs. proximity.....	41
Conclusion.....	43
Summary & Conclusion.....	44
Research Question & Hypothesis.....	46
Chapter 2: Choosing the Analytic Method.....	57
Choice of Statistical Models.....	59
Multilevel Modeling.....	61
Growth Curve Analyses.....	64
Generalized Estimating Equations.....	67
Latent Class Trajectory Modeling.....	70
Growth Trajectory Mixture Survival Analysis.....	74
Cox Regression Survival Analysis with Time-varying Predictors.....	75
Modeling re-assessments, rather than modeling change.....	76
Model advantages.....	78
Addressing the timing of assessments.....	79
Examining risk associations, rather than conducting risk assessment.....	83

Effect size estimates.....	90
Measures analogous to R^2	93
Time-dependent measure of prediction accuracy.....	95
Summary.....	98
Chapter 3: Methodology & Dataset Preparation.....	99
Method.....	100
Participants.....	100
Measures.....	102
Recidivism.....	102
Static risk.....	105
Dynamic risk and protective factors.....	106
Demographic information.....	107
Psychometric Research on DRAOR Data.....	108
Factor structure and measurement invariance.....	108
Predicting future recidivism outcomes.....	113
Detecting change across time.....	115
Summary.....	116
Structuring the Data for Analysis.....	117
Defining the unit of analysis.....	117
Defining units of discrete time.....	120
Defining baseline.....	121
Choosing which assessment would define the discrete period....	122
Truncating the follow-up period.....	124

Lagging recidivism information to ensure prospective prediction.....	124
Truncating sequences due to long gaps between assessments....	124
Imputing assessments during weeks when no assessments occurred.....	126
Summary.....	128
Limitations.....	128
Data nesting within supervision officers.....	129
Differences in fidelity across supervision officers.....	129
Unknown inter-rater reliability between supervision officers.....	132
Summary.....	136
Strengths.....	137
Chapter 4: Basic Model Building.....	138
Descriptive Statistics.....	139
Life Table.....	139
Central Tendency and Imputation Strategies.....	147
Subscale Correlations and Internal Consistency.....	153
Change Across Time, Intraclass Correlations, and Stability Coefficients.....	158
Testing Model Assumptions.....	166
Assumptions of General Regression Models.....	166
Normality.....	167
Homoscedasticity and linearity.....	172
Multivariate outliers.....	172

Multicollinearity.....	175
Assumptions and Decisions Specific to Cox Regression Models.....	175
Equivalence between censored and remaining cases.....	175
Choosing a representation of time.....	180
Proportionality of hazards.....	183
Model Building.....	195
Final Model Checks.....	200
Potential Non-linear Effects.....	203
Presence of Highly Influential Cases.....	205
Unobserved Heterogeneity.....	206
Chapter 5: Testing Prediction with Proximal vs. Distal Assessments.....	215
Baseline versus Proximal Assessments.....	215
Averaging versus Proximal Assessments.....	229
Different Averaging Strategies versus Proximity Only.....	233
Multivariate Prediction Models.....	241
Chapter 6: Discussion.....	250
Possible Explanations for the Pattern of Results.....	253
Limitations & Future Directions.....	261
Recommendations.....	266
References.....	269
Appendix.....	304

List of Tables

Table 1.1. <i>Comprehensive Theories that Explain Both Crime Onset and Criminal Actions.....</i>	8
Table 1.2. <i>Publications Reporting Retrospective Matched Designs or Prospective Multi-Assessment Designs to Assess the Relationship Between Dynamic Risk Factors and Criminal Recidivism.....</i>	47
Table 1.3. <i>Dynamic Risk Factors and Risk Instruments Tested for Change and Predictive Validity at Re-assessment in the Community.....</i>	50
Table 2.1. <i>Longitudinal Statistical Models and the Features They Can Accommodate.....</i>	60
Table 3.1. <i>Psychometric Properties (Cronbach's Alpha and Pearson Correlations) of DRAOR Data Reported in Past Publications.....</i>	109
Table 3.2. <i>Associations Between DRAOR Scores and Recidivism Outcomes (Point Biserial Correlations and Area Under the Curve) Reported in Prior Publications.....</i>	114
Table 3.3. <i>Logical Strategies for Choosing Among Multiple Assessments in One Discrete Time Period and Imputing Scores in Missing Time Periods (as Recommended by Singer & Willett, 2003).....</i>	127
Table 4.1. <i>Life Table Describing the Number of Weeks Offenders Remained in the Community Following Release From Incarceration During</i>	

<i>a One-year Follow-up Period.....</i>	<i>140</i>
Table 4.2. <i>Descriptive Statistics (Means and Standard Deviations) for DRAOR</i>	
<i>Subscales Displayed for Each Week of the One-year Follow-up Period, With</i>	
<i>Percentage of Non-imputed Values.....</i>	<i>148</i>
Table 4.3. <i>Model Fit Indices For Cox Regression Time-Varying Predictor Models Using</i>	
<i>DRAOR Subscales to Predict Time to Recidivism, Comparing Six Dataset</i>	
<i>Building Strategies.....</i>	<i>152</i>
Table 4.4. <i>Psychometric Properties (Internal Consistency, Inter-subscale Correlations,</i>	
<i>and Correlations with Static Risk) of DRAOR Scores, Displayed for Each Week of</i>	
<i>the One-year Follow-up Period.....</i>	<i>154</i>
Table 4.5. <i>Proportion of Change Across DRAOR Subscale Assessments, Displayed For</i>	
<i>Each Week-to-Week Transition During the One-year Follow-up Period.....</i>	<i>159</i>
Table 4.6. <i>Descriptive Statistics (Observed Range, Baseline Mean, Mean Change, and</i>	
<i>Mean Net Change) Across DRAOR Assessments During the One-year Follow-up</i>	
<i>Period.....</i>	<i>163</i>
Table 4.7. <i>Statistics for Diagnosing Distribution Normality of DRAOR Subscale Scores</i>	
<i>(Skewness Ratios, Kurtosis Ratios, and Number of Extreme Outliers), Displayed</i>	
<i>for Each Week of the One-year Follow-up Period.....</i>	<i>168</i>
Table 4.8. <i>Independent Samples t-Tests Comparing the Static Risk and DRAOR Scores</i>	
<i>of Censored (During a Prior Week) and Uncensored Non-recidivists (From a</i>	
<i>Subsequent Week), Displayed for Each Week-to-Week Comparison During the</i>	
<i>One-year Follow-up Period.....</i>	<i>176</i>

Table 4.9. <i>Comparison of Alternative Smooth Polynomial Representations for the Main Effect of Time in a Baseline Discrete-time Hazard Model</i>	181
Table 4.10. <i>Results for Fitting Three Discrete-time Hazard Models Predicting Criminal Recidivism Using DRAOR Subscales, Including Interaction Terms Representing Score*Time</i>	192
Table 4.11. <i>Results of a Discrete-time Hazard Model Using Static Risk (RoC*RoI) to Predict Survival Time to Criminal Recidivism</i>	197
Table 4.12. <i>Results of a Discrete-time Hazard Model Using DRAOR Stable Subscale to Predict Survival Time to Criminal Recidivism</i>	198
Table 4.13. <i>Results of a Discrete-time Hazard Model Using DRAOR Acute Subscale to Predict Survival Time to Criminal Recidivism</i>	201
Table 4.14. <i>Results of a Discrete-time Hazard Model Using DRAOR Protect Subscale to Predict Survival Time to Criminal Recidivism</i>	202
Table 4.15. <i>Results of Fitting Four Discrete-time Hazard Models Predicting Criminal Recidivism Using DRAOR Subscales, Including Quadratic Predictor Terms</i>	204
Table 4.16. <i>Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and Static Risk (RoC*RoI) Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, as Estimated Through Sensitivity Analyses (Lin et al., 1998)</i>	209
Table 4.17. <i>Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and DRAOR Stable Scores Under the Assumption</i>	

<i>of Varying Degrees of Unobserved Heterogeneity, as Estimated Through Sensitivity Analyses (Lin et al., 1998).....</i>	212
Table 4.18. <i>Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and DRAOR Acute Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, as Estimated Through Sensitivity Analyses (Lin et al., 1998).....</i>	213
Table 4.19. <i>Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and DRAOR Protect Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, as Estimated Through Sensitivity Analyses (Lin et al., 1998).....</i>	214
Table 5.1. <i>Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Stable Subscale Score.....</i>	218
Table 5.2. <i>Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Acute Subscale Score.....</i>	223
Table 5.3. <i>Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Protect Subscale Score.....</i>	225
Table 5.4. <i>Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Stable Subscale Score, Including a Rolling Mean.....</i>	231

Table 5.5. <i>Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Acute Subscale Score, Including a Rolling Mean.....</i>	234
Table 5.6. <i>Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Protect Subscale Score, Including a Rolling Mean.....</i>	236
Table 5.7. <i>Model Fit (Akaike Information Criteria) and Effect Size (R-Squared, and Concordance) Indices for Cox Regression Models with Time-varying Predictors Using Different Specifications of DRAOR Subscales as Predictors, Including Differing Averaging Strategies.....</i>	240
Table 5.8. <i>Results of Fitting Three Multivariate Discrete-time Hazard Models to Predict Time to Criminal Recidivism, by Entering DRAOR Baseline or Dynamic Subscale Scores Following Static Risk Scores (RoC*RoI).....</i>	242
Table 5.9 <i>Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and Static Risk (RoC*RoI) Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, Estimated from the Final Multivariate Model.....</i>	247
Table 5.10. <i>Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and DRAOR Acute Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, Estimated from the Final Multivariate Model.....</i>	248
Table 5.11. <i>Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and DRAOR Protect Scores Under the Assumption</i>	

<i>of Varying Degrees of Unobserved Heterogeneity, Estimated from the Final</i>	
<i>Multivariate Model.....</i>	249

List of Figures

<i>Figure 3.1.</i> Participant flowchart describing inclusion criteria for analysis.....	118
<i>Figure 4.1.</i> Observed hazard rate at each week of the follow up, with a regression line of best fit superimposed onto the graph.....	145
<i>Figure 4.2.</i> Observed cumulative proportion of non-recidivists across the one-year follow-up period.....	146
<i>Figure 4.3.</i> Logit-transformed observed hazard rates at each week of the one-year follow-up period, with linear and quadratic regression lines of best fit superimposed onto the graph.....	182
<i>Figure 4.4.</i> Logit-transformed observed hazard rates at each week of the one-year follow-up period, separating the sample by a median split on RoC*RoI scores.....	184
<i>Figure 4.5.</i> Scatterplot of Schoenfeld residuals resulting from a Cox regression model with static risk (RoC*RoI) as the single predictor, with line of best fit superimposed onto graph.....	186
<i>Figure 4.6.</i> Logit-transformed observed hazard rates at each week of the one-year follow-up period for scores that represent ± 1 SD on DRAOR <i>stable</i> subscale.....	187
<i>Figure 4.7.</i> Logit-transformed observed hazard rates at each week of the one-year follow-up period for scores that represent ± 1 SD on DRAOR <i>acute</i> subscale.....	188
<i>Figure 4.8.</i> Logit-transformed observed hazard rates at each week of the	

one-year follow-up period for scores that represent ± 1 SD on	
DRAOR <i>protect</i> subscale.....	189
<i>Figure 4.9.</i> Scatterplot of Schoenfeld residuals resulting from a Cox regression	
model with DRAOR <i>stable</i> subscale as the single predictor, with line of best fit	
superimposed onto graph.....	191
<i>Figure 4.10.</i> Scatterplot of Schoenfeld residuals resulting from a Cox regression	
model with DRAOR <i>acute</i> subscale as the single predictor, with line of best fit	
superimposed onto graph.....	193
<i>Figure 4.11.</i> Scatterplot of Schoenfeld residuals resulting from a Cox regression	
model with DRAOR <i>protect</i> subscale as the single predictor, with line of best	
fit superimposed onto graph.....	194
 <i>Figure 5.1.</i> A single participant's observed values on the DRAOR <i>stable</i> subscale	
across time.....	216
<i>Figure 5.2.</i> Time dependent Area Under the Curve (AUC) statistics graphed across	
the one-year follow-up period for baseline DRAOR <i>stable</i> scores and dynamic	
DRAOR <i>stable</i> scores.....	221
<i>Figure 5.3.</i> Bootstrapped 95% confidence intervals overlaid onto Figure 5.2...	
221	
<i>Figure 5.4.</i> Time dependent Area Under the Curve (AUC) statistics graphed across	
the one-year follow-up period for baseline DRAOR <i>acute</i> scores and dynamic	
DRAOR <i>acute</i> scores.....	227
<i>Figure 5.5.</i> Bootstrapped 95% confidence intervals overlaid onto Figure 5.4...	
227	

<i>Figure 5.6.</i> Time dependent Area Under the Curve (AUC) statistics graphed across the one-year follow-up period for baseline DRAOR <i>protect</i> scores and dynamic DRAOR <i>protect</i> scores.....	228
<i>Figure 5.7.</i> Bootstrapped 95% confidence intervals overlaid onto Figure 5.6....	228
<i>Figure 5.8.</i> A single participant's observed values on the DRAOR <i>stable</i> subscale across time, with a depiction of a rolling mean score overlaid onto the graph.....	230
<i>Figure 5.9.</i> A single participant's observed values on the DRAOR <i>stable</i> subscale across time, with a depiction of a moving rolling mean score overlaid onto the graph.....	239
<i>Figure 5.10.</i> Time dependent Area Under the Curve (AUC) statistics graphed across the one-year follow-up period for static risk scores (RoC*RoI) and baseline DRAOR subscale scores, combined.....	244
<i>Figure 5.11.</i> Time dependent Area Under the Curve (AUC) statistics graphed across the one-year follow-up period for static risk scores (RoC*RoI) and dynamic DRAOR subscale scores, combined.....	246
<i>Figure 5.12.</i> Bootstrapped 95% CIs overlaid onto Figure 5.11.....	246

Introduction

Two primary goals are addressed in this dissertation. The first goal is to describe a statistical model that has been recently gaining exposure within correctional psychology. *Cox regression survival analysis with time-varying predictors* has important advantages when applied to research questions within corrections, and is uniquely suited to simultaneously address two objectives within the same model: (a) determine which variables predict recidivism following release from incarceration, and (b) model re-assessments of an offender's features and circumstances as he or she returns to and remains within the community. Thus, the first goal of this dissertation is to demonstrate the application of this statistical model with corrections re-entry data.

Second, this dissertation describes a strategy for testing the concept of *dynamic risk*. The term *risk factor* refers to demographic, situational, or intra-individual characteristics that are statistically associated with criminal recidivism, and *dynamic risk factors* are variables that can theoretically change across time. By definition, as a dynamic risk factor varies across time, the offender's likelihood of criminal behavior should vary along with it. Thus, a risk assessment instrument that measures dynamic risk variables *in a dynamic way across time* is hypothesized to show "real-time" predictive validity for short-term recidivism. This dissertation describes tests of this hypothesis, and describes how a Cox regression model can be used to examine whether more temporally proximal ratings enhance prediction compared to more distal ratings.

This dissertation is organized to allow readers to focus their attention on the goals relevant to their interests. Chapter 1 reviews prior literature on dynamic risk, and describes the theoretical reasons why repeatedly measuring dynamic variables should enhance decision-making within correctional supervision. Chapter 2 addresses the statistical modeling of re-assessment data by (a) describing a researcher's choice of statistical models when attempting to analyze re-assessment data, and (b) introducing the features of the Cox regression model with time-varying predictors. Chapter 3 describes the dataset used within this dissertation, as well as the measurement and procedures of the study from which the dataset was drawn.

The remaining chapters address whether it is meaningful for an agency to continually re-assess offenders with a purportedly dynamic risk measure. Chapter 4 describes the initial descriptive results of the re-assessment data, and presents the basic prediction models using Cox regression with time-varying predictors. Chapter 5 specifically addresses model-building strategies for testing the hypothesis that distal assessments (i.e., initial baseline scores) will demonstrate weaker predictive validity compared to models that include more proximal assessments (i.e., re-assessments).

In conclusion, Chapter 6 provides a general discussion of the results drawn from the analyses, comments on the implications of the findings, and proposes some directions for future research.

Chapter 1: Theory, Literature, and Hypothesis

Variables that are statistically associated with future criminal offending are termed *risk factors*, and the assessment of an individual's risk factors plays a critical

role in all aspects of correctional practice. Structured consideration of risk factors enhances offender management, and represents current best correctional practice (Ægisdóttir et al., 2006; Andrews, Bonta, & Wormith, 2006; Hanson, 2009; Smith, Gendreau, & Swartz, 2009). Indicators of an offender's past criminal history, current age, and psychosocial variables statistically related to future crime should be considered during correctional decision-making, including estimating risk of recidivism following release, allocating rehabilitation and supervision resources, and defining targets for programs (Andrews & Bonta, 2010).

Risk factors have been classified into *static* (historical, non-improving) and *dynamic* (theoretically changeable) domains (Bonta, 1996; Zamble & Quinsey, 1997). By definition, a dynamic risk factor should logically exhibit at least two features. First, dynamic factors should show change across time. Second, due to this change, assessments of dynamic factors that are proximal in time to criminal events should be more closely associated with criminal behavior than assessments that are comparatively distal. The distinction between *static* and *dynamic* specifies the potential for change, whereas Hanson and Harris (2000; Hanson, Harris, Scott, & Helmus, 2007) further categorized dynamic factors into *stable* and *acute* domains. This distinction recognizes that some dynamic risk factors should be conceptualized as relatively enduring traits (*stable*), while others may arise and pass in short time frames, immediately prior to the commission of a crime (*acute*). These sub-categories specify the temporal proximity of the change in relation to observed criminal behavior.

Thus, risk factors are currently divided into a three-tiered conceptual system based on both the rapidity of potential change, and proximity to criminal action. Static risk factors are the most distal and unchanging variables, and are only useful for classifying an offender's degree of risk for future criminal behavior. Static risk factors indicate *risk status*. Stable dynamic risk factors are changeable, but generally slow to change, and typically are best characterized as clinical targets within rehabilitation programs. Acute dynamic risk factors are highly proximal to criminal action and signal the imminent circumstances, behaviors, and cognitions that are related to an offender taking criminal action. For this reason, acute factors indicate *risk state*.

This conceptual system corresponds well with existing and recently emerging knowledge of criminal behavior, but the evidence base to support this framework is currently weak. This chapter broadly reviews correctional research on dynamic risk, and argues that new research on dynamic risk factors has potential to impact at least three aspects of the field and correctional practice, by contributing to (a) theory building, (b) intervention (offender management and rehabilitation), and (c) prediction. This chapter concludes that there is limited literature examining whether hypothesized dynamic risk factors change across time, yet there is even less research examining whether re-assessments of dynamic risk factors improve prediction beyond prior assessments. Thus, the hypothesis that assessment proximity enhances prediction represents an important question for correctional practice. In addition, new research examining this hypothesis would contribute to the field's theoretical understanding of criminal behavior.

Dynamic Risk Factors are Important Within Theories of Criminal Behavior

Farrington (2011) argued that an adequate explanation of criminal behavior answers two related, but distinct questions. The first question addresses the onset of criminal behavior (i.e., “Why do some individuals offend?”). Answers to this question require focusing on variables that differentiate individuals from each other. The second question concerns within-individual differences regarding why crime recurs (i.e., “Why do offenders commit an offence?”). Thus, theories of criminal behavior should aim to address two levels of understanding: (a) the relatively stable individual differences that distinguish offenders from non-offenders, and (b) the changing circumstances that influence an individual to choose to take criminal action in a specific situation (while foregoing criminal action in another).

This distinction between levels of explanation parallels the conceptual hierarchy of static, stable, and acute risk factors. Specifically, static risk factors are largely historical variables that describe the background features related to the onset and emerging patterns of criminal behavior. By contrast, dynamic risk factors are more likely to be the features that cause criminal actions because they logically vary alongside an individual offender’s personal pattern of criminal behavior, and are subjectively salient to the individual at the moment of criminal involvement.

However, risk assessment has developed around the central concern of maximizing prediction, and because of this, the identification of risk factors has leaned more heavily toward finding static predictors that are practical to measure (between-individual features), without as much attention toward *understanding*

criminal behavior through within-individual features (Skeem & Monahan, 2011; Ward & Beech, 2014). Thus, less is known about dynamic factors, and there is room to develop a more psychological understanding of how risk (or propensity toward criminal behavior) manifests itself in offenders' lives, defined by Hanson (2009, p. 177) and Mann, Hanson, and Thornton (2010) as the identification of "psychologically meaningful risk factors". In contrast to static factors, dynamic factors are more likely to have inherent psychological meaning, which gives dynamic factors some predominance when theorizing the causes of criminal behaviors.

Farrington's (2011) separation of between- and within-individual explanations and Mann and colleagues' (2010) emphasis on the phenomenological experience of committing a crime suggest theory should attempt to describe *how* risk factors might causally impact behavior. Risk variables should not be automatically assumed to be causal simply because they share an association with criminal behavior, and many risk factors are better conceptualized as markers of an underlying (actually causal) propensity (Polaschek, 2012; Ward & Beech, 2014).

Specifically, it is unclear how most known risk variables manifest themselves to propel a criminal action. For example, *younger age at first arrest* is a statistical risk factor of future crime, but how this fact translates into a psychological phenomenon that impacts an individual's decision to commit a new crime is obscure. Early habit formation or foundational developmental experiences may be relevant elements within this risk-crime connection, but on its own, age at first offence is an atheoretical predictor. Some dynamic risk variables, such as substance misuse or

antisocial attitudes, offer more conceptually accessible causal pathways to crime, but a risk variable on its own cannot be considered synonymous with the psychological phenomena directly related to criminal action, in the absence of further conceptual development (Ward & Beech, 2014).

Comprehensive Theories of Criminal Behavior

Few theorists have attempted to specify complete frameworks of general criminal behavior that include causal explanations of both stable individual differences and immediate decisions to commit crime (van der Laan, Blom, & Kleemans, 2009). Yet, three sets of authors have built remarkably similar explanatory theories, despite writing from different theoretical traditions and drawing data from distinct populations. These theories are referenced and summarized in Table 1.1. Note that the rows within Table 1.1 differentiate levels of explanation, by categorizing risk factors based on their proximity to criminal action.

Andrews and Bonta (2010) specifically drew upon learning and cognitive theories from psychology to build their corrections-based model of criminal behavior. As described in their Personal, Interpersonal, and Community-Reinforcement (PIC-R) perspective, they argue that the immediate antecedent to criminal action is the offender's perception that the benefits of crime outweigh the costs (see Column 2 in Table 1.1). In other words, crime results when an individual defines a particular situation as likely to reward the commission of criminal behavior. The model suggests perceptions are derived from both the immediate situation (i.e., any situational cues that suggest crime would be a legitimate course of action), and the individual's personal characteristics (attitudes, self-efficacy, and

Table 1.1

Comprehensive Theories that Explain Both Crime Onset and Criminal Actions

Level of Proximity to Crime Event	Theories of Criminal Behavior		
	Personal, Interpersonal, and Community-Reinforcement Perspective (PIC-R; Andrews & Bonta, 2010)	Integrated Cognitive Antisocial Potential Theory (ICAP; Farrington, 2011)	Social Schematic Theory (SST; Simons & Burt, 2011)
Background Variables that Influence General Exposure to Crime-Supportive Values	Characteristics of family of origin, personal temperament, neighborhood features, socialization	Impulsivity, social strain, interpersonal modeling, socialization	Community disorganization, harsh parenting, deviant peers, racial discrimination
Mediating Factors that Influence Encountering Perceived Criminal Opportunities	Cognitive features favorable toward crime, situational factors, criminal associates	Motivation, situational factors	Criminogenic knowledge schemas, specifically hostile view of relationships, concern with immediate gratification, and cynical view of conduct norms
Immediate Proximal Causes During the Moment of Criminal Action	Definitions favorable to criminal behavior, perceived costs and benefits of crime, features of the setting	Perceived cost and benefits of crime, perceived probability of detection, cognitive repertoires of behavior, decision-making processes favorable to choosing crime	Situational interpretations that legitimize or compel criminal behavior

peers). In addition, Andrews and Bonta argued that variables such as personal temperament, family of origin, neighborhood, and childhood experiences are important for shaping an individual's attitudes and current peer groups, which in turn directly impact decisions regarding crime. This perspective draws on and combines a variety of empirical research, and borrows specifically from Theory of Reasoned Action (Ajzen & Fishbein, 1969) and its more recent extension, Theory of Planned Behavior (TPB; Ajzen, 1991; for meta-analyses, see Conner & Armitage, 1998; Hagger, Chatzisarantis, & Biddle, 2002; and, see Andrews & Bonta, 2010, for more detail on the theories that influenced the PIC-R perspective).

Although Farrington (2011) and Simons and Burt (2011) drew upon traditional criminology theories to build their comprehensive theories of criminal behavior, the similarities with Andrews and Bonta's (2010) PIC-R perspective are much more pronounced than the dissimilarities. Of note, all three frameworks emphasize offender decision-making in the moment of criminal action, and identify aspects of offender cognition as the core proximal antecedents to decisions to commit crime. Broadly, all three frameworks suggest offenders view criminal behavior as favorable or legitimate. Specifically, Andrews and Bonta (2010) defined offender cognition as antisocial attitudes, criminal self-efficacy, and perceived social approval for crime, Farrington (2011) as motivational factors, including beliefs about what will be gained from committing a crime, and Simons and Burt (2011) as social schemas, or simplified cognitive shortcuts that summarize the nature of past social interactions. In addition, all three theories argue that criminal action results

from the interaction of offender cognition and features of the situation, placing importance on the context of the immediate environment.

Of the three theories, Simons and Burt's (2011) Social Schematic Theory (SST) is arguably the most explicit about which cognitive elements mediate the effect of offenders' background characteristics on their immediate decisions to commit crime. SST focuses on three social schemas as mechanisms that translate various disadvantaging circumstances into perceptions that criminal action is beneficial. Specifically, these schemas are (a) hostile views of relationships, (b) discounting the value of delaying gratification until the future, and (c) cynical views of conventional norms. In particular, SST hypothesizes that these schemas emerge as rational responses to childhood environments that were saturated with criminal models, weak social ties, and overt exploitation. Once these schemas have developed, SST views criminal behavior as a rational response to situations that individuals will inevitably interpret (or misinterpret) as (a) an encounter with someone who is likely exploit them, (b) an opportunity to claim a benefit that is doubtful to exist in the future, or (c) a trivial violation against a community that has failed to preserve equality and moral authority.

Simons, Burt, and colleagues (i.e., Simons & Burt, 2011; Simons, Burt, Barr, Lei, & Stewart, 2014; Simons & Barr, 2014) have gathered empirical support for their framework using longitudinal data collected with community-based juveniles measured repeatedly until young adulthood. Their results have supported their framework, such that changes in background characteristics (parenting practices, community features) were positively related to the development of the three

hypothesized schemas (Simons & Burt, 2011). In turn, the schemas were related to interpretations of situations (perceived provocations and opportunities), and these interpretations were directly related to criminal delinquency (Simons et al., 2014). Further, Simons and Barr (2014) demonstrated that delinquents who engaged in high-quality romantic relationships in young adulthood simultaneously reduced their criminal behavior, and that simultaneous changes in social schemas could account for approximately one-third of the partnership-crime reduction effect. In summary, this set of research suggests that background features indirectly influence all stages of criminal involvement (onset, repeat offending, and cessation) through social schemas (pro-crime perspectives), which directly impact criminal action, if they are accessible to the individual in an immediate situation.

Pro-Crime Cognitions: The Primary Antecedent of Crime

This evidence lends support to the conclusion that dynamic risk factors in the form of offender cognitions may be among the explicit mechanisms, or antecedents, that cause criminal actions. Frameworks such as Andrews and Bonta's (2010) PIC-R perspective and Simons and Burt's (2011) social schematic theory encourage hypotheses about how background features influence general beliefs about crime, which then translate into perceptions of situations that result in decisions to commit crime. Similarly, situational action theory (Wikström & Treiber, 2009; Wikström, Ceccato, Hardie, & Treiber, 2010) hypothesizes that individuals (with varying propensities) and situations (with varying opportunities for crime) interact to propel individuals' choices to commit crime. Similar to the frameworks listed in Table 1.1, situational action theory considers the main individual-level antecedents

of crime to include beliefs about the value of crime (morality), as well as ability (or willingness) to control oneself in the situation.

Thus, all of these frameworks emphasize offender perception, and view static risk factors as markers that point to the presence of pro-crime beliefs. In other words, an individual's background features (static risk factors) derive their power to predict future criminal behavior by signaling a higher likelihood that an offender's immediate social and cognitive features will support the commission of a crime (Simons & Burt, 2011). Walters (2015) reported additional evidence for this conclusion, using longitudinal data to demonstrate that cognitive features (discounting of the future, and preference for risk-taking) accounted for variance between past criminal behavior and future criminal behavior, in part. Despite differences in samples, measurement, and methodology, multiple authors appear to be narrowing in on explanations of criminal action that focus on pro-crime cognitions as a primary motivating dynamic risk factor.

Although recent empirical research and conceptual writings have advanced the theoretical frameworks described in Table 1.1, current theory arguably falls short of Farrington's (2011) objectives, remaining under-developed and lacking explicit evidence. This is particularly the case regarding the field's understanding of the psychological, subjective experience of committing a crime, and how immediate decisions link to known risk factors (Yang & Mulvey, 2012). Researchers could assist new advancements by further (a) identifying the nature of pro-crime cognitions, (b) specifying the mechanisms whereby general cognitions develop and directly influence offenders' decisions to commit crime, and (c) creating explicit

measurement of these cognitive features. At present, measurement of these variables is often indirect, and the link between cognitions and behavior needs further conceptual development.

A Phenomenological Approach to Dynamic Risk

Theorizing psychological explanations of crime requires taking subjective experience into account, or in other words, relying on a phenomenological approach of human decision-making. There are precedents for this within both general psychology and correctional psychology. For example, Szasz (1961) rejected explanations of human behavior that emphasized general propensities or abstract classifications, preferring instead a humanistic approach that examined individuals' decision-making processes and their perceived goals. Szasz preferred the metaphor of a "game", arguing that individuals make behavioral choices consistent with the social games they believe they are playing, their perceptions of what it means to "win", and the "rules" they believe constrain their choices. In other words, the psychological meaning of a decision is closely tied to questions such as (a) "What outcome am I seeking within this interaction?", (b) "What options are available to me within this situation?", and (c) "What are the means whereby I am able to attain my goal?"

Specific to criminal behavior, a phenomenological approach assumes that individuals enter situations with their own understanding of the options, roles, and goals that determine whether criminal behavior should be enacted. The questions posed above de-prioritize the typical focus on individuals' propensity toward antisocial conduct, instead using the framework of the "game" to understand why

individuals perceive opportunities as necessitating criminal behavior. Consistent with Szasz (1961), Simons and Burt (2011) described their social schemas as rules of a game. They also suggested that schemas are “lessons” that offenders learn about the nature and meaning of social interactions (Simons et al., 2014, p. 656). Similarly, Ward and colleagues’ (e.g., Ward & Marshall, 2004; Ward & Gannon, 2006) Good Lives Model (GLM) theorizes criminal behavior in terms of goals and rules. Specifically, GLM uses the concept of *primary goods* to describe the psychological outcome offenders seek to gain from committing a crime, encapsulated in the question: “What human value is the offender seeking to attain through criminal action?” GLM defines *secondary goods* as the behaviors, or the means, through which the offender attempts to attain what he or she values, which can be described as the “rules” the individual follows to achieve his or her goals (see also Barnao & Ward, 2015). Rather than explaining behavior in terms of abstract risk potential, these metaphors focus on individuals’ understanding of their goals, personal roles, and perceived options.

Thus, correctional psychology has paid some attention to building frameworks that attempt to describe and explain criminal behavior in terms of psychology-informed, in-the-moment decisions, rather than distal background influences. Szasz’s (1961) preference for a phenomenological approach over a propensity approach parallels the current dissatisfaction within theories of criminal behavior that emphasize background risk over a person-based understanding of criminal action. A purely psychological, phenomenological approach must take into

account the context where the behavior is enacted, and how the actor subjectively perceives the situation (in terms of roles, goals, and means).

Yet, the practical need within corrections to regularly and easily assess risk for recidivism has resulted in a more detailed understanding of the historical markers that indicate potential for crime, and a less detailed understanding of the direct causal influences on criminal action. Making the link between practice and theory largely rests on advancing the field's understanding of the inter-connections among static risk factors, dynamic risk factors, and crime. Dynamic risk factors are a critical piece within a full theoretical understanding of crime; this means research must attend to the features that occur at the time of criminal action, by measuring proximal, dynamic factors that may influence an offender's immediate decision to commit a crime. Thus, there is danger that theories of criminal action will remain stagnant and under-developed until there is more research that *dynamically* measures dynamic risk factors, with measurements taken a short time prior to criminal actions.

Dynamic Risk Factors are *Theoretically* Causes of Criminal Behavior

Some researchers have framed their search for empirically validated dynamic risk factors as research that establishes the *causes* of criminal behavior (e.g., Howard & Dixon, 2013). Although it is an important endeavor to demonstrate that longitudinal changes in dynamic risk factors either precede or vary alongside changes in criminal behavior, it is largely incorrect to suggest such evidence alone demonstrates causality (Ward & Beech, 2014). In particular, because psychosocial or environmental variables logically and ethically cannot be randomly assigned to

individuals, questions of causality are technically beyond the reach of correctional psychology. The closest approximation involves actively manipulating hypothesized dynamic risk factors (i.e., through rehabilitation), and demonstrating that both (a) change has occurred at the individual level due to the rehabilitation, and that (b) changes in dynamic risk factors are associated with reductions in criminal behavior (Douglas & Skeem, 2005; Kraemer et al., 1997). However, as described in the next section, there is a meager amount of such evidence in the rehabilitation literature.

In the absence of the ability to experimentally manipulate risk variables, observing longitudinal change can help specify which immediate features and circumstances are associated with crime behaviors. Tracking which characteristics change as individuals move from a non-crime state to a state of crime is an important element for describing what may have influenced an immediate decision to commit a crime. However, terminology should emphasize the *association* between risk variables and criminal behavior, even as researchers attempt to build evidence for theoretical frameworks that seek to explain causes of criminal behavior.

Dynamic Risk Factors are Important for Intervention

Classifying risk variables into *static* and *dynamic* domains largely began for the purpose of constructing risk assessments that would be comprehensive in their ability to both predict recidivism and define appropriate targets for intervention (Bonta, 1996). Combining these two purposes recognizes that (a) correctional agencies should seek to reduce offenders' potential for crime in addition to measuring it, and (b) static, atheoretical predictor variables are unhelpful for case management. As such, within correctional practice, dynamic risk factors are

necessarily at the forefront of any practitioner's decisions, and understanding dynamic, intra-individual change is critical for case management. For example, correctional agencies are asked to comment on whether a particular offender has experienced improvements, and this requires some clinical assessment of acute, evolving, ongoing risk status. In other words, intervention without attention to dynamic factors reduces clinicians to being another mechanism of control within the criminal justice system, rather than agents of change (Philipse, Koeter, van der Staak, & van den Brink, 2006). However, current research is limited in its ability to provide adequate direction to practitioners about continuous offender management, because little is known about which risk factors change and which dynamic changes are necessary for risk reduction (Coid et al., 2011; Day, 2013).

Dynamic Risk Factors are the Primary Focus Within Offender Management

Rehabilitation is a core goal within corrections, and even in the absence of explicit evidence, any attempt to manage offenders must begin with identifying a dynamic risk factor to target (i.e., a variable presumed or known to be associated with a reduction in crime, given its ability to predict crime, and ability to change). Consistent with existing terminology at the time they proposed their rehabilitation framework, Andrews and Bonta's (2010) original work used the term *criminogenic need* to define dynamic risk factors targeted in programs. Some writers appear to misunderstand that *criminogenic need* and *dynamic risk factor* are, and ought to be, synonymous concepts (e.g., see Baird, 2009; Caudy, Durso, & Taxman, 2013; Gottfredson & Moriarty, 2006; Taxman & Caudy, 2015, as examples of authors who seem to have misinterpreted this point, Ward, 2015 for a critique, and Andrews &

Bonta, 2010, for the most recent edition of the text that establishes the definition and nature of the concept of *criminogenic need*). The main confusion appears to stem from authors observing research findings where measures of purported criminogenic needs failed to predict recidivism. Unfortunately, many jurisdictions assume that measures that predict recidivism in one population will also predict recidivism in another, or that certain need factors should be targeted in their programs simply because this has been done elsewhere. The authors cited above are correct to point out that these situations represent severe failures to correctly implement effective practices, or conduct due diligence to ensure that assessments and rehabilitation targets are meaningful for the population in their jurisdiction. However, failures of implementation do not automatically suggest weaknesses in the conceptual framework, and Andrews and Bonta (2010) clearly state that variables that are not associated with recidivism have no business being the primary targets within rehabilitation programs.

Dynamic risk variables are important within intervention because Andrews and Bonta's (2010) rehabilitation theory posits that changes in dynamic risk factors underlie successful programs; specifically, their "approach assumes that the connection between treatment and recidivism is mediated by intermediate change in criminogenic need factors" (Andrews, Hoge, Bonta, Gendreau, & Cullen, 1990, p. 425). This position appears to be reasonable, based on some existing data, but this is also perhaps the clearest example within correctional psychology of practice needing to function without explicit evidence of core assumptions (Olver, Beggs Christofferson, Grace, & Wong, 2014; Olver & Wong, 2011; Polaschek, 2012;

Woessner & Schwedler, 2014). The existing body of literature most strongly supports the conclusion that (a) rehabilitation programs that target dynamic risk factors are more effective for reducing recidivism than programs that target factors unrelated to recidivism, and (b) offenders who attend well-designed programs exhibit lower recidivism rates, as a group, compared to groups of non-participating offenders (Andrews & Bonta, 2010). However, as highlighted above, case decisions would benefit from more individual-level knowledge, rather than only knowing that an offender was exposed to interventions that likely produced some change in the group, on average.

Meta-reviews of Intra-Individual Change During Rehabilitation

Case-level knowledge requires research that examines individual change in risk factors during the process of rehabilitation, but this type of research is unfortunately rare in correctional psychology. Thus, there is thin evidence to support Andrews and colleagues' (1990) assumption that within-individual decreases in risk factors are the reason some program completers successfully reduce their criminal activity following program participation (for reviews, see Banse, Koppehele-Gossel, Kistemaker, Werner, & Schmidt, 2013; Beggs, 2010; Helmond, Overbeek, Brugman, & Gibbs, 2014; Lloyd & Serin, 2014; Scott, 2004; Serin, Lloyd, Helmus, Derkzen, & Luong, 2013). Change measurement has been lacking in this field, and conclusions currently rely on evidence from less rigorous cross-sectional studies, which amounts to applying knowledge about between-level explanations of criminal behavior to the management of within-individual change (Farrington, 2011).

The authors of the reviews cited above attempted to summarize studies that explicitly measured the relationship between within-individual program change and post-program recidivism, and each emerged with only a handful of high-quality studies. Yet, the collection of existing evidence shows some important consistency, despite the reviewers varying in whether they focused on general offenders, substance misusing offenders, violent offenders, sexual offenders, or family violent offenders. Specifically, change in beliefs about crime (pro-crime cognitions) was the most dependable predictor of reduced recidivism, across the most primary studies, and across all offender sub-types. For each offender sub-type, researchers were most likely to measure changes in offence-specific beliefs. For example, lower recidivism was observed (a) among general offenders who reported reductions in general beliefs about crime (see Serin et al., 2013), (b) among substance misusing offenders who reported reductions in beliefs about the consequences of drug use (e.g., Millson, Weekes, & Lightfoot, 1995), and (c) among sex offenders with child victims who reported reductions in beliefs supportive of offending against children (see Beggs, 2010).

Also, among general, substance misusing, violent (Serin et al., 2013), sexual (Beggs, 2010), and family violent offenders (Scott, 2004), reductions in negative emotionality (anger, hostility, sadness, stress, or psychological distress) emerged as a potentially important predictor of recidivism at the individual level. Finally, there was a generally consistent finding among violent (Lloyd & Serin, 2014) and sex offenders (Beggs, 2010) that increases in personal insight during rehabilitation were related to reduced recidivism. Other change variables were less consistently

related to post-program recidivism, but limited findings showed significant and meaningful reductions in antisocial peers (among general offenders; e.g., Kroner & Yessine, 2013), increases in assertiveness and empathy (among sex offenders; Beggs, 2010), and reduced substance misuse (among general offenders, sex offenders, and family violent offenders; Serin et al., 2013; Beggs, 2010; Scott, 2004, respectively) were related to lower recidivism.

Although the search for empirically validated dynamic change variables in the rehabilitation literature has revealed some conceptual consistency across offender sub-types, studies have generally reported far more null or weak findings, compared to significant or meaningful findings. Notably, within each study's pool of questionnaires used to assess pro-criminal cognitions, many researchers found that only a few of the pre-post measures showed relationships with recidivism (e.g., Simourd, Olver, & Brandenburg, 2015). Also, some pre-post measures were an effective predictor within some programs, but not within other similar programs, without evident explanations for these discrepant findings (Serin et al., 2013).

In general, studies published during the years following these reviews have utilized higher-quality methodologies, yet findings have remained decidedly inconsistent. For example, among studies examining change in sex offenders attending rehabilitation, one study reported that change scores predicted recidivism beyond static risk scores (Beggs & Grace, 2011), but this finding is balanced by non-significant results in both large sample (Barnett, Wakeling, Mandeville-Norden, & Rakestrow, 2013; Wakeling, Beech, & Freemantle, 2013) and small sample programs (Nunes, Pettersen, Hermann, Looman, & Spape, 2014; Pettersen et al.,

2015; Woessner & Schwedler, 2014). Similarly, Olver, Kingston, Nicholaichuk, and Wong (2014) found only a few subscales predicted recidivism in their sample of sex offenders, while most pre-post measures did not; however, change in sexual deviance was a consistent predictor across two samples (Olver, Beggs Christofferson, et al., 2014). Also, one study with violent offenders found changes in attitudes and anger did not predict post-program recidivism (Klepfisz, O'Brien, & Daffern, 2014), whereas change scores in other programs predicted violent recidivism (De Vries Robbé, de Vogel, Douglas, & Nijman, 2015; Lewis, Olver, & Wong, 2012; Novaco & Taylor, 2015).

Future Directions for Dynamic Risk Research in Rehabilitation

Despite disappointingly inconsistent findings, the recent rehabilitation literature has improved efforts to advance understanding of the person-level processes within programs. However, in terms of establishing whether changes in dynamic risk factors may cause reductions in crime, existing studies still largely lack the adequate design to test whether (a) change is associated with participation in the program, and (b) change is associated with lower recidivism at the individual level. In particular, the majority of existing individual-level studies are limited by excluding a control group from their design, and random assignment studies rarely measure within-individual change. It is also most typical for within-individual studies to include small sample sizes.

Furthermore, conclusions remain limited due to the field's lack of access to well-validated measures of dynamic risk variables. Validation of measures of dynamic risk is a critical first step toward providing convincing evidence that

manipulations of dynamic risk were successful at the individual level. As a result, it is reasonable to hypothesize that the inconsistencies within this area of literature may be primarily due to measurement techniques. Some have pointed out that the field has yet to develop a consistent operational definition of pro-criminal beliefs, which likely leads to inconsistencies in research findings, as well as inconsistencies regarding which assessment procedures are implemented in various criminal justice agencies (Caudy et al., 2013; Taxman, Caudy, & Pattavina, 2013). However, a good operational definition of criminal cognitions will arguably be multi-dimensional, because there are a number of empirically supported theoretical frameworks that each define pro-criminal attitudes in a unique way (e.g., rationalizations, techniques of neutralization, criminal self-efficacy, outcome expectancies, cognitive distortions, moral disengagement, etc.; see Helmond et al., 2014). Thus, the lack of consistent operational definition reflects the existence of a large variety of questionnaires; yet, this variety accurately reflects the current state of theory, which emphasizes that offender thinking styles are multifaceted.

In addition to working to better validate measurement techniques, there are at least three further ways researchers may advance the literature on within-person change in rehabilitation. First, greater attention should be paid to simultaneous changes in self-report cognitions and validated dynamic risk scores. For example, Devlin and Gibbs (2010) reported that both risk scores and cognitions changed during treatment, and that these changes were moderately correlated with each other. However, only pre-post change in pro-criminal beliefs was related to

recidivism, suggesting important rehabilitative improvements were better assessed by self-report questionnaires, compared to a formal risk measure.

Second, some researchers have demonstrated that the predictive ability of program change interacts with degree of offender risk. In particular, change scores predicted sexual recidivism for high-risk offenders only, and predicted violent recidivism for high-psychopathy offenders only (Olver & Wong, 2009; Olver, Wong, Nicholaichuk, & Gordon, 2007). Similarly, Quinsey, Jones, Book, and Barr (2006) showed that changes in dynamic risk among offenders living in the community were more strongly related to recorded incidents among high risk individuals, compared to lower risk clients. Researchers who have access to samples that are large and diverse should similarly investigate whether non-significant findings may be the result of measuring change scores among lower risk clients.

Finally, several researchers have piloted measures intended to assess client strengths during rehabilitation, and it is an important question whether protective factors increase during programs and whether these increases are associated with lower recidivism rates. In particular, Bergeron and Miller (2013) showed that, on average, offenders who completed rehabilitation programs self-reported reduced dynamic risk and increased protective factors at the end of the program. Similarly, dynamic risk and protective factors functioned in the same way together in another program where both risk and protective change scores predicted violent recidivism (de Vries Robbé et al., 2015). However, in one of the few studies that randomly assigned participants to rehabilitation and control groups, intent-to-treat analyses showed no significant group differences in participant-level changes in dynamic risk

and protective factors, although youth who completed the program showed increases in some protective factors (Baglivio & Jackowski, 2013). It remains unclear whether measures of protective factors will add to knowledge about within-individual change in rehabilitation. However, since the program change literature is in its infancy, it would be worthwhile for future researchers to incorporate a multi-construct approach to change measurement (Serin et al., 2013; Woessner & Schwedler, 2014).

Dynamic Risk Factors *May* Be Important for Prediction

The research described in the sections above suggests it is important to consider dynamic factors when theorizing the causal influences on crime, and intervening in offenders' lives through rehabilitation. Whereas rehabilitation research considers dynamic factors as mechanisms of change, the risk assessment literature examines whether measures of dynamic risk factors have utility for prediction. Logically, if dynamic risk factors are reduced through program participation, risk to re-offend is also reduced, and for prediction purposes, this additionally implies that the second, post-program assessment should be a more accurate indicator of an individual's future risk. Although this is a meaningful hypothesis, the current debate within the risk assessment literature has instead centered on whether dynamic factors should be assessed *at all*, not whether dynamic factors should be *re-assessed* to gain a more up-to-date forecast. Thus, it is less clear whether purportedly dynamic factors are valuable for prediction.

One extreme position in this debate argues that assessment of dynamic risk factors should be abandoned for a sole focus on static risk factors, which are more

clearly defined, easily measured, and powerfully predictive (Gottfredson & Moriarty, 2006). That the predictive ability of dynamic risk factors is in question represents a critical shortcoming, because without evidence that a dynamic factor is statistically associated with future crime, that variable should not be considered a risk factor, should not be incorporated into theories of criminal behavior, and should not be given weight as a target for intervention.

Below, this section argues that only multiple assessment research designs can adequately address whether (truly) dynamic risk factors have value for prediction. These research designs have been especially rare. This section reviews the small body of literature that has attempted to examine the relationship between assessment timing and prediction accuracy, which includes both single assessment and multi-assessment designs. Prior to the review of the research findings, this section describes the correctional perspective on risk assessment using dynamic variables.

Dissimilarities Between Rehabilitation and Risk Assessment Research

Broadly, it is somewhat artificial to distinguish between the rehabilitation and the risk assessment literatures, when attempting to understand dynamic risk. The purpose of both literatures is to estimate the relationship between dynamic risk factors and recidivism, and predict reoffending with case-level knowledge. However, there are three important dissimilarities.

First, the rehabilitation literature examines changes during incarceration, with few exceptions (e.g., studies of community based rehabilitation, and studies that use institutional behavior as the recidivism outcome). As a result, the

rehabilitation literature uses change scores as static predictors of recidivism in the community, since changes in scores are no longer examined once the offenders are released post-program. By contrast, studies that examine change in risk factors among offenders in the community are better able to assist supervision decisions, and examine “real-time” prediction of recidivism among offenders continually at risk to reoffend.

Second, rehabilitation studies are limited to including only features of the individual (beliefs, traits, emotions) in their definition of dynamic risk factors, whereas community based studies can additionally incorporate information about environmental influences (neighborhood, social factors). A variety of research supports the inclusion of context factors in risk assessment. For example, information about incarcerated offenders’ post-release plans adds incremental prediction above static scores, suggesting the context in which offenders will be released may be independent of static risk (e.g., Dickson, Polaschek, & Casey, 2013; Scoones, Willis, & Grace, 2012). Similarly, dynamic aspects of context (stability in housing, family features, and use of leisure time) accounted for some of the variance between static scores and recidivism outcome (Kroner, Gray, & Goodrich, 2013). Thus, although some researchers make a conceptual distinction between re-entry context and dynamic risk factors, environmental variables appear to behave similarly to, and may simply be another form of dynamic risk (Scoones et al., 2012).

Indeed, context variables are often included in the definition of dynamic risk, or, in many cases, a dynamic risk factor can be best defined as an interaction between individuals and their context. For example, variables such as *social support*,

access to victims, and *use of free time* reflect both individual choices, and existing opportunities. Including context or context-individual interactions in dynamic risk assessment is consonant with the theoretical frameworks reviewed in the first section above. Also, in practice, a context-free estimation of risk has only limited value for clinicians attempting to understand and manage an offender's risk (Campbell, French, & Gendreau, 2009; Morgan, Kroner, Mills, Serna, & McDonald, 2013; Müller-Isberner, Webster, & Gretenkord, 2007; Yang, Wong, & Coid, 2010).

Third, the rehabilitation literature has traditionally examined dynamic change using measures of self-reported psychosocial factors rather than relying on formal risk assessments, with few exceptions (e.g., the Violence Risk Scale; Olver et al., 2007). In contrast, community based studies typically examine formal measures of risk that may incorporate the psychosocial variables measured in programs, but usually include a larger variety of risk domains. These measures were primarily built for the purpose of maximizing prediction and establishing long-term risk probability. Thus, despite including dynamic risk factors, these measures rarely have been employed for repeated measurements, nor has research suggested these measures are particularly sensitive to detecting change across intra-individual assessments (Kroner et al., 2007; Morgan et al., 2013). For example, the Level of Service Inventory-Revised (LSI-R; Andrews & Bonta, 1995) is one of the earliest formal risk measures to incorporate conceptually dynamic variables, but the LSI-R has only recently been studied for its ability to detect change (e.g., Brooks Holliday, Heilbrun, & Fretz, 2012; Devlin & Gibbs, 2010; Labrecque, Smith, Lovins, & Latessa, 2014; Morgan et al., 2013; Raynor, 2007; Schlager & Pacheco, 2011; Vose, 2008).

When statistically significant, results from these studies have been consistent with hypotheses (e.g., decreases in LSI-R risk scores were related to lower recidivism among at-risk offenders; Andrews & Robinson, 1984; Raynor, 2007), but findings have been generally weak in magnitude (e.g., Vose, 2008).

More Community Based Dynamic Risk Research is Necessary

As a result of these limited and weak results, the current literature provides little guidance to community based correctional agencies regarding whether they should regularly re-assess offender clients using formal risk measures that include dynamic risk domains. Theoretically, and by definition, re-assessment of dynamic risk factors should be valuable for providing more accurate, up-to-date information about re-entering offenders, and a number of researchers argue that dynamic factors should be routinely measured throughout supervision to both test this hypothesis, and assist correctional decision-making (Day, 2013; Hanson & Harris, 2000; Kroner et al., 2007; Zamble & Quinsey, 1997). However, the field has conducted too few studies on the predictive ability of dynamic risk factors to provide adequate support for the core conceptual assumptions about dynamic risk factors. More research is necessary.

In particular, in both rehabilitation and re-entry contexts, the status of dynamic factors as *conceptually* dynamic has been taken for granted (Philipse et al., 2006), without explicit tests of whether dynamic factors change or can enhance prediction when re-assessed. Although dynamic variables do not need to provide incremental validity over static variables to be useful in practice for rehabilitation and risk management (Olver et al., 2007), their exact definition insinuates that

dynamic risk factors should add to prediction over static factors, by (a) introducing more psychologically relevant content into the assessment, (b) expressing the magnitude of recent reductions in risk via change scores, and/or (c) providing a more accurate assessment of current functioning, due to assessment proximity. In order to establish the predictive value of dynamic factors among re-entering offenders, temporal proximity is the key underlying assumption, and it would be a waste of resources for an agency to re-assess items that appear dynamic, but behave statically in actuality. Arguably, the proximity hypothesis can only be tested in a sample of offenders that is continually at risk to re-offend, after release from incarceration.

Is there Evidence that Re-assessments Enhance Prediction?

Single-occasion measurement strategies.

The most common methodology for examining the predictive value of dynamic risk domains involves measuring theoretically dynamic variables at a single point in time (e.g., van der Knaap, Alberda, Oosterveld, & Born, 2012). This research relies on the logic that a single, baseline assessment can conceptually measure static factors that were solidified earlier in time, while measuring the current state of the client's dynamic factors. Using this methodology, there has been inconsistent support for dynamic risk factors adding to prediction over static items. For example, in one study, dynamic items (client denial, poor self-management, social isolation, low empathy) did not predict incrementally over static items (Philipse et al., 2006). Similarly, static background variables were more strongly associated with juvenile delinquency compared to current dynamic variables (peers,

substance use; van der Laan et al., 2009). Also, although another study demonstrated that some dynamic risk factors predicted above static risk variables, prediction was weak (Caudy et al., 2013).

Inconsistent results have also been observed using formal risk measures. Two studies found prediction was reduced when using the full version of the LSI-R, compared to a shortened, static-variable-only version (Austin, Coleman, Peyton, & Johnson, 2003; Raynor, 2007). Also, meta-analyses of risk assessment have either suggested weak (Campbell et al., 2009; Yang et al., 2010) or modest (Hanson & Morton-Bourgon, 2009) incremental prediction of dynamic risk factors over static factors. However, counterbalancing these non-significant and weak findings, a number of primary studies have demonstrated that some dynamic variables can predict recidivism above static variables (e.g., Beech, Friendship, Erikson, & Hanson, 2002; Desmarais, Nicholls, Wilson, & Brink, 2012; McDermott, Edens, Quanbeck, Busse, & Scott, 2008; McGrath & Thompson, 2012; Thornton, 2002).

Yet, even when baseline dynamic variables incrementally predict above static risk factors, it remains unclear whether enhanced prediction is due to adding new content areas to the assessment, or due to the dynamic variable's theoretical proximity to criminal action (Kroner et al., 2007). As a result, this approach has been correctly criticized for neither describing the relationship between change and recidivism, nor providing evidence that assessment proximity is related to enhanced prediction (Olver et al., 2007; Woessner & Schwedler, 2014). Although relatively unsophisticated research methodologies help lay the basic foundation of knowledge within the literature, establishing predictive validity for dynamic risk factors must

begin to take time into account, thus extending beyond tests demonstrating that dynamic factors (a) independently predict recidivism, (b) incrementally predict recidivism above static risk factors, or (c) predict recidivism when measured at only one point in time.

On the other hand, Walters (2002, 2006) utilized creative measurement strategies within a single-occasion design to distinguish past and current risk factors. First, Walters (2002) asked participants to report past criminal thinking and current criminal thinking within a single assessment. Results were consistent with the hypothesized time frame built into the two sets of items, such that past criminal thinking had a relatively stronger relationship with indicators of criminal history, whereas current criminal thinking was more strongly related to future recidivism. Using the same measure, Walters (2006) found criminal thinking measures only weakly predicted post-release recidivism when assessments were older than two years prior to the offenders' release. This suggests these assessments had a finite "shelf life", and that assessment proximity was associated with stronger prediction.

Segmenting the follow-up.

Other researchers have similarly utilized one baseline assessment, but examined proximity by testing whether the risk measure predicted immediately ensuing recidivism events more strongly, compared to events later in the follow-up period. For example, Ullrich and Coid (2011) used an assessment of dynamic factors to predict recidivism incidents separately during each year post-release. In each year, recidivists during that particular year were compared to non-recidivists who

remained crime-free throughout the entire follow-up period. Variables that were associated with recidivism were consistently related to recidivism across the entire time period, with the exception of *access to housing*, which significantly predicted recidivism only in the first year of follow-up. With a similar analytic strategy, Lofthouse and colleagues (2014) used dynamic variables to predict outcome incidents exclusive to each month within their sample's follow-up period, but results did not suggest that prediction varied substantially from month to month. However, in another study, dynamic items were significant predictors of recidivism when the follow-up was limited to 1-3 months, whereas prediction did not remain significant for 6-month outcomes (Chu, Thomas, Ogloff, & Daffern, 2013).

Comparing post- versus pre-program assessments.

Typically, the rehabilitation literature has emphasized the magnitude of pre-post-program change as the prediction variable of interest. Yet, among the many studies that have failed to demonstrate that change scores predict reoffending, some offer evidence that more proximal scores may be better predictors compared to more distal scores. For example, post-program scores have significantly predicted recidivism, even when pre-program and change scores have not (e.g., Pettersen et al., 2014; Wakeling et al., 2013; however, see also Helmus, Hanson, Babchishin, & Mann, 2013).

Post-program scores represent more up-to-date information about offenders' dynamic risk factors, and this proximity to recidivism may enhance prediction. However, because these studies measure offenders immediately after program participation, these assessments are not solely measuring more proximal

information, but scores potentially also reflect the degree of client insight and cooperation at the end of the program. Thus, it remains unclear whether the proximity of the assessment alone enhanced prediction, or if post-program measures may be better predictors because they incorporate additional latent information about client insight. This is especially a criticism among studies where only program completers were assessed post-program (which is the most typical design). The inclusion of a no-participation control group would better support the conclusion that assessment proximity itself enhanced prediction.

Additional indirect evidence from the rehabilitation literature suggests that post-program functioning may be better able to predict recidivism compared to pre-program functioning. Olver and Wong (2011) showed that a static risk measure was not an effective predictor of recidivism among offenders who exhibited the most change during the program, compared to the measure's predictive ability among offenders showing little program change. This suggests the static measure may have been failing to capture important proximal information about risk, specifically among successful clients, due to its historical nature. Similarly, the predictive ability of historical risk factors was reduced when used to predict recidivism among a group of program completers, compared to non-completers (Studer & Reddon, 1998). It is unclear from these study designs whether dynamic factor information should be weighted more heavily compared to static factor information following program participation, and it should be noted that some tests only approached statistical significance. Also, a meta-analysis with a large sample found no evidence that average static scores or prediction accuracy differed between samples of

individuals who attended rehabilitation, compared to routine correctional samples (Hanson, Thornton, Helmus, & Babchishin, 2015). Although it is an intriguing hypothesis, it remains unclear whether static prediction degrades among participants who experienced substantial change during rehabilitation.

Retrospective matched designs.

Retrospective matched designs employ interviews or file reviews to examine the circumstances that occurred just prior to recorded recidivism events. This methodology establishes the importance of proximal risk factors by examining whether the individual and environmental factors that occurred prior to a criminal action were different from (a) the circumstances of a similar non-recidivist (a single-time-point, between-individual design), or (b) the circumstances of the same recidivist at an earlier point in time (a multiple-time-point, within-individual design).

Zamble and Quinsey (1997) largely pioneered this approach for identifying proximal acute risk, using a between-individual design. They interviewed offenders as they returned to incarceration for a new recidivism event and non-recidivists remaining crime-free in the community. Their results suggested a number of dynamic risk factors differentiated those who committed new criminal action, including negative emotionality, substance use, lack of social support, and poor use of leisure time (including association with criminal peers).

Quinsey, Coleman, Jones, and Altrows (1997) used both between- and within-person analyses with retrospective data from a sample of offenders who met the criteria for major psychiatric diagnoses. In their within-individual analyses,

results indicated the dynamic risk factors that increased prior to a criminal event included impulsive behaviors, social withdrawal, negative or extreme emotionality, cognitive distortions favorable toward crime, psychosis, poor compliance with supervision, and poor coping skills. Using similar samples and measurement, Quinsey and colleagues (Quinsey, Book, & Skilling, 2004; Quinsey et al., 2006) found additional support for comparable dynamic factors.

With a sample of sex offenders, Hanson and Harris (2000) conducted retrospective interviews with supervision officers, and coded files written prior to recidivism events. Overall, recidivists exhibited more problems compared to non-recidivists, but recidivists also demonstrated deterioration from earlier levels of functioning prior to a recidivism event. Identified dynamic risk factors included increased substance use, negative emotionality, social problems, victim blaming, and poor self-management, which are generally similar to Quinsey and colleagues' findings. The most important acute risk factors for these sex offenders included sexual preoccupation, access to victims, non-acknowledgement of personal risk for recidivism, anger, and using medication to manage risk. Change scores indicated increased anger was related to recidivism. Static, stable, and acute risk factors (defined conceptually) were each incrementally associated with recidivism, with a set of stable factors (denial of risk, poor social influences, and beliefs of sexual entitlement) sharing relatively more variance with outcome compared to static and acute risk factors.

Lindsay and colleagues (2004) demonstrated a retrospective within-individual design with a small sample of offenders characterized by intellectual

delays. They compared ratings recorded on the day of an incident to the day prior to the incident, as well as a randomly selected day at some point prior to the incident (the control observations). These within-individual analyses showed that incident days (the day of the incident, and the day before) were characterized by significantly higher ratings of negative emotionality, poor self-regulation, antisocial behavior, psychotic thinking styles, and poor compliance, compared to control days. Finally, in a sample of high-risk individuals meeting the criteria for a psychiatric diagnosis, participants were more likely to report alcohol use on days prior to a violent incident, compared to other days during the follow-up period (Mulvey et al., 2006). By contrast, reported marijuana and other substance use were unrelated to the timing of violent incidents.

Within-individual retrospective matched designs are valuable because they compare proximal versus distal assessments. In particular, these designs are most clearly useful for identifying the variables that occur at the time of criminal actions, but are limited because they are unable to prospectively estimate which individuals are most likely to re-offend, and when recidivism is likely to occur. These designs involve *a priori* selection of two assessments, based on their timing, rather than a comparison of a single prospectively rated imminent assessment against all other potential assessments within the follow-up period. For this reason, it is unclear if differences between the two assessments are inflated by how the assessments were selected. In addition, the retrospective nature of the analyses raises concerns about participant recall bias (since the timing of the recidivism event is typically known at

the time of the interview or file coding). For these reasons, prospective designs provide a more rigorous way to examine the proximity hypothesis.

Prospective designs with re-assessment data.

In the broader field of criminology, there are a large variety of prospective re-assessment studies examining criminal behavior across time. However, many of the existing multi-wave studies do not share the primary goals of correctional psychology. In particular, the criminology literature often conducts longitudinal research with (a) non-offender juvenile populations, in order to identify the correlates of delinquency onset (e.g., Chen, 2010; Childs, Sullivan, & Gullledge, 2011) or (b) juvenile offenders, in order to describe the developmental differences between normative, rule-breaking adolescents and long-term, serious offenders (e.g., Cardwell et al., 2015; Shulman, Cauffman, Piquero, & Fagan, 2011). In addition, the longitudinal literature has often focused on patterns of criminal behavior across many years or decades (e.g., Kazemian, Farrington, & Le Blanc, 2009; Monahan & Piquero, 2009), with specific attention to patterns of onset, escalation, and cessation of crime (see Piquero, 2008, for a review). By nature of this literature's life-course-level timeframe, these studies may not include measures of dynamic risk factors, or will typically define dynamic risk factors as features of childhood, family of origin, and early personality traits.

This contrasts with the objectives of correctional psychology, which attempts to predict single instances of re-offending using relatively rapidly changing situational and intra-personal features that (a) are amenable to management or intervention, and (b) occur within the period when adult offenders are re-entering

the community from incarceration. Multi-wave studies that address these objectives are uncommon.

Even more uniquely, few prospective multi-wave re-entry studies have attempted to test whether re-assessment timing is related to more accurate prediction. Rather, these studies have examined (a) change processes and their relation to recidivism (e.g., Babchishin, 2013; Blanchard, 2013; Labrecque et al., 2014; Michel et al., 2013), (b) the general relationship between dynamic factors and recidivism (e.g., Greiner, Law, & Brown, 2014), and (c) whether repeated assessments are incrementally related to recidivism beyond static risk scales (e.g., Brown, St. Amand, & Zamble, 2009; Morgan et al., 2013). Although these research questions (change, predictive validity, and incremental prediction) represent important goals, examining assessment proximity requires testing whether a dynamic measure can add incremental prediction beyond *its own prior measurement*.

Lagged prediction models. To test this, three studies used lagged prediction models, which determine if variables measured proximally to violent behavior are more strongly related to outcome, compared to measures from all other assessment periods (Buchanan, Nich, Douglas, Babuscio, & Easton, 2013; Skeem et al., 2006; Yang, Mulvey, Loughran, & Hanusa, 2012). These authors conducted multi-wave assessments of individuals meeting the criteria for psychiatric diagnoses, and identified hostility, alcohol use, and lack of support as proximal indicators of violence. One study did not find a proximal prediction effect for psychosis or negative emotions (Skeem et al., 2006), whereas another study found a proximal prediction effect for moderate levels of psychosis (versus high or

low levels), and negative emotions, only when alcohol use was low (whereas high alcohol use was consistently related to violence; Yang et al., 2012).

Survival analysis. Three additional studies used the analytic strategy described in Chapter 2 of this dissertation (i.e., Cox regression with time-varying predictors; Brown et al., 2009; Howard & Dixon, 2013; Jones, Brown, & Zamble, 2010). First, Howard and Dixon (2013) explicitly tested whether re-assessments added predictive validity over initial assessments of the same variables. Their follow-up included 12 discrete three-month periods that spanned four years. All dynamically assessed variables significantly added predictive validity above initial assessments, including personal features (pro-criminal beliefs, anger, and alcohol misuse) and re-entry features (accommodation, employment, and access to psychiatric care).

These findings are unique for providing statistically significant support for the conclusion that re-assessment enhances prediction of recidivism outcome. However, when evaluating the importance of these findings, note that the substantial size of Howard and Dixon's (2013) sample ($N = 196,493$ individuals, collectively providing 2.7 million assessments) diminishes the weight that should be put on traditional tests of statistical significance. Notably, the effect sizes did not increase when comparing prediction using the re-assessment models versus the baseline-only models. Thus, Howard and Dixon's (2013) statistically significant findings did not appear to have practical importance for prediction accuracy.

Similarly, Brown and colleagues (2009) did not find differences in prediction effect sizes when they compared a baseline model versus a re-assessment model,

even though the re-assessed variables demonstrated statistical significance in the regression models. However, using the same sample with a longer follow-up time (four ratings across six months, rather than three ratings taken in the first three months), the difference in effect size between the baseline-only and re-assessment models approached statistical significance ($p < .10$; Jones et al., 2010). Note, however, that although these authors incorporated re-assessments into their design, their baseline and re-assessment prediction models did not include exactly the same dynamic variables. This makes it unclear whether re-assessments showed a trend to increasing prediction accuracy because these assessments added new content, or due to the proximity of the assessment to criminal behavior.

Still, Jones and colleagues' (2010) research supports the value of re-assessment. Although dynamic variables measured at baseline demonstrated the poorest prediction, and static variables showed the strongest prediction, when static and re-assessed dynamic variables were combined together in one model, the effect size was appreciably enhanced (again, $p < .10$ when comparing model effect size). This supports the conclusion that (a) re-assessment of dynamic variables adds to prediction when incorporated with static variables, and that (b) the true predictive value of dynamic factors is underestimated when dynamic variables are measured only at baseline.

Averaging vs. proximity. Finally, Hanson and colleagues (2007) similarly struggled to find consistent evidence that proximal assessments enhanced prediction above prior assessments, among a sample of sex offenders. In particular, when examining conceptually stable dynamic factors, proximal assessments

enhanced prediction over initial assessments in only one out of ten tests (using different definitions of recidivism, and different lengths of follow-up). Similarly, ratings of conceptually acute dynamic factors taken close to recidivism were highly correlated with prior ratings, and did not incrementally improve prediction. Not only did these theoretically acute variables remain largely stable in actuality, the average of a larger number of re-assessments was more strongly related to recidivism, compared to the average of a lower number of re-assessments. Although this finding supports the conclusion that re-assessment enhances prediction, this effect is logically due to combining more information about the offender across time, rather than assessing the offender's risk state immediately prior to recidivism.

In a re-assessment of Hanson and colleagues' (2007) data, Babchishin (2013) used a different approach to test the incremental prediction of re-assessments over baseline scores. This approach involved estimating individual-level patterns of change from observed acute risk factor scores, and then imputing model-derived risk scores within each two-month interval in the possible 12-month follow-up. Overall, results suggested there was incremental prediction of re-assessments over baseline assessments for sexual recidivism, but not violent recidivism (i.e., sex and violent recidivism, combined). Using the broadest definition of recidivism (sex, violent, and all other types of recidivism), re-assessments demonstrated added prediction over baseline scores. Similar to Hanson and colleagues' (2007) analysis, proximal assessments enhanced prediction for some definitions of outcomes, but not all. However, unlike prior work, Babchishin's (2013) approach suggested a proximity effect for acute factors in addition to stable factors.

Babchishin's (2013) results suggest some promise for the importance of re-assessments, but these analyses may not directly test the proximity hypothesis because (a) re-assessment scores were model-derived, not observed scores, and, more importantly, (b) this analytic strategy does not explicitly match re-assessment scores to the timing of recidivism. The most conservative interpretation of these results suggests that rate of change information enhances prediction over baseline assessments, but would caution about concluding that proximal re-assessment enhances prediction, because the timing of recidivism events was not matched to assessments within the analysis.

Conclusion. Thus, there is limited prospective research that has examined the predictive validity of temporally proximal dynamic risk factors in comparison to more distal assessments. In addition, the most compelling current evidence suggests that re-assessments become important only when they are either combined (a) with static measures (Brown et al., 2009; Jones et al., 2010), or (b) into averages with other re-assessments (Hanson et al., 2007). However, with so few existing studies, it is impossible to determine how much each finding may have been influenced by unique features of the study from which it emerged, including type of sample, timing of assessments, length of follow-up, or assessment protocols. For example, the authors of one prospective study were unable to test any hypotheses about assessment proximity because none of the measures were related to recidivism; at the end of the study, there was evidence that the measurement strategy had been less than optimal (Morgan et al., 2013).

To advance research on the value of re-assessment, the field also requires better validation of assessments of dynamic risk factors. Currently, limited findings create a situation where one set of authors can recommend relying less on self-report measures and more on ratings from supervision officers (Morgan et al., 2013), another set of authors can recommend relying less on ratings from supervision officers and more on self-report measures (Brown et al., 2009), and there is little additional evidence to contribute to this debate.

Summary & Conclusion

This chapter argues that understanding dynamic factors remains important for theory and correctional practice, even if research findings supporting predictive accuracy are currently limited. Critics have suggested that including dynamic factors in risk assessment increases the complexity of the prediction task, and creates unclear, multi-construct measures that predict recidivism to a lesser degree (Gottfredson & Moriarty, 2006). This position assumes that risk prediction and risk reduction should be separate enterprises (as suggested by Baird, 2009; Caudy et al., 2013), yet in practice, risk reduction cannot occur without (a) knowledge of dynamic factors that predict recidivism, (b) accurate assessment of these factors, and (c) accurate re-assessment (if these factors are truly dynamic). The field must rely on knowledge from risk assessment research to justify correctional intervention, and prediction accuracy must ultimately determine which dynamic risk factors deserve attention. Thus, the research question of interest is not *if* dynamic risk factors are related to recidivism, but rather determining *which* changeable factors are related to recidivism, and *how* they manifest into an

individual's risk state. Certainly, risk tools that combine the goals of prediction and management may confuse correctional decision-makers (Caudy et al., 2013); it is a complex task to apply the results of these dual-role instruments within case-level decisions. But, in a functioning correctional system, it is artificial to separate prediction from efforts to manage offenders (Philipse et al., 2006).

In addition, as argued in this chapter, developing a more mature theoretical understanding of criminal behavior requires new research attention toward acute dynamic risk factors. Instead, the prediction literature has suffered from a sole reliance on historical factors, or has used an atheoretical approach to select dynamic items for actuarial instruments. Ignoring the psychological meaning of risk factors through atheoretical selection limits the field's ability to understand the phenomena being predicted (Krauss, Sales, Becker, & Figueredo, 2000). In this way, investigating the predictive value of re-assessed dynamic factors *selected by theory* is a key way to begin feeding critical knowledge into the field's theories of criminal behavior.

Within both the rehabilitation and prediction studies that have attempted to identify dynamic risk factors, there is a concerning high ratio of non-significant to significant findings. Yet, there is also some consistency regarding which theoretically dynamic variables are related to recidivism, since most dynamic risk variables have found statistically significant support in more than one study. This consistency is visually displayed in Table 1.2 and Table 1.3. In particular, Table 1.2 lists the studies that have either used a retrospective matched design or a prospective prediction design, and specifies the main defining features of these

studies. Table 1.3 displays which dynamic variables have been tested for associations with recidivism, and whether findings were statistically significant (designated with a ✓) or non-significant (designated with a ✗). Table 1.3 also indicates the number of studies with identical results (i.e., a “2” that follows a checkmark represents that two studies demonstrated this same finding). Finally, superscripted alphabetic letters are placed next to checkmarks to allow readers to match the findings within Table 1.3 to the primary studies listed in Table 1.2.

Overall, as displayed in Table 1.3 and reviewed earlier in this chapter, there is some consistency across methodology to suggest that personal features (e.g., pro-criminal beliefs, hostility/anger, substance misuse) and re-entry factors (e.g., employment, accommodation, access to social support) can change, predict recidivism, add incremental prediction over static variables, and may be proximal indicators of immediate risk. This consistency encourages additional studies of dynamic prediction using these dynamic factors.

Research Question & Hypothesis

Correctional practice would benefit from more evidence that (a) dynamic factors change across time, and that (b) the re-assessment of dynamic factors enhances prediction of re-entry recidivism. These two hypotheses represent the defining features of dynamic risk factors. In order to add to the small existing body of research on dynamic risk, and to address some of the questions raised in this chapter’s literature review, this dissertation focuses on a standardized risk assessment measure (Dynamic Risk Assessment for Offender Re-entry, or DRAOR; Serin, 2007) used to prospectively re-assess offenders who were at risk in the

Table 1.2

Publications Reporting Retrospective Matched Designs or Prospective Multi-Assessment Designs to Assess the Relationship Between Dynamic Risk Factors and Criminal Recidivism

Source	Analytic Approach	Population	% male	Baseline N	Follow-up Length	# of Ratings	Study Code
Retrospective matched designs							
Hanson & Harris (2000)	within-individual	sex offenders	100	409	-	-	a
Lindsay et al. (2004)	within-individual	intellectually delayed offenders	-	5	-	-	b
Mulvey et al. (2006)	within-individual	psychiatric clients	48	132	-	-	c
Quinsey et al. (1997)	within-individual	psychiatric offenders	100	111	-	-	d
Quinsey et al. (2004)	within-individual	psychiatric offenders	100	57	-	-	e
Quinsey et al. (2006)	within-individual	psychiatric offenders	88	531	-	-	f
Zamble & Quinsey (1997)	between-individual	general offenders	100	347	-	-	g

Change and re-assessment prediction designs						
Blanchard (2013)	GEE	psychiatric offenders	51	235	5 months	6 ratings (max) h
Brown et al. (2009)	Cox regression	general offenders	100	136	3 months	3 ratings i
Jones et al. (2010) †	Cox regression	general offenders	100	127	6 months	4 ratings j
Buchanan et al. (2013)	GEE	psychiatric clients	100	75	12 weeks	4 ratings k
Greiner et al. (2014)	Cox regression	female offenders	0	497	2 years	4 ratings l
Hanson et al. (2007) Stable	logistic regression	sex offenders	100	799	1 year	2 ratings m
Hanson et al. (2007); Babchishin (2013) Acute	logistic regression	sex offenders	100	701	6 months	7 ratings (median) n
Howard & Moore (2009)	correlation	general offenders	-	1862	2 years	2 ratings o
Howard & Dixon (2013)	Cox regression	general offenders	87	196,493	4 years	12 ratings p
Labrecque et al. (2014)	logistic regression	general offenders	80	828	1 year	2 ratings q
Michel et al. (2013)	GEE	psychiatric offenders	96	246	2 years	5 ratings r

Morgan et al. (2013)	logistic regression	general offenders	100	133	6 months	7 ratings	s
Skeem et al. (2006)	SEM	psychiatric clients	48	132	6 months	26 ratings	t
Yang et al. (2012)	logistic regression	psychiatric clients	55	587	1 year	6 ratings	u

Note. Study Code = code for readers to link publications in Table 1.2 with results presented in Table 1.3. GEE = generalized estimating equations. SEM = structural equation modeling. Stable = Stable-2007 (Hanson et al., 2007). Acute = Acute-2007 (Hanson et al., 2007). #This publication reports on an extended follow-up of the dataset in the publication listed directly above, i.e., Brown et al. (2009).

Table 1.3

Dynamic Risk Factors and Risk Instruments Tested for Change and Predictive Validity at Re-assessment in the Community

Variable	Retrospective matched designs	Change across time	Change related to recidivism	Prediction of re-assessments over static	Proximal prediction over prior
Number of samples	7	7	5	5	6
Personal features					
Impulsivity	✓ (1 ^d)	✓ (1 ^q)	✓ (1 ^r)	-	✗ (1 ^k)
Pro-crime attitudes	✓ (4 ^{adef})	✓ (4 ^{lro}) ✗ (1 ⁱ)	✓ (2 ^{ro})	✓ (1 ⁱ) ✗ (1 ^s)	✓ (1 ^p) ✗ (2 ^{kp})
Anger/hostility	✓ (4 ^{abdg})	-	-	✓ (1 ⁿ)	✓ (2 ^{pt})
Negative emotions	✓ (2 ^{ae}) ✗ (1 ^f)	✓ (2 ^{il})	-	✓ (2 ⁱⁿ) ✗ (1 ^s)	✓ (1 ^u) ✗ (3 ^{ktu})
Positive emotions	-	✗ (1 ⁱ)	-	-	-
Empathy	✓ (1 ^f)	-	-	✗ (1 ^s)	-
Sexual preoccupation	✓ (1 ^a)	-	-	✓ (1 ⁿ)	-
Stress/well-being	✓ (1 ^g)	✓ (3 ^{iro})	✓ (1 ^o) ✗ (1 ^r)	-	✗ (2 ^{kt})
Coping ability	✓ (3 ^{abd})	✓ (1 ⁱ)	-	✗ (1 ^s)	-
Psychosis	✓ (3 ^{bdf}) ✗ (1 ^e)	✗ (1 ^r)	-	✗ (1 ^s)	✓ (1 ^u) ✗ (3 ^{ktu})
Use of leisure time	✓ (1 ^g)	✓ (1 ⁱ)	-	-	-
Substance misuse	✓ (2 ^{ag})	✓ (1 ⁱ) ✗ (1 ⁱ)	-	✓ (2 ⁱⁿ) ✗ (1 ⁱ)	-
Alcohol use	✓ (2 ^{eg})	✓ (1 ^o)	✓ (1 ^o)	-	✓ (3 ^{pu}) ✗ (1 ^k)

Drug use	✓ (1 _g)	✗ (1 _c)	✓ (1 _o)	✓ (1 _o)	-	✓ (1 _k)
Finances		✗ (1 _g)	✓ (2 _o)	✓ (1 _o)	-	-
Social variables						
Lack of social support	✓ (4 _{adeg})	✗ (1 _f)	✓ (4 _{lro})	✓ (1 _o)	✗ (1 _r)	✓ (2 _{in})
Antisocial peers	✓ (1 _g)		✓ (2 _o)	✗ (1 _i)	✓ (1 _o)	-
Re-entry variables						
Accommodation	-		✓ (1 _o)	✗ (1 _i)	✓ (1 _o)	✓ (1 _p)
Employability		✗ (2 _{ag})	✓ (3 _{lo})	✓ (1 _o)	✓ (1 _i)	✓ (1 _p)
Education	-		✓ (1 _o)	✓ (1 _o)	-	-
Life problems	✓ (1 _g)		✓ (1 _i)	-	✓ (1 _i)	-
Victim access	✓ (1 _a)		-	-	✓ (1 _n)	-
Community features		✗ (1 _a)	✓ (2 _{lr})		✗ (1 _r)	✗ (1 _k)
Unrealistic plans	-		✓ (1 _r)	✓ (1 _r)	-	✓ (1 _k)
Clinical variables						
Access to clinical care	-		-	-	-	✓ (1 _p)
Lack of insight/denial	✓ (3 _{aef})		✓ (1 _r)		✗ (1 _s)	✓ (1 _p)
Noncompliance	✓ (4 _{abdf})	✗ (1 _e)	✓ (2 _{lr})	✗ (1 _i)	✓ (1 _r)	✗ (1 _k)
Responsivity/alliance		✗ (2 _{bf})	✓ (1 _r)	✓ (1 _r)	-	-
Health	✓ (1 _g)			✗ (1 _i)	-	-

Risk instruments				
HCR-20 (C & R scales)	-	✓ (1 ^h)	✓ (1 ^h)	-
START (Vulnerability)	-	✓ (1 ^h)	✗ (1 ^h)	-
START (Strength)	-	✓ (1 ^h)	✓ (1 ^h)	-
LSI-R	-	-	✓ (1 ^q)	✗ (1 ^s)
Stable-2007	-	-	-	✓ (1 ^m)
Acute-2007	-	✓ (1 ⁿ)	✓ (1 ⁿ)	✓ (1 ⁿ)

Note. Within each row, a dynamic risk variable may have been assessed in a different way by different authors; results are combined based on conceptual similarity, and do not necessarily reflect identical measurement. ✓ = statistically significant finding. ✗ = statistically non-significant finding. "-" = results not reported. Numbers in parentheses equal the number of publications that reported this finding. Superscripts represent study codes so readers can link results reported in Table 1.3 to publications listed in Table 1.2. HCR-20 = Historical Clinical Risk Management-20 (Webster, Douglas, Eaves, & Hart, 1997). START = Short-term Assessment of Risk and Treatability (Webster, Martin, Brink, Nicholls, & Desmarais, 2009). LSI-R = Level of Service Inventory-Revised (Andrews & Bonta, 1995). Stable-2007 (Hanson et al., 2007). Acute-2007 (Hanson et al., 2007).

community. DRAOR includes dynamic risk factors identified by theory, and includes both psychosocial and context variables.

Prior work with DRAOR (Hanby, 2013, see further discussion in Chapter 3) established this measure's ability to predict recidivism using baseline assessments, and investigated patterns of intra-individual change in DRAOR scores across time. Hanby's (2013) analyses provided support for one of the two important features that define a dynamic instrument (i.e., scores are observed to change across time). Using the same dataset, this dissertation extends this prior work by examining the proximity hypothesis (i.e., is the observed change relevant for prediction purposes, such that re-assessments enhance prediction over prior assessments?).

The over-arching hypothesis for this project is that re-assessment enhances prediction. However, there are at least two disparate reasons why re-assessments may enhance prediction. On a conceptual level, the definition of dynamic factors logically demands that the most recent assessment should be the best predictor of recidivism because it is the most proximal assessment of the offender's current functioning and circumstances.

Hypothesis 1.1. In a population of at-risk offenders, the most proximal assessment is most strongly related to future recidivism, compared to all prior assessments.

By contrast, on a measurement level, prior research (Hanson et al., 2007, specifically) suggests it may be equally reasonable to hypothesize that re-assessments enhance prediction, but only when combined with other assessments. In other words, larger averages of assessments may provide more accurate single

point estimates of an offender's behavior and functioning. Under this assumption, combined estimates are expected to be more accurate, and thus stronger predictors of recidivism.

Hypothesis 1.2. In a population of at-risk offenders, larger combinations of assessments taken across time are more accurate predictors of recidivism, compared to smaller combinations of assessments.

This second hypothesis is similar to the first because both suggest that re-assessment should be valuable, in practice. However, Hypothesis 1.2 is conceptually distinct from the first because it emphasizes that the value of re-assessments lies within aggregating to attain more accurate information, rather than more proximal information. Within the correctional literature, there appear to be few prior studies to guide this hypothesis, beyond Hanson and colleagues' (2007) findings, as reviewed above. Providing an additional, but equivocal example, Quinsey and colleagues (2006) showed that average scores of dynamic factors assessed across the follow-up period differentiated between recidivists and non-recidivists, but these authors did not examine differences in prediction accuracy by prospectively testing predictors of varying proximity, or averaging strategies.

However, from the assessment literature, generalizability theory (see Webb, Shavelson, & Haertel, 2007) assumes multiple assessments are necessary to accurately identify stable differences among individuals (which establishes the measure's *generalizability*), and multiple within-individual assessments must be averaged before the mean score approximates the individual's "true" average (which establishes the measure's *dependability*). More broadly, generalizability

theory suggests that any single measurement (e.g., a risk assessment score) is associated with a variety of sources of measurement error, and that repeated measurements are required to parse out, and control for the various sources of error. Potential sources of measurement error depend on the measurement situation, but can include (a) which test item was used to assess the construct (e.g., which measurement items are chosen to evaluate dynamic constructs), (b) which rater scored the test item (e.g., which supervision officer completed the assessment), and (c) which testing occasion was scored (e.g., which re-assessment was used). Combining assessments across items, raters, and occasions is an important way to reduce measurement error, and to attain more accurate estimates of the construct being measured.

Thus, aggregating multiple assessments is expected to enhance actuarial prediction because an average of combined ratings removes the variance of different situations (measurement occasions). This allows a score to represent only the component of an individual's behavior that is consistent across situations. Measurement aggregation was the key strategy that resolved early debates about whether personality constructs had value for understanding individuals' behaviors (Epstein, 1979, 1983). This body of research demonstrated that a single personality rating poorly predicts a single instance of behavior, but prediction accuracy increases substantially when multiple ratings are averaged together.

This dissertation examines the value of re-assessment, and Chapter 5 presents analyses that test both hypotheses about why re-assessment might be valuable, by examining the relationship between recidivism and (a) the most

proximal assessment, and (b) averages of multiple assessment occasions across time. Although Hypotheses 1.1 and 1.2 are stated in a way that suggests these are competing explanations, it is also reasonable to hypothesize that both proximity and averaging have mutual effects on prediction accuracy. In this case, it is expected that prediction will increase with larger averages of *proximal* assessments, such that prediction will reach an apex with greater averaging until the averaging strategy begins to degrade, once averages start incorporating assessments that are too distal to be meaningful. It is important to note that because there is little prior work examining the prediction effects of re-assessment, the research in this dissertation is largely exploratory.

Prior to presenting the analyses that test the core hypotheses of this dissertation in Chapter 5, Chapter 2 describes the analytic strategy, Chapter 3 describes the data, and Chapter 4 presents the basic statistical models that justify further exploration of these hypotheses.

Chapter 2: Choosing the Analytic Approach

There are a number of research design and modeling issues a researcher must consider when attempting to statistically analyze change across multiple re-assessments (Lloyd & Serin, 2014), for example, (a) measurement reliability, and stability of construct validity, (b) timing and uniformity of assessment schedules, (c) shape of development across time, (d) heterogeneity of change patterns among participant sub-groups, and (e) accounting for bi-directional, mutually evolving processes. Andrews (1995) outlined a field-wide, multi-decade research agenda when he wrote: "Overall, exploration of the predictive validities of assessments of change remains a major issue and is, perhaps, the major issue for the development of theory and practice in the psychology of crime." (p. 45). Fourteen years later, Brown and colleagues (2009) argued that establishing a gold-standard statistical approach for analyzing offender change remains a compelling issue that still needs resolution. A few years later, Michel and colleagues (2013) stated there is still no consensus in the field regarding the best statistical approach for evaluating dynamic risk factors.

A single gold standard may be difficult to define because a corrections researcher with longitudinal re-assessment data has a large suite of statistical models from which to choose. Each potential model answers a slightly different research question, and it is a respectable strategy to combine approaches to build knowledge about change, as some have advocated and effectively demonstrated (e.g., Brown et al., 2009). Certainly, it lends credibility to a research conclusion when

many modeling approaches reveal the same finding, or can distinguish the heterogeneous features of what amounts to multifaceted change process.

At the same time, although each statistical model has its own purposes and advantages, the demarcation between various longitudinal models has been quickly disappearing over time as software packages evolve; it has become easier to program more flexibility into a model, and provide options that were not traditionally available before (Preacher, Wichman, MacCallum, & Briggs, 2008). In essence, each statistical approach was developed within particular research traditions with their own concerns and conventions, but greater multi-disciplinary communication has allowed for unprecedented cross-pollination, such that many analytic approaches are beginning to converge (Preacher et al., 2008). For example, some statistical models (e.g., multilevel modeling and growth curve analyses) provide identical results under certain conditions (e.g., see Willett & Bub, 2004), underscoring the similarities underlying what appears, at first, to be an extensive list of analytic options.

This chapter builds support for the statistical approach adopted in this dissertation. First, this chapter briefly describes six statistical models (or classes of statistical models) that criminologists and correctional psychologists have used to analyze re-assessment data (defined as three or more measurements per individual), focusing on the advantages and disadvantages of each. These discussions are not meant to be exhaustive. Instead, these sections are designed to provide interested readers with enough of a conceptual overview of each technique to decide if it could

be appropriately applied with their re-assessment data, and if the technique would directly and readily answer their core research questions.

This discussion helps justify why Cox regression with time-varying predictors was chosen to model the data used in this project. This chapter specifies the features that make this model generally attractive for analyzing longitudinal re-entry data, and explain why these features were particularly appropriate for investigating the hypothesis presented in this dissertation. The final sections of this chapter also introduce various practical issues involved in building Cox regression models with longitudinal data, and present cautions about over-interpreting results from these models. Finally, this chapter presents two measures of effect size that are appropriate for Cox regression models, and are recommended for use.

Important features of each statistical model are displayed in Table 2.1, which lists the advantages that make each type of model particularly attractive for longitudinal analysis, depending on the features of the dataset, and the researcher's goals. Note, however, that the models listed in Table 2.1 can often be (but not always) adapted to accommodate additional features, as required. As stated above, many of these models were developed (or are typically used) in a way that is more narrow than necessary. The sections below provide references for further reading about these nuances.

Choice of Statistical Models

A critical feature of re-assessment data is lack of independence (because measures taken twice from the same individual are more closely related than two measures from separate individuals). As a result, commonly used statistical models

Table 2.1

Longitudinal Statistical Models and the Features They Can Accommodate

Features the Statistical Model Can Readily Accommodate	Predictors					
	Multilevel Modeling	Growth Curve Analyses	Generalized Estimating Equations	Latent Class Trajectory Modeling	Growth Trajectory	Mixture Survival Analysis
Large number of within-individual assessments	✓				✓	✓
Missing data or uneven measurement schedules	✓		✓		✓	✓
Describe patterns of change	✓	✓		✓		
Describe mutually evolving processes		✓	✓			
Non-random drop-out prior to end of follow-up					✓	✓

are inappropriate for re-assessment data (Fitzmaurice & Ravichandran, 2008; Singer & Willet, 2003). Instead, as described in the sections below, regression-based models have been adapted in multiple ways to account for non-independence. However, it is generally true that models designed to appropriately account for within-individual change (e.g., growth models) are a distinct class of models from those designed for the purpose of predicting events (e.g., logistic regression). In many ways, these two goals do not naturally fit into a single analytic approach, because there are statistical challenges associated with combining change analyses (longitudinal re-assessment) with prediction models (estimating the likelihood of one binary outcome). None of the models described below manage to easily or fully blend both goals into one approach, but despite this, each has its own advantages for researchers attempting to estimate the association between re-entry re-assessment data with recidivism outcomes.

Multilevel Modeling

Broadly, multilevel models are designed to account for data that are clustered into groups. For longitudinal analyses (see Hox, 2010; Singer & Willett, 2003), within-individual re-assessments are logically clustered within individuals. Longitudinal multilevel models incorporate *time* as an explicitly modeled variable that allows the researcher to describe patterns of change (i.e., the direction and rate of change in scores across time), and describe how participants may differ from each other in change across time. For researchers examining dynamic risk, multilevel models are poised to answer the following questions.

- What was the average rate of change in dynamic risk scores across time?

- Did individual offenders significantly vary in their average risk score during the follow-up?
- Did individual offenders significantly vary in the degree to which their dynamic risk scores changed over time?
- If there was variation in change across time, did features of individual offenders explain which offenders changed at a different rate than others? For example, did older offenders' risk scores decrease more rapidly than younger offenders? Were recidivists observed to show increases in risk scores whereas non-recidivists showed decreases?

Multilevel models are characterized by two important advantages: (a) there is no upper limit to how many times a participant can be assessed, and (b) participants who missed an assessment do not have to be removed from analysis, which means researchers do not necessarily have to impute missing data, and may have an unbalanced assessment schedule (i.e., research designs where it was never intended to measure all participants on the same schedule). Still, multilevel models more heavily weight information that comes from participants with more assessments stretched across longer periods of time, and from sequences of assessments that occurred within the typical time frame of the study (i.e., a participant with only two assessments, which were recorded several months after most individuals were no longer observed, will not be weighted heavily in the statistical solution). Also, for some research questions, multilevel modeling offers the important (and unique) advantage that several levels of clustering can be modeled; for example, multilevel modeling is readily adapted to examine within-

individual changes (Level 1) within participants (Level 2) who were assigned to different rehabilitation groups (Level 3).

On the other hand, multilevel models cannot readily incorporate two simultaneous change processes (for example, by addressing how changes in risk might have impacted later, or simultaneous, changes in protective factors). However, if it is legitimate for a researcher to instead conceptualize the mutual change process as an interaction between two variables (for example, by hypothesizing that protective factors increased across time at a different rate, depending upon how individuals were rated on risk), multilevel modeling is an appropriate strategy to investigate the interaction.

There is another disadvantage inherent within multilevel models that potentially make them a limited choice for re-entry data. Although multilevel models allow unbalanced measurement schedules and missing data, these models assume that missing data are missing at random. This assumption is untenable when it is a defining feature of the dataset that each series of re-entry assessments ceases at the time of recidivism or revocation. Multilevel models do not readily account for systematic truncation, and including *recidivism status* as a person-level (Level 2) variable can be best justified under the condition that all participants' re-assessments (Level 1) were recorded *only* during a period of time prior to *any* occurrence of recidivism in the sample.

Several authors have recently demonstrated using multilevel models with correctional or criminological datasets, including Babchishin (2013), Hanby (2013), and Orrick (2012).

Growth Curve Analyses

Growth curve analyses (see Preacher et al., 2008) represent a special case of structural equation models (SEM). Broadly, structural equation models provide a highly flexible framework that can be adapted to test a seemingly endless variety of models, because they are designed to use a simplified set of structural paths to represent the relationships between many observed variables. For longitudinal data, growth curve modeling adapts SEM by characterizing scores from each time point as indicators of an underlying latent pattern of change. Growth curve models describe these latent constructs (i.e., participants' average scores during the follow-up, and average change in scores across time), and can additionally be used to examine differences in change patterns among participants. Thus, growth curve analyses can be used to answer all of the bullet-point questions posed in the section immediately above, and for some datasets, multilevel models and growth curve analyses can be viewed as interchangeable approaches for addressing the same research questions. However, growth curve analyses are characterized by some unique advantages and disadvantages.

In terms of advantages, compared to multilevel modeling, growth curve analyses can more readily answer complex questions about (a) how changes in one variable may be related to later outcomes (i.e., recidivism), and (b) mutually evolving change processes. In particular, it is relatively straightforward to adapt growth curve analyses into a *parallel process model*, which allow researchers to examine hypotheses about how changes in one variable may be related to simultaneous (or later) changes in another variable (see Little, Preacher, Selig, &

Card, 2007; McArdle, 2009; McArdle & Hamagami, 2004). For examining whether proximal re-assessments predict outcomes after controlling for earlier assessments, cross-lagged models can be used to estimate the relationship between recidivism and a risk score taken at Time 3 while accounting for the predictive effect of risk scores measured at Times 1 & 2, as well as other predictors in the model. Thus, for researchers examining dynamic risk, complex growth curve analyses are poised to answer the following questions.

- Were changes in dynamic risk related to subsequent outcomes, or other variables that were measured after change?
- Were changes in dynamic risk related to contemporaneous changes in protective factors?

These compellingly complex and flexible features of growth curve models are unfortunately tempered by some important disadvantages. In particular, growth curve models are best applied to data collected using a panel design (i.e., when measurement occasions were balanced across participants) with relatively few time points. In terms of number of assessments, it is typically seen as ideal when models can include more than four or five assessments per individual (which is similarly recommended when using multilevel modeling). Although there is not a strict upper limit to the number of assessments each participant may have recorded in the dataset, growth curve models may not easily converge when assessments exceed ten to twelve measurements per individual, and will certainly have computational difficulties at very high numbers of assessments.

A second reason growth curve models are typically used with datasets that include only a small number of re-assessments is because these models require that all participants were assessed on the same schedule, with no missing data; it becomes increasingly problematic to have a complete dataset as the number of measurement occasions increase. Whereas incomplete data and scattered measurement occasions are not problematic for multilevel models, growth curve models require both that no data are missing, and that any imputed data points were missing at random. This limitation arises from the way *time* is imbedded into the growth curve model, and because growth curve models treat each measurement occasion as a separate observed variable.

Thus, it is highly likely that researchers with re-entry data would first need to conduct data imputation in order to retain all participants for growth curve analyses. Advances in data imputation strategies make this task relatively straightforward and rigorous, but for re-entry data, it is unlikely that missing data points can be considered missing at random. In the same way as discussed in the section above, growth curve models are not designed to account for systematic truncation, and cannot account for individuals dropping out of the dataset due to recidivism. This limitation makes growth curve models potentially inappropriate for re-entry data, and most appropriate for data collected during a period when participants were not at risk for recidivism, or when measurement continued on schedule even after recidivism had occurred. This could include study designs where repeated instances of recidivism events were recorded for each participant,

such as long-term, multi-year studies where one of the time-dependent variables is defined as *number of arrests per year*.

Some authors that have used complex growth curve modeling with correctional or criminological datasets include Chen (2010) who fit a parallel process model, and Mulvey and colleagues (2006) who fit a cross-lagged model.

Generalized Estimating Equations

Generalized estimating equations (see Burton, Gurrin, & Sly, 1998; Ghisletta & Spini, 2004; Hanley, Negassa, Edwardes, & Forrester, 2003; Lipsitz & Fitzmaurice, 2009) represent an analytic *approach* (rather than a *model*, per se) that extends regression models to account for clustered, correlated data. Depending on the researcher's goals, the generalized estimating equations (GEE) approach may be particularly attractive for a re-assessment dataset because it can explicitly model the *effects* of change. For example, a researcher can use GEE to test whether increases in a risk factor were associated with subsequent increases in the likelihood of recidivism outcome, across all time periods. Thus, GEE examines two mutually occurring processes, and is especially useful when one process is best described using a dichotomous indicator (or a count variable). As such, GEE is unique in its ability to incorporate binary longitudinal data (i.e., presence/absence of criminal behavior recorded in multiple time periods); this can be applied to correctional datasets where participants were at risk for recidivism during each defined time period within the follow-up, and remained at risk for another recidivism event even following prior recidivism events. For researchers examining

dynamic risk, generalized estimating equations are appropriate for answering the following question.

- Across the sample of offenders, were observed increases in risk factors related to a higher proportion of recidivism events in subsequent time periods?

Generalized estimating equations are a flexible approach and can be adapted to account for unbalanced datasets that include participants with large numbers of re-assessments. However, the features that make GEE a compelling statistical approach will best emerge when a researcher has a dataset with (a) a large number of participants that each have (b) a relatively small number of assessments that (c) were recorded on the same schedule.

Also, GEE represents an attractive approach mostly when all individuals in the dataset were observed to change in generally the same way across time. In other words, GEE does not readily account for sub-populations that exhibited different patterns of change (Liang & Zeger, 1986), unless GEE is adapted for this purpose. Thus, unlike multilevel modeling and growth curve analyses, GEE generally does not explicitly allow for or examine variations in participants' patterns of change. Instead, GEE only takes into account within-individual clustering of re-assessments in order to more accurately model what is considered the primary concern of the approach: the average relationship between two variables within the full sample, across individuals, and ignoring intra-individual changes. In essence, GEE considers it a nuisance that data are within-individual re-assessments, given that the purpose is to estimate the degree to which a predictor

variable is related to recidivism, or the degree to which *average* changes in a predictor variable are related to *rates* of recidivism, across the whole sample. In addition, GEE must be adapted to account for the (likely) possibility that outcomes in one time period were related to predictor ratings in a later time period (i.e., a recidivism event observed in Time 2 would reasonably affect how a rater would score risk factors in Time 3).

Finally, and most important, similar to the two models described in the sections above, GEE does not readily account for data that are missing for systematic reasons. Instead, the GEE approach assumes that data are missing completely at random (i.e., patterns of missing data are unrelated to what would have been observed, or values previously observed in the dataset). As stated before, this is a particularly problematic assumption with most re-entry datasets, in which assessments typically ceased at the time of recidivism. However, it is possible to adapt GEE to attempt to account for systematic dropout from observation, using an approach that re-weights cases based on risk of dropout.

Despite these numerous constraints, GEE offers a potentially attractive approach for analyzing re-entry data, given that GEE combines both the modeling of change and logistic-based prediction, thus blending two important analytic goals into a single model. There are few examples of GEE with data from correctional psychology, but Blanchard (2013) and Michel and colleagues (2013) have demonstrated its use with dynamic risk assessment data.

Latent Class Trajectory Modeling

In contrast to GEE approaches (which ignore how participants' scores in the sample may change at different rates, or in different directions), the main purpose of latent class trajectory modeling (also called *growth mixture modeling*; see Muthén & Asparouhov, 2009) is to explicitly model unique subgroups in the data that share different patterns of change across re-assessment data (i.e., shared trajectories). Although multilevel modeling and growth curve analyses model how individuals' scores may vary around the mean intercept (often defined as the baseline score, or average score during the follow-up), and how individuals' patterns of scores may vary from the average pattern of change, these two approaches constrain all change patterns to follow the same shape. By contrast, latent class trajectory modeling extends simple growth curve models to allow researchers to identify distinct groups with distinctly shaped (and positioned) trajectories across time. Membership in these subgroups is not pre-defined, but is empirically derived, based on how individuals' distributions (trajectories of change) share similar parameters (level and shape) with others in the sample.

For example, this approach could derive four subgroups from a re-entry dataset, each identified as having (a) consistently low risk scores throughout the follow-up, (b) consistently high scores throughout the follow-up, (c) gradually declining risk scores across the follow-up, or (d) rapidly declining risk scores across the follow-up. Whereas simple growth curve analyses quantify differences among participants on a continuous scale (i.e., to what degree is increased age related to more quickly declining risk scores across time?), latent class trajectory modeling

identifies individual differences by using discrete groups. Thus, this approach models how participants may differ from each other not just in degree, but in kind, as expressed in the following research questions.

- Within the sample, was there evidence to suggest changes in dynamic risk followed unique trajectories? Can discrete subgroups of offenders be identified, based on different patterns of change across time that emerged from the data?
- If unique trajectory subgroups were identified, did these groups significantly differ from each other in their individual features, and/or outcomes?

There are a number of compelling reasons for researchers to consider whether unique trajectories may be present within their re-entry re-assessment data. Logically, if this hypothesis is tested, but no substantially different trajectories emerge, a researcher has empirical justification for using more simplified models that assume patterns of change among all participants follow a similar shape. On the other hand, if different trajectories emerge, a researcher has discovered an important source of variation within the data that should be explicitly modeled, either with complex approaches, or by separating out and examining each subgroup with less complex modeling approaches.

Overall, it is probable that discrete trajectories may be identified in re-entry re-assessment data, depending on the dataset's sample, time frame, and measurement schedule. Theoretically, a population of offenders will at the very least include both recidivists and non-recidivists (at least one latent class of each),

who are hypothesized to generally exhibit patterns of stably high and declining risk, respectively. Also, visually, inspection of re-entry re-assessment data typically tends to show a finite number of patterns that may be empirically distinct from each other.

Despite these compelling reasons, latent class trajectory modeling is arguably best used for the one task it is designed to address (i.e., identifying trajectories), and shares many of the same limitations of growth curve analyses (the model it extends), specifically: (a) problems with model convergence when number of assessments is large, (b) assumption that missing data are not missing for systematic reasons, and (c) restricted to predicting only outcomes that occurred following completion of re-assessment data collection, unless participants are at risk for recidivism during every time period, and outcomes are treated as a time-dependent variable. In terms of this last limitation, there are two ways latent class trajectory modeling can combine the tasks of analyzing change and prediction outcomes. As implied in the limitation, this approach can be used to ask, for example: (a) were discrete trajectories of change identified during the first three months of re-entry related to recidivism outcomes in the subsequent three months?, or (b) were discrete trajectories of outcomes (i.e., a group with high rates of revocations across three years vs. a group with moderate and declining rates of revocations) associated with significantly different risk scores?

Researchers in criminology have widely embraced latent class trajectory methods to specifically address how individuals' patterns of criminal behavior change across time (see Nagin & Land, 1993, for early development and application

of the technique, and Piquero, 2008, for a review of existing studies). This approach has typically used officially recorded or self-reported rates of criminal acts that were gathered over long periods of time (i.e., many years or decades), but has rarely examined predictor variables that are of interest for short-term correctional management (i.e., dynamic risk factors). Although this body of research provides the broader life-course context for understanding the shorter-term change that occurs during re-entry, the substantive questions are different. In correctional psychology, there are few examples of researchers examining latent trajectories, but Quinsey and colleagues (2006) represent a rare exception; these authors demonstrated this technique with dynamic risk re-entry data. For general recidivism outcomes, their results suggested a two-group solution (high risk and low risk), and, for violent recidivism outcomes, a one-group solution.

Growth Trajectory Mixture Survival Analysis

Growth trajectory mixture discrete-time survival analysis (see Muthén & Masyn, 2005) represents a unique approach for incorporating both change analyses and outcome prediction within the same statistical model, by blending two analytic traditions. Mixture modeling, the first component of the analysis, is exactly as described in the section immediately above, which empirically separates longitudinal data into sub-populations, or trajectory groups. The second component of this analysis relies on traditional Cox regression survival analysis to test the association between change trajectories and time to recidivism. In essence, this approach first (a) derives knowledge of unique trajectories from the data, and then (b) uses an indicator of participants' group membership (associated with a discrete

pattern of change) as a predictor of subsequent outcome events, while taking into account time at risk.

As a solution to the “*change or prediction, but not both simultaneously*” dichotomy, merging two gold-standard statistical models with separate purposes into one “super-model” represents an intriguing approach. As such, this model gains many advantages, albeit with one critical caveat: the model is designed to separate analysis of change (using mixture modeling) and prediction of outcomes (using Cox regression) into different time frames. Thus, during the first time frame, the model explicitly examines trajectories of change; at this time, measurement must be balanced, with no missing data, nor a high number of assessments, nor systematic dropout. Then, in the second time frame, the model gains the advantages of Cox regression, where a large number of assessments can be used from an unbalanced schedule, and systematic dropout can be explicitly modeled. (See the next section for more detailed discussion of these advantages.) Thus, this approach requires that processes of change precede the time period when participants are at risk for recidivism. For examining re-entry, this is a potentially critical limitation; however, this approach would represent an appropriate strategy for the following research questions.

- While incarcerated individuals attended rehabilitation, were they observed to show discrete patterns of change in antisocial attitudes throughout program attendance, and were these discrete patterns related to recidivism, after release?

- Were sub-groups of re-entering individuals observed to show discrete patterns of changes in mood during the first two months following release, and, among individuals who remained in the community for more than two months, were these patterns of change related to subsequent recidivism during the rest of the follow-up period?

Growth trajectory mixture survival analysis has been applied outside of criminology and correctional psychology (such as within education research, see Muthén & Masyn, 2005, for examples), but a search for how this technique may have been applied with data from offender samples revealed no currently existing studies.

Cox Regression Survival Analysis with Time-varying Predictors

Cox regression survival analysis (see Singer & Willett, 2003; Willett & Singer, 1991, 1993; Willett, Singer, & Martin, 1998) is a widely used statistical model within correctional psychology, and represents a gold-standard approach for modeling the association between predictor variables and subsequent dichotomous outcomes. This approach is uniquely designed to predict events that terminate participation in the study (i.e., individuals are removed from the study at the point they experience recidivism), while still retaining all information collected from participants prior to this termination *and* retaining all information from individuals who did not experience the predicted outcome (i.e., those who stayed crime-free throughout the follow-up). For re-entry datasets, this feature represents a substantial advantage, because the model explicitly incorporates a critical aspect of the phenomenon of interest (i.e., ongoing management of offenders until some are removed from the

community due to recidivism, while others cease being observed only due to research design).

Modeling re-assessments, rather than modeling change.

The five approaches described in the sections above focus primarily on describing change, but can be adapted to examine recidivism outcomes. By contrast, basic Cox regression is at its core a prediction model that can be readily adapted to include re-assessments throughout the follow-up period (i.e., the computational formulae does not require adjustments in order to incorporate re-assessments). In essence, the model continually updates each individual's predictor score as assessments occur across time, always using the timeliest information in its estimate of the association between the predictor and outcome. The Cox model is unique for its ability to examine short-term associations between predictors and outcomes, with the definition of "short-term" tied to how the researcher decides to set the time frame for prediction (see discussion below; Dekker, Mutsert, van Dijk, Zoccali, & Jager, 2008). Unlike growth trajectory mixture survival analysis, Cox regression with time-varying predictors blends the modeling of re-assessments into the same time period during which it simultaneously predicts and observes outcomes.

Despite readily including re-assessments, Cox regression emphasizes the task of prediction over describing change processes, and does not explicitly describe how participants changed across time. Making this distinction between a *change* model and a *re-assessment* model may appear esoteric, but this nuance has critical importance when considering what conclusions can be drawn from model results.

Unlike multilevel modeling or growth curve analyses, the inherent design of Cox regression does not allow researchers to draw conclusions about how individuals changed, test whether participants differed in their patterns of change, nor examine how change processes were related to outcomes. Further, the model is also limited in that it cannot include mutually evolving change processes (except in the form of interaction effects).

These limitations result from the way each individual's sequence of scores is only *implicitly* retained in the Cox model, yet these sequences are neither described nor examined as the unit of analysis. Instead, the Cox model uses knowledge about the ordering of scores within each sequence to pull out and analyze information about the *timing* of assessments (and how these were related to observed events). In this way, the model is constrained to focus on the relationship between outcomes and proximal scores, rather than on the relationship between outcomes and individuals (and their patterns of change¹).

This distinction introduces an important benefit for a prediction model, because the core analytic question changes from “*who* is at risk?” to “*when* is an individual at risk?” As a re-assessment model (not a change model), Cox regression assumes that the association between a predictor and outcome remains constant, and that individuals are sometimes characterized by that predictor variable, and sometimes not. For example, individuals can change membership from one group to

¹ A researcher may employ a Cox model to investigate the association between change processes and recidivism events by building time-varying predictor variables that express the amount of observed change from assessment to re-assessment (i.e., change scores). Still, overall, Cox regression has limited value as a model for describing change because it is not directly equipped to examine patterns of growth across multiple scores (and, a pattern across three or more scores is necessary for informative modeling of change).

another (e.g., a re-entering offender may obtain employment three months after release, but become unemployed one month after that), while the effect of group membership can stay the same (e.g., employed status is always related to reduced likelihood of recidivism). Thus, Cox regression with time-varying predictors draws conclusions about *scores* and outcomes, rather than *individuals* and outcomes, but accomplishes this by incorporating information about within-individual change to make more precise estimates of a predictor's association with outcome.

Although the Cox model addresses *when* individuals are at risk, rather than *which* individuals are at risk, this does not imply that knowledge of participant-specific patterns of change is unimportant when working with Cox regression models. Rather, Chapter 4 describes and displays a number of descriptive and diagnostic approaches that a researcher should employ to both justify using Cox regression models, and to assist in the interpretation of model results, by providing the broader context about *who* in the sample is changing, and *how*. As stated earlier, understanding change processes is a complex endeavor that requires (at least) several modeling approaches.

Model advantages.

The Cox regression model's emphasis on the timing of assessments (rather than change across assessments) makes it a particularly appropriate model for addressing the hypothesis proposed in this dissertation. Among the longitudinal models reviewed in this chapter, Cox regression is unique for allowing researchers to draw conclusions about the relative importance of assessment proximity. This is evident when considering how Cox regression with either time-varying or time-

constant variables is the exact same computational model. Thus, the model can easily be used to compare an approach that updates scores across time to an approach that does not update scores (i.e., does including re-assessments improve model accuracy?).

As stated above, a critical advantage of the Cox model is its ability to analyze data characterized by systematic, meaningful dropout. An additional advantage is that there is no upper limit to how many re-assessments can be incorporated; Cox regression can readily incorporate a large number of data points per participant. Also, it is not required that assessment schedules be balanced across participants. These features free a researcher from controlling for individual differences in how many assessments were recorded, or imputing missing data to fit a pre-specified assessment schedule.

Addressing the timing of assessments.

Although unbalanced assessment designs do not pose problems for the model, a researcher must still consider the assessment schedule when preparing a dataset for analysis with Cox regression. The temporal spacing of assessments is important because this approach is designed to model the association between assessment timing and events. This means researchers must concern themselves with *when* assessments were relevant during the follow-up period, and at what point these assessments ceased to be accurate reflections of participants' current states. Especially for the purpose of addressing the hypothesis in this dissertation, it was necessary to explicitly define what amount of time would be considered "close proximity" between assessments and events. (See Chapter 3 for further discussion

of the decision criterion used for this project.) In other words, Cox regression with time-varying predictors requires researchers to in some way address the age of their data (de Bruijne, le Cessie, Kluin-Nelemans, & van Houwelingen, 2001), either by (a) using logic to determine how the data will be placed into temporal sequence across the follow-up period, (b) explicitly exploring multiple approaches to how the data could be placed into the follow-up period, by comparing different models, or (c) specifying time-varying predictors that quantify the age of other time-varying predictors.

Further, an even more restrictive feature of Cox regression models is the requirement that at the time of *every observed recidivism event*, both recidivists and non-recidivists remaining in the dataset must display an observed score on the predictor(s) in the model. This is required because the model evaluates whether recidivists' scores at the time of recidivism differ systematically from the scores of non-recidivists, who must be present in the dataset (i.e., not yet censored) and provide scores for comparison (i.e., were observed on predictors of interest). Thus, although the Cox regression approach readily allows participants to be inconsistently observed throughout the follow-up period (and, in fact, the model technically ignores any data drawn from periods when no participants experienced recidivism), scores must be recorded on a schedule that is frequent enough to logically extrapolate an appropriate value on the predictor for all participants at every point of recidivism.

In practice, this constraint is often addressed by placing data into discrete time periods. A researcher may choose how to define the units of time in which the

follow-up period is measured (depending on the precision of the actual data collection schedule); Cox regression can be used to predict events that occur across milliseconds in the same way it can be used to predict events across decades. When assessment and event data are collected in small units of time (days), combining the data into larger discrete periods (weeks or months) can typically be justified, depending on the researcher's goals. If a researcher chooses to segment the follow-up period using larger units of time based on how frequently data were actually observed for each participant, and how frequently events were recorded, she will logically be able to ensure that all participants were rated on predictor variables in every period of observed recidivism.

Note, however, that when each participant's follow-up period is divided into exactly the same discrete time intervals (i.e., all time periods are defined equally across all participants), the model can be analyzed using a strategy that is no longer *technically* a Cox regression model. The main implication of this difference (both practically and mathematically) is that the equally spaced, discrete-time model requires the researcher to explicitly choose and model a variable that represents *time*, whereas a true Cox model does not include a specification of time. (This choice is discussed in Chapter 4.)

Although the Cox model has a certain computational advantage, because it is allowed to ignore the estimation of one (or sometimes more) parameter(s), there are numerous practical advantages that result from choosing a non-Cox approach (named the *discrete-time hazard model*). First, the discrete-time hazard model is a variant of the logistic regression model (adapted to have multiple intercepts rather

than one intercept), and can thus be analyzed using standard logistic regression software (Willett & Singer, 1991). This is an advantage because (a) researchers tend to be familiar with employing logistic regression approaches and interpreting their coefficients, and (b) some software packages require a relatively burdensome amount of syntax programming to fit a Cox regression model with time-varying predictors.

For practical purposes, choosing a discrete-time hazard model requires researchers to organize their data into a *person-period format* (see an example in the Appendix). For computational purposes, a researcher who chooses the logistic approach is also choosing to specify the logit transformation as the link function in the formulae. Note that other link functions can be chosen, but in correctional psychology, the logit transformation is likely to be an appropriate choice.²

A second important practical advantage to choosing the discrete-time hazard model is that violations of model assumptions can be more easily tested and readily remediated in this framework, compared to a true Cox approach (Willett & Singer, 1993). This is advantageous because model violations are common within observed data, and researchers should arguably begin analysis assuming that it will likely be necessary to include additional parameters in final models to account for some model violations. (See Chapter 4 for a discussion of diagnostic and remediation approaches.) Model diagnostics in a discrete-time framework are straightforward,

² Singer and Willett (2003) recommend that researchers consider the complementary log-log transformation when hazard rates are high in some time periods. However, it would be highly unusual for re-entry data to display hazard rates that are high enough for differences to emerge between models using these different link functions.

but would be unwieldy (in terms of both cognitive strain and software programming) when working with a true Cox model.

Thus, there are important benefits if a researcher can decide it is legitimate to structure the observed assessments into discrete time periods that are equal for all participants. For the project in this dissertation, equally spaced discrete time periods were chosen (i.e., weeks). Although the underlying computation is slightly different, the Cox regression and the discrete-time hazard models are essentially identical under these conditions, and, of course, results were essentially identical when models in this dissertation were run using either approach.

In summary, whether spacing time periods equally or individually, researchers building time-varying survival models must make decisions about whether it is reasonable to assign a validly observed predictor score to each participant at the time of every recidivism event (or, impute scores during periods when participants were not observed in actuality). In practical terms, a true Cox model requires imputation only at times of observed recidivism, but in the special case of the Cox model where discrete periods are equally spaced, score assignment or imputation is required during every discrete time period that occurred prior to recidivism or censoring. Choice regarding how to impute these scores can be addressed through an empirical, data-driven decision-making process, as described in more detail in Chapters 3 and 4.

Examining risk associations, rather than conducting risk assessment.

The above discussion of the basic features of Cox regression serves to specify what conclusions can be drawn from model coefficients, and what conclusions

would be inappropriate (i.e., using Cox regression to state that a pattern of change was related to recidivism). To make two broader points about the practical meaning that can be derived from these models, the subsequent section will discuss effect size estimates for Cox models, while this section briefly highlights the important difference between two often-confused activities in correctional psychology: (a) building prediction-focused statistical models, and (b) establishing accurate risk assessment.

A recently emerging body of research has put fine-tuned focus on (what has turned out to be) a critical difference between (a) finding an association between measures of risk factors and recidivism outcomes, and (b) applying knowledge of that statistical relationship to make reliable, consistent predictions about expected rates of recidivism across samples, jurisdictions, etc. In essence, it is relatively simple to find an association between test scores and outcomes, and a formidable research enterprise to accurately match an expected likelihood of an outcome (derived from prior research with test scores) to an observed outcome rate in a future sample (Hart, 1999).

On the one hand, many authors (e.g., Campbell et al., 2009; Coid et al., 2011; Desmarais & Singh, 2013; Skeem & Monahan, 2011) have concluded that all risk instruments reach similar levels of moderate prediction accuracy. Results from these studies are best interpreted to suggest that all risk instruments share an ability to accurately rank offenders along a continuum of reoffending likelihood. Yet, actual (absolute) rates of recidivism outcomes vary to an important degree from sample to sample (Fazel, Singh, Doll, & Grann, 2012; Singh, Fazel, Gueorguieva, &

Buchanan, 2014), such that participants from different samples sharing the same scores on the same risk instruments are not guaranteed to exhibit the same rates of recidivism. Thus, whereas a risk instrument may show cross-sample stability in its relative rankings (Babchishin, Hanson, & Helmus, 2012a; Hanson, Babchishin, Helmus, & Thornton, 2013), it is unlikely that base rates and absolute recidivism rates would remain stable across samples (Hanson, 2009; Litwack, 2001; Rogers, 2000), and absolute rates have indeed been shown to be unstable (Helmus, Hanson, Thornton, Babchishin, & Harris, 2012). This is highly problematic when considering that correctional practice is brimming with risk instruments, but few implemented instruments have been examined in more than one sample (Desmarais & Singh, 2013; Schwalbe, 2007), and different studies of the same instrument often define recidivism outcome in different ways, precluding comparison (Yang et al., 2010).

These limitations do not invalidate the existing body of risk assessment research, but rather imply that researchers should replicate, extend, and evolve existing approaches to instrument validation. Furthermore, practitioners should become well versed in being cautious about giving consumers and legal decision-makers the impression of certainty. In particular, predicted outcomes from risk instruments have always been probability statements about the likelihood that an offender can appropriately be classified into an observed group with a known (calculated) average recidivism rate (Baird, 2009; Wright, Clear, & Dickson, 1984); this definition requires clarity and caution when making a decision about a single individual (Desmarais & Singh, 2013; Fazel et al., 2012; Maden, 2003; see also Hart, Michie, & Cooke, 2007).

There are at least two stages of uncertainty, however. It is uncertain that the individual likely belongs to the classification group, and it is uncertain if that group's observed recidivism rate would replicate in another group, even when these groups share key similarities. The inherent uncertainty of risk assessment does not mean it should be abandoned, but rather that all estimates should be accurately qualified (Hanson, 2009); arguably, building in explicit acknowledgement of the uncertainty when conducting risk assessment has made it a more valuable enterprise (Simon, 2005), and has served to highlight important directions of future research.

The critical knowledge gap, and a source of confusion within correctional psychology, can possibly be summarized as a general tendency to mistake *prediction* for *precision*. Prediction, which can be established using Cox regression models, is an important feature of a risk instrument. Logically, an instrument that successfully predicts has good construct validity, because it measures a latent construct that is truly associated with the outcome of interest. Prediction is the most rigorous test of construct validity because it eliminates any possibility that the outcome may have influenced how the scale items were rated.

However, although variability on a risk measure may (and should) be related to the predicted outcome, this finding merely expresses that the (imprecisely measured) latent construct and outcome have an association, such that higher scores on the scale are generally related to higher likelihood of recidivism. Although this establishes prediction, the precise nature of the association remains unknown, because the scale used to measure risk is an arbitrary metric (Blanton & Jaccard, 2006a, 2006b). By contrast, a fully non-arbitrary risk scale would be precise to the

degree that each score on the scale would be associated with its own precisely known behavioral manifestation, and/or real-world outcome.

There are a number of reasons the metrics of risk scales are inherently arbitrary. First, as with assessment of any psychological construct, there is an unknown relationship between scale items and the underlying construct that it measures. As such, it is impossible to tell if a one-point increase in a dynamic risk instrument corresponds to a “true” one-unit increase in the construct that has been labeled “dynamic risk”. This limitation occurs because the underlying construct is not directly observable, and as a result, its true scale is unknown.

It is even possible for a non-arbitrary measure to be used as an arbitrary metric for an underlying construct. As an example, suppose a researcher collects information about the number of each offender’s antisocial peers. This metric is non-arbitrary; Harold associates with three friends who have a criminal record, and three chairs would be required for each to take a seat. However, if the researcher is using this metric to measure the *influence of antisocial peers on an individual’s propensity to offend*, it is evident that the metric has now become arbitrary for this construct. If Harold were to gain a fourth antisocial peer, it is unclear if this would have no impact, a small impact, or a large impact on his underlying propensity toward, or his likelihood of criminal behavior.

Demonstrating that the metrics of risk scales have an unknown relationship with underlying propensity, Babchishin and colleagues (2012b) entered highly correlated static risk measures into prediction models and found evidence of incremental prediction. This approach effectively re-calibrated the risk metric in a

way that served to decrease measurement error and re-weight items to enhance the metric-prediction relationship (for example, by down-weighting sexual deviance items when predicting non-sexual re-offending). In addition, averaging these correlated risk measures resulted in improved estimates of absolute recidivism rates across samples, compared to scores from single measures. Thus, after combining measures, the new metric was more precisely related to observed outcomes (but still not perfectly calibrated), and this greater precision between scale scores and real-world outcomes made the scale somewhat less arbitrary.

Second, as also demonstrated by Babchishin and colleagues' (2012b) findings, most contemporary risk measures feature an additional limitation by being composite measures that atheoretically combine multiple risk factors. Measures that assess multiple latent constructs have the important advantage of being better predictors of recidivism. The critical disadvantage of this approach is that there is no longer any reason to expect that scores on a risk measure will correspond in an identifiable way with a unitary underlying construct. Implicitly, all existing risk measures are attempts to measure a highly abstract, multifaceted latent construct: *propensity toward criminal behavior*. This enterprise is doomed to only ever give rise to arbitrary metrics because each scale is designed to assess multiple latent factors that correlate with each other in unknown ways to an unknown degree. Given that it is not even expected to correspond to a single dimension, a one-unit increase on a risk scale is certainly highly arbitrary (and, thus, its true relationship with the latent phenomenon is imprecisely known, and can only be inferred from behavioral outcomes, rather than any manifestation of the phenomenon itself).

An atheoretical composite measure may not only complicate the correspondence between a risk metric and its latent construct, but also obscure it. For example, Barbaree and colleagues (Barbaree, Langton, & Blanchard, 2007; Barbaree, Langton, Blanchard, & Boer, 2008; Barbaree, Langton, Blanchard, & Cantor, 2009) demonstrated that many actuarial risk items did not predict recidivism for the expected reason. These items were associated with outcome not because they measured the psychological constructs they appeared to measure at face value, but because they were proxies for young age. In other words, their research showed that purely atheoretical selection of the best predictors of future crime automatically selects variables that are correlated with younger age, leading to a confounding of age and risk scores. When the influence of age is embedded within many actuarial items, it can appear that more latent constructs underlie propensity for criminal behavior than is actually the case.

To develop risk instruments measured with less arbitrary metrics, Hanson and colleagues (e.g., Babchishin, Hanson, & Helmus, 2012b; Hanson et al., 2013) argue that the long-term solution requires increased attention toward construct validity. In other words, it is a critical endeavor to specify the latent constructs measured within risk scales, and the observable experiences (both behavioral manifestations and outcomes) that are associated with variations on those latent constructs. When a risk measure represents a construct, rather than a composite propensity, researchers can work toward defining a less arbitrary metric, by establishing (through the consensus of replication) which observable referents signal where individuals are placed along that latent construct.

In summary, the main purpose of this section is to acknowledge where Cox regression models fit within the broader goal of successful risk assessment. When using scales with arbitrary metrics, such models are ideal for understanding theoretical relationships between constructs and outcomes (Blanton & Jaccard, 2006a, 2006b). For the purpose of the project in this dissertation, Cox regression is appropriate for testing whether there is a theoretically closer relationship between outcomes and proximal scores, compared to distal scores. Cox models helpfully (a) establish whether a risk instrument has good construct validity by demonstrating predictive validity, and (b) estimate the shared association between an underlying risk construct and observed outcomes, even if no precise meaning can be ascribed to the magnitude of this association (i.e., coefficients do not provide rigorous scale cutoffs that are useful for future forecasting across samples). Thus, results from these models are not immediately applicable in practice, without additional research.

Effect size estimates.

Within correctional psychology, researchers have largely ignored instituting an effect size for Cox regression survival analysis (with either time-invariant or time-varying predictors). Defining an effect size for Cox regression survival models is not as straightforward as directly applying effect size calculations used in more familiar, widely used statistical models. The outcome of interest in survival models is distinct because it combines both a binary outcome (i.e., did the individual recidivate?), and a continuous outcome (i.e., how many days occurred before recidivism was observed?). This feature complicates the definition of an effect size,

and requires adapting existing effect size strategies for Cox regression (Heagerty & Zheng, 2005).

For models with continuous outcomes, effect sizes are typically reported as an R^2 statistic, or as the percentage of variance within the dependent variable that is accounted for by the predictor variable (Cohen, 1988). For models with longitudinal binary outcomes, researchers typically estimate the strength of the effect in terms of prediction accuracy, by plotting true positive rates (sensitivity) against false positive error rates ($1 - \text{specificity}$) using receiver operating characteristic (ROC; Swets, Dawes, & Monahan, 2000) curves. Researchers summarize accuracy across all values of the predictor variable and report a global accuracy statistic, the area under the curve (AUC).

For Cox regression, these widely used approaches are not fully satisfactory because neither technique can adequately incorporate the longitudinal nature of survival models. In particular, directly adapting R^2 for survival models does not take censoring into account. Although follow-up time is a continuous variable, some individuals' follow-up times represent time-to-recidivism-event, whereas other follow-up times represent time-to-censoring. Similarly, ROC prediction models are arguably limited because these calculations treat the follow-up period *in toto*. In practice, prediction accuracy may vary across the follow-up period, such that a measure may better predict events early in a follow-up period compared to events later in a follow-up period. The typical approach to AUC statistics does not take into account changing, or rolling, prediction accuracy rates. This is a particularly critical limitation when utilizing Cox regression with time-varying predictors, since core

features of this model include treating both (a) the recidivism status of individuals and (b) the predictor variable as features that update continuously across the follow-up period.

Although Cox regression is commonly used within the correctional psychology literature, researchers have not yet adopted a common standard for presenting effect sizes for these models. To describe the strength of a finding, researchers typically only report the hazard ratios generated by the Cox model. The hazard ratio is not necessarily uninformative for determining the strength of an effect, because a reader can interpret the statistic as the increased likelihood that participants will experience the recidivism event, for each unit increase in the predictor variable. However, hazard ratios are problematic as a common-metric effect size because their interpretation is always tied to the scale of measurement of the predictor variable. For researchers interested in comparing the impact of different predictor measures, comparison of hazard ratios is an inadequate solution. Although there are exceptions, correctional researchers also tend to report AUC statistics only when developing risk assessment instruments. When Cox regression is employed to test the theoretical relationships between variables and recidivism, AUCs are generally omitted, and the model's prediction accuracy is rarely discussed.

It is relatively recent that statisticians have addressed the issue of effect size in survival models. Notably, none of the most widely used statistics packages have incorporated effect size options while running Cox regression. This reflects the fact that no single estimation approach has gained full consensus across disciplines, especially in the social sciences, and a variety of approaches remain under debate

and development (Schemper & Henderson, 2000). In addition, including time-varying predictors in Cox models further restricts available effect size options; only some effect size measures for Cox regression can be extended to allow time-varying predictors. For this reason, a number of effect size options are disqualified for use with the models in this dissertation (e.g., Kent & O'Quigley, 1988; Schemper & Henderson, 2000). However, the following discussion outlines two suitable options that are particularly recommended because they can be applied with all Cox models.

Measures analogous to R^2 . Xu and O'Quigley (1999; see also O'Quigley & Xu, 2001; O'Quigley, Xu, & Stare, 2005) developed an effect size measure for Cox regression that is effectively analogous to the familiar R^2 statistic in use with linear regression. Similar to an R^2 statistic, Xu and O'Quigley's (1999) measure can be interpreted as the proportion of explained randomness associated with the predictors in the model. As a proportion, when no variance is explained by the model, $R^2 = 0$, and when the chronological order of event times fully correspond to variation in predictor values, $R^2 = 1$. Intermediate values provide the degree to which predictor variance is associated with survival times, in terms of proportion of variance explained. Thus, the statistic expresses the relative importance of the predictor measure, and is useful as an effect size.

Conceptually, Xu and O'Quigley's (1999) R^2 expresses whether predictor values are ordered sensibly (versus randomly) across the series of chronological failure (event) times. When developing risk assessment measures, researchers hypothesize that higher values will be associated with earlier recidivism event times, and lower values will be associated with longer survival. If event times correspond

fully randomly with values on the risk assessment measure, the strength of the effect is poor or nonexistent.

More specifically, given the observed order in which participants experienced recidivism events across the follow-up period, Xu and O'Quigley's (1999) R^2 statistic expresses how much variation *in the predictor variable* is explained by the known order of survival times. This interpretation may appear counter-intuitive as it reverses the typical approach to an R^2 statistic, which usually expresses the amount of variation explained in a dependent variable, given an independent variable. However, there are key advantages to this approach (O'Quigley et al., 2005; Xu & O'Quigley, 1999). In particular, computation is simplified, since information routinely generated during calculation of the Cox model can be re-purposed to calculate the R^2 statistic. Second, this feature allows the inclusion of time-varying predictor variables. This flexibility is arguably a distinct advantage over a similar R^2 statistic (developed by Kent & O'Quigley, 1988), which offers the more intuitive interpretation (where R^2 = the proportion of explained randomness in failure times, given the known predictor values). However, O'Quigley et al. (2005) demonstrated there is close equivalence of the two measures when both are applied to the same Cox model. The advantage of being able to employ the same effect size measure in either time-varying or time-invariant circumstances is arguably crucial for consistency across models. For this reason, O'Quigley et al. (2005) recommended Xu and O'Quigley's (1999) version of the R^2 statistic.

Computationally, Xu and O'Quigley's (1999) R^2 statistic employs a ratio that quantifies the impact of including a predictor in the model versus a baseline (no predictor) model (O'Quigley & Xu, 2001). More specifically, observed predictor values are compared to expected values (i.e., residuals, which in the case of Cox regression are squared Schoenfeld residuals), and the resulting ratio compares the (a) average discrepancy between observed and expected values when the predictor variable is included to the (b) average discrepancy when the predictor variable is omitted. This description is fully correct when using a model not affected by censoring (i.e., all individuals experience the event during the follow-up period), but in practice, the residuals must be differentially weighted to take censoring into account.

Because the computation of the R^2 statistic compares a fitted value to an observed value, Potapov, Adler and Schmid (2015) recommend using a split-half re-sampling technique when calculating Xu and O'Quigley's measure. In other words, they recommend developing the Cox regression model using the first half of available cases, then using that model to obtain fitted values for the second half of available cases. The R^2 statistic is then calculated using the participants (and fitted values) derived from the second set of available cases. This approach provides a more conservative, but also more defensible, estimate of the R^2 statistic, since using the same data to both (a) develop the model and (b) test discrepancy values, likely artificially minimizes discrepancies.

Time-dependent measure of prediction accuracy. Heagerty and Zheng (2005) extended the typical use of receiver operating characteristic (ROC) and area

under the curve (AUC) statistics by developing time-dependent calculations of sensitivity and specificity. Their approach involves generating a series of time-specific AUC statistics that are updated as event information evolves across the follow-up period. One important appeal of this extension is that this approach models changes in predictor accuracy across time. For example, time-specific AUC statistics may degrade across the follow-up period, such that the risk measure better predicts recidivism event times that occur early after baseline compared to later in the follow-up period.

In addition to graphing changes in AUC across time, a single summary statistic can be calculated that is interpreted as the weighted average of time-specific AUC statistics across the follow-up period. This summary statistic is closely equivalent to the *concordance index* (also called Harrell's *c*, the *c*-index, or simply, *c*; Harrell, Califf, Pryor, Lee, & Rosati, 1982), an effect size indicator that researchers in the medical sciences widely use and report. As a weighted average, the *c*-index can be interpreted as an AUC statistic that is adjusted to take into account follow-up times. Specifically, like an AUC, the *c*-index is the probability that any two randomly selected individuals were correctly ordered on the predictor measure (i.e., that a recidivist was rated as higher risk than a non-recidivist on the risk scale), but qualifies this by clarifying that the two randomly selected individuals were only compared if they survived without recidivism or censoring *up to that specified point in the follow-up time* (see also Harrell, 2001; Harrell, Lee, & Mark, 1996; Kattan, 2003). In other words, the predictor values of two randomly selected individuals are compared only if these individuals survived long enough to enter together into

the risk set at that point in time (Heagerty & Zheng, 2005). Note that Heagerty and Zhang's summary statistic is different from Harrell's c only in how each pair of cases is weighted (Saha, 2009). In practice, this difference is unlikely to be noticed, and both statistics are named *concordance* indices.

Thus, a time-dependent AUC has a slightly different interpretation compared to a typical AUC, when baseline assessments are used to predict recidivism events without controlling for follow-up time. As a result, time-dependent AUC statistics better reflect the longitudinal nature of Cox regression models by treating a measure's prediction accuracy as an element that continually updates across the follow-up time. Further, this extension of AUC is appealing in its ability to accommodate time-varying predictor variables. Because time-dependent AUC statistics allow event information to update throughout the follow-up period, they can also naturally incorporate predictor values that similarly update across time. When time-varying measures are used, the global accuracy summary AUC value is specifically interpreted as the weighted average of time-specific AUC statistics, calculated by using continuously updated event *and* predictor information.

More formally, Heagerty and Zheng's (2005) time-dependent ROC statistics are based on *incident/dynamic* definitions of sensitivity and specificity. These terms refer to how participant information (cases and controls) is updated across the follow-up time. The term *cases* refers to participants who experienced a recidivism event at some point during the follow-up period, whereas *controls* are participants who were censored (or, were not observed to experience a recidivism event).

Under the incident/dynamic definition, at any one point in time, cases are defined as

those who recidivated *during that specified time period only* (incident) rather than as all individuals who have recidivated up to that point (cumulative). Similarly, controls are defined as individuals who are considered censored *during that specified time period* (dynamic) rather than limiting controls only to those who were fully censored across the entire follow-up period (static).

In practice, this means that any single participant may be counted among the controls in an early time period, and subsequently considered to be a case in a later time period. Thus, the calculation is described as using *incident sensitivity* and *dynamic specificity*, due to the way it defines cases as *incident cases* and controls as *dynamic controls*. Note that this feature mimics the way participants are treated within Cox regression, where individuals contribute information to the survival model's maximum likelihood equation as a control participant up until the point that they contribute information as a case participant, at which point they exit the analysis (or, may exit due to censoring; Heagerty & Zheng, 2005; Singer & Willet, 2003).

Summary. These two effect size options have appealing properties, practical interpretations, and important utility for comparing results across Cox models (whether employing models with time-varying or time-constant predictors). For these reasons, these effect size indices were used alongside the analyses within this dissertation. Within tables displayed in Chapters 4 and 5, Xu and O'Quigley's (1999) index is referenced as " R^2 (XO)", and Heagerty and Zheng's (2005) concordance index is referenced as the "*c*-index". Time-dependent AUC analyses are also graphed

within various figures shown in Chapter 5. Also, readers wishing to calculate these effect size indices with their own data will find guidance provided in the Appendix.

Some discuss effect size indices as if they express the “real world” impact of the results. This is not the case (Blanton & Jaccard, 2006a; Yeaton & Sechrest, 1981). Also, transforming a model’s effect into an effect size does not make the scale of interest any less arbitrary (Blanton & Jaccard, 2006a). These effect size indices have inherent limitations (i.e., AUC estimates will vary depending on the range of risk scores observed in the sample; Hanson, 2008), and are most valuable when considered alongside model coefficients, rather than in isolation. But, both indices remain appealing as a result of having easy interpretations and practical meaning, by nature of being expressed as proportions. Specifically, R^2 (XO) can be readily understood as the proportion of total variance shared between the predictors and outcome, and the *c*-index can be readily understood as the proportion of correctly ranked participants.

Chapter 3: Methodology & Dataset Preparation

This chapter introduces the dataset analyzed in this dissertation, including the jurisdiction, participants, measures, and data collection procedure. For more information about the dataset, and the development of the dynamic risk measure (Dynamic Risk Assessment for Offender Re-entry, or DRAOR; Serin, 2007), see Hanby (2013; also discussed later in this chapter). After introducing the dataset, this chapter also reviews prior research that establishes the psychometric properties of DRAOR, and highlights limitations of the measure and study design.

Method

Participants

Participants were drawn from a dataset that included all offenders released on community supervision after incarceration (i.e., parole) in New Zealand during a two-year period. In New Zealand, all offenders sentenced to two or more years in prison must report to a supervision officer for at least six months following release (Yesberg, Scanlan, Hanby, Serin, & Polaschek, 2015). Thus, offenders who are incarcerated for their full sentences are supervised in the community for six months only, whereas offenders released prior to their sentence expiry are supervised for greater than six months.

At the fiscal year starting April 1, 2010, a new dynamic assessment measure (i.e., DRAOR) was implemented into standard practice within the New Zealand Department of Corrections, requiring community supervision officers to rate all individuals on their caseloads. Officers rated offenders at every “quality contact”, defined as any meeting of sufficient length to assess the items.

Between April 1, 2010 and March 31, 2012, a total 97,188 assessments were recorded for 3498 offenders. Three assessments had an incorrect date and were deleted. An additional 196 offenders (5.3%) were released on supervision during this time period, but assessments were never recorded for these offenders. It is unknown why no assessments occurred; it is possible these offenders recidivated and were returned to incarceration before their first meetings with their supervision officers, but this is speculation. Neither recidivism data nor static risk scores were available for offenders who were not assessed.

Among the 3498 offenders with dynamic assessments, 77 offenders were excluded because assessments could not be matched to their supervision start dates. For 11 of these 77 offenders, assessments began at implementation (in April 2010), but since their supervision began sometime prior to this, assessments could not be matched to date of release. These 11 offenders represented 202 assessments. Similarly, another 190 assessments recorded early in the implementation period did not have a recorded release date; these assessments were removed, but assessments associated with a later release date for these same offenders were retained.

For the remaining 66 removed offenders, assessments could not be matched to date of release for other reasons (see further justification of these decision criteria below). Seven offenders were removed because all assessments were recorded much earlier than their release date. It is likely that these prior assessments were associated with a prior release date that was not recorded in the dataset. The remaining 59 excluded offenders did not have a recorded assessment within four weeks of release. As discussed below, baseline assessments were defined as assessments that took place within 28 days of release (four discrete seven-day periods). These 66 offenders represented 1675 assessments that were removed from analyses.

The 3421 offenders included in analyses (M age at release = 35.1, SD = 11.7, range = 17-86 years) were primarily male (92.8%). Date of birth was missing for two participants, and gender was missing for three participants. Ethnicities were identified by country of origin, specified as Aboriginal (52.4% total; Cook Islands,

2.5%; Fiji, 0.2%; New Zealand Maori, 49.5%; New Zealander, 0.1%) or countries located in Africa (0.4%), Asia (1.6%), Europe (2.7%), White individuals from Ireland, the British isles, and British colonies (Australia, New Zealand, United States; 34.9%), the Middle East (0.5%), South America ($n = 1$), or South Pacific islands (7.3%).

Ethnicity information was unspecified for five individuals.

Participants were single (never married, 65.5%; divorced, 3.0%; widowed, 0.6%; legally separated, 8.2%) or married (legally, 6.5%; common-law, 10.1%). Marital status was unknown or unspecified for 210 individuals (6.1%). Individuals were incarcerated due to convictions for technical violations (including alcohol and traffic offences, 2.8%), substance offences (14.4%), non-violent offences (fraud, property violations, etc.; 27.0%), non-sexual violent offences (39.3%), or sexual violent offences (16.4%).

Measures

Recidivism.

Recidivism follow-up continued until July 18, 2012, about four months following the end of data capture for assessments. For criminal recidivism, the dataset recorded the date the offence occurred (rather than date of detection or return to custody; personal communication, N. Wilson, July 3, 2013). This provides an accurate metric for the timing of the outcome behavior. However, for breaches of supervision conditions and technical violations, it is unclear if the date of recidivism was (a) the date the behavior of the breach took place, (b) the date that the supervision officer became aware that problematic behavior had occurred, or (c) the

date that the supervision officer decided to record a breach. This likely varied among technical violation events.

Dates of new offences did not always coincide with the end of ongoing dynamic assessments by supervision officers. There were at least three examples of this. First, some recidivism events were non-serious technical violations that did not result in re-incarceration. In these cases, assessments continued in the community uninterrupted, even after a breach had taken place. Second, some recidivism events were not detected quickly, and assessments continued until the offender was eventually apprehended and removed from the community. Third, for some offenders, supervision was completed a substantial amount of time before the recidivism event occurred. In the most common situation, an offender completed a six-month supervision order (at which time dynamic re-assessments ended), but reoffended some weeks later.

Thus, there were complex ways that sequences of multiple recidivism events could be matched to sequences of assessments. This project used a simplified examination of recidivism by focusing on the earliest recidivism event for each offender. The definition of recidivism within this project includes both technical violations and new criminal actions. Although technical violations represent less serious recidivism events, these violations have been shown to significantly correlate with future convictions for new offences (Gray, Fields, & Maxwell, 2001; Polaschek, Yesberg, Bell, Dickson, & Casey, 2014 as cited in Yesberg et al., 2015). Recidivism sequences were examined, and the date of an offender's first chronological recidivism event, whether technical violation or criminal reoffence,

was coded as their date of *any recidivism*. Additional recidivism events that occurred following the first recidivism event were not included in the analyses presented here.

Within the two year and four month follow-up period, there were 1538 recidivists among the 3421 offenders, accruing 1754 recidivism events. Among these, 893 offenders only recorded a technical violation, 429 offenders recorded only a criminal offence, and 216 offenders recorded both. For this project, the 1109 technical violations were defined broadly as crimes without an identifiable victim, which included true technical violations (administrative, parole act offences, etc.), alcohol offences (breaking restrictions on alcohol intake, drinking in public, etc.), administrative driving offences, and drug possession offences. The 645 criminal offences were defined as crimes against others, including 366 (56.7%) non-violent crimes (fraud, theft, property destruction, etc.), 268 (41.6%) non-sexual violent offences (threats, robbery, assault, family violence, etc.), and 11 sexual violent offences (1.7%).

For this project, the length of the potential follow-up period was truncated to one year. In addition, rules were established regarding how distal a dynamic assessment could be to a recidivism event before the assessment was no longer considered proximal (i.e., within a six week range, see discussion below). As a result, recidivism events were not counted for analytic purposes if these events occurred (a) after one year had elapsed, or (b) seven or greater weeks following an offender's last dynamic assessment. Although these decision rules represent a loss of known outcome information, the analyses presented within this project aim to determine

whether a dynamic risk assessment can predict relatively concurrent recidivism events. This purpose requires removing non-concurrent recidivism events.

However, despite these decision rules narrowing the outcome information, 76.6% of all recidivism events ($n = 1344$) were included in analyses (62.4% of these were technical violations, 37.6% criminal violations; see below for more detail).

Static risk.

The *Risk of Reconviction***Risk of Re-imprisonment* measure (RoC*RoI; Bakker, Riley, & O'Malley, 1999) is the New Zealand Department of Corrections' population-developed tool. RoC*RoI contains demographic and criminal history information (16 items) and accurately predicts new convictions within five years following release that result in a custodial sentence (AUC = .76; Bakker et al., 1999; Nadesu, 2007). Demographic items include gender and current age. Criminal history items include indicators of criminal career length (age at first offence, total time incarcerated, etc.), crime frequency (recent and total), and crime seriousness (recent and total).

Scores range from 0 to 1, with higher scores indicating higher risk to reoffend. Generally, scores between 0 and 0.49 are accepted to indicate low risk (less than 50% probability of return to incarceration), scores between 0.50 and 0.64 indicate moderate risk, and scores between 0.65 and 1.0 range from high to very high risk. Any specified score can be interpreted as the likelihood over the next five years that the individual will be reconvicted *and* returned to incarceration (i.e., a RoC*RoI score of 0.55 represents a 55% likelihood of return to incarceration). The dataset included total scores only, precluding calculation of internal consistency.

Dynamic risk and protective factors.

Dynamic Risk Assessment for Offender Re-entry (DRAOR; Serin, 2007; Serin, Mailloux, & Wilson, 2012) is designed to provide supervision officers with a structure for comprehensive assessment of in-community features and circumstances among general adult offenders, and to track changes in these variables as offenders re-enter and remain in the community. DRAOR owes its item content to research evidence from the correctional, desistance, re-entry, and rehabilitation literatures, with item selection most heavily drawn from Andrews and Bonta (2010), Hanson and Harris (2000), Hanson and colleagues (2007), Maruna (2001), and Sampson and Laub (2005). Hanby (2013) discussed item selection at length.

Specifically, items are organized conceptually into three domains reflecting *stable* dynamic risk (six items), *acute* dynamic risk (seven items), and *protective* factors (six items). The six *stable* items include traits and behavior patterns conceptualized as relatively enduring, but potentially dynamic. Two attitude items include *sense of entitlement* (inflated self worth) and *attitudes toward authority* (antagonistic to conventional authorities). Two trait items include *impulse control* (highly impulsive) and *problem solving* (poor consideration of consequences). Two behavior pattern items include *peer associations* (spends time only with antisocial others) and *attachment with others* (behaves callously toward others).

The seven *acute* items include situations, moods, and behaviors that may change rapidly. Four situation items include *living situation* (lack of accommodations), *employment* (unemployed), *interpersonal relationship* (high

conflict romantic relationship, or no relationship), and *access to victims* (opportunity to offend). Two mood items include *negative mood* (exhibiting depressed or anxious mood), and *anger/hostility*. The final item assesses current problematic *substance use*.

The six *protective* items assess prosocial perceptions and prosocial connectedness. Three items assess internal perceptions, including *prosocial identity* (views self as oriented toward non-criminal goals), *cost/benefits* (views prosocial options more favorably than criminal actions), and *responsive to advice* (views prosocial individuals as valuable resources). Three social connectedness items assess *social support* (availability of prosocial others), *high expectations* (encouragement and investment from prosocial others), and *social control* (internal investment in prosocial goals approved by others).

All items are scored on a three-point scale. Each risk item is scored as *no problem* (0), *slight/possible problem* (1), or *definite problem* (2) and each protective item is scored as *not an asset* (0), *slight/possible asset* (1), or *definite asset* (2).

Supervision officers rate each item based on interviews as well as any additional information available from collateral sources, official reports, or other caseworkers (e.g., program providers).

Psychometric properties are discussed in a subsection below.

Demographic information.

Date of birth, gender, country of origin, marital status, and index offence were gathered from New Zealand Department of Corrections' Integrated Offender Management System (IOMS).

Psychometric Research on DRAOR Data

Several past projects have validated DRAOR as a promising tool for risk assessment, and dynamic re-assessment. These validation projects arose from the implementation of DRAOR into community supervision practice throughout Iowa (see Chadwick, 2014; Smeth, 2013) and throughout New Zealand (see Hanby, 2013; Yesberg & Polaschek, 2014; Yesberg et al., 2015). Table 3.1 displays a list of existing research reports on DRAOR. This table provides the defining features of each study, the inter-correlations among DRAOR subscales, and internal consistency estimates.

This section reviews the research listed in Table 3.1 by focusing on the psychometric properties that are necessary to justify building the statistical models presented in this dissertation. In particular, there are at least three psychometric properties that help justify utilizing DRAOR measurements in a longitudinal, re-assessment Cox regression model: (a) invariant measurement across multiple occasions, (b) association with future recidivism outcomes, and (c) individual-level change in scores across time.

Factor structure and measurement invariance.

As described above, DRAOR is divided into three subscales, incorporating two dimensions: (a) the direction of the association between items and recidivism (*protective* factors vs. risk factors), and (b) the theoretical likelihood the item will change in the long- vs. short-term (*stable* vs. *acute* risk factors). Three projects used factor analytic strategies and have largely confirmed the intended three-subscale structure of DRAOR, albeit with some mixed results. Specifically, the distinction

Table 3.1
Psychometric Properties (Cronbach's Alpha and Pearson Correlations) of DRAOR Data Reported in Past Publications

Source	Sample	Location	<i>n</i>	Cronbach's α			Pearson's <i>r</i>		
				Stable	Acute	Protect	Stable-Acute	Stable-Protect	Acute-Protect
Serin & Prell (2012)	probationers & parolees	Iowa	563	-	-	-	.74	-.70	-.66
Hanby (2013)	parolees	NZ	3372	.81	.62	.84	.61	-.71	-.59
Smeth (2013)	male sex offenders	Iowa	212	-	-	-	.72	-.57	-.47
Chadwick (2014)	probationers & parolees	Iowa	391	.81	.62	.86	.73	-.53	-.49
Yesberg & Polaschek (2014)	high risk male parolees	NZ	299	-	-	-	-	-.50	-

Note. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale. NZ = New Zealand. "-" = results not reported.

between the proposed factors separating *stable* and *acute* risk items has not consistently emerged.

For example, an exploratory factor analysis (EFA) loaded all risk items together, separate from protective factors, creating a two-factor scale (Chadwick, 2014). Hanby's (2013) analyses suggested that the original DRAOR structure fit her data well, but that model fit could be improved if some risk items were allowed to cross-load, or were re-located from the *stable* subscale to the *acute* subscale, or vice versa. By contrast, using a confirmatory factor analysis (CFA) with data from an exclusively high-risk sample, Yesberg and Polaschek (2014) found the original subscale structure did not fit their data well. Instead, similar to Hanby (2013), a principal components analysis suggested some crossover between the *stable* and *acute* subscales. In addition, the *acute* risk subscale items loaded conceptually into two new subscales defined by either external circumstances or internal features (Yesberg & Polaschek, 2014).

Factor analytic procedures and statistics establishing internal consistency (i.e., Cronbach's alpha) were designed to test whether a collection of items measure a unitary underlying psychological construct. Some have criticized the convention of applying these statistical procedures with measures intended for risk assessment (e.g., Baird, 2009), for at least two reasons. First, the purpose of risk measures is to predict reoffending, and some have argued that the focus should be on creating measures that maximize prediction, rather than creating measures that assess a conceptually distinct, unitary construct. Indeed, these two goals are usually in conflict with each other, since prediction is likely maximized by aggregation of

variables into a well-rounded, multi-faceted assessment (Kroner et al., 2007).

Second, leading theories of criminal behavior (such as the theories discussed in Chapter 1) do not conceptualize risk potential as a unitary dimension, but rather as a dynamic interaction between multiple internal and external variables. For these reasons, it is problematic to rely on factor analysis exclusively for establishing whether a risk assessment instrument has utility. Note, however, that risk instruments that intend to describe a set of psychologically relevant constructs related to risk should exhibit a factor structure that conforms to the proposed constructs.

Yet, applying factor analysis to DRAOR data is arguably a limited approach, because each DRAOR item was chosen to be a single indicator of a personal or external variable, and subscales were chosen that combine items that have similar temporal relationships to recidivism. As such, there is no compelling conceptual reason why the items would necessarily load together onto the subscales as designed, given that DRAOR does not propose to contain three distinct psychological phenomena. Despite this, the factor analyses described above largely justify proceeding with research using the original intended DRAOR subscales, while keeping in mind that there may not be a clear conceptual distinction between *stable* and *acute* risk factors.

On the other hand, for the present project, it is arguably more important that DRAOR demonstrate consistent factor structure, i.e., factor invariance across measurement occasions. *Measurement invariance* refers to empirical evidence that scale items measure the same underlying construct(s) both at the initial assessment

occasion, and at later re-assessments. There are many reasons scores on the same risk scale taken at different times may unintentionally measure different constructs (see Babchishin, 2013). For example, Brown and colleagues (2009) suggested that it may be misleading to compare scores on the same scale if the measure is first assessed in prison (e.g., measuring anticipated leisure problems) and later assessed in the community (e.g., measuring actual observed leisure problems). Despite rating the same items, scores may not assess the same psychological construct, if the time between assessments involves a change in circumstances, raters' understanding of item content, social expectations of behavior (e.g., a measure of biting behaviors would assess a relatively normative interaction among toddlers, but a rare and bizarre behavior among adults), etc. For this reason, measurement invariance is a critical assumption when examining change in scores within longitudinal analyses. If a measure's latent construct varies across time (i.e., the latent constructs at each time point are not equivalent), a researcher cannot take for granted that changes in risk scores represent changes in a variable of interest.

To establish measurement invariance of DRAOR assessments across time, Hanby (2013) conducted two CFAs on re-assessments taken at different time points subsequent to initial baseline scores. Results indicated the original subscale structure of DRAOR remained a good fit of the data at these two time points. Similarly, Yesberg and Polaschek (2014) reported that their CFA results did not substantially differ when analyzing either the first or third DRAOR assessment in the sequence of re-assessments. These results support the assumption that DRAOR scores have the same meaning at each re-assessment, and this allows for changes in

scores to be considered indicators of changes in the underlying construct. Note, however, that there are more rigorous ways to test the assumption of measurement invariance, specifically using longitudinal exploratory factor analysis (see Asparouhov & Muthén, 2009) to explicitly test whether scale items' factor loadings and item thresholds remain invariant across all periods (i.e., by comparing model fit when loadings are fixed versus free; see Babchishin, 2013; Bergeron & Miller, 2013, for examples of how this statistical technique has been used in correctional psychology).

Predicting future recidivism outcomes.

When assessed at a single point in time (i.e., at the baseline of a supervision order), there is evidence that DRAOR subscale scores predict future violations and offences while under community supervision. A summary of outcome results is provided in Table 3.2. Each of these studies reported the relationship between DRAOR scores and more than one definition of recidivism. In summary, Table 3.2 shows that DRAOR scores (a) significantly predict general recidivism (technical violations and new criminal offences) to a moderate to strong degree, (b) significantly, but weakly predict new offences only (excluding technical violations), and (c) are not associated with sexual recidivism outcomes.

In addition, almost all tests have been successful when examining whether DRAOR adds incremental prediction of general recidivism beyond static risk scores, using Cox regression survival analyses or logistic regression. Both Hanby (2013) and Yesberg and Polaschek (2014) used a total score that combines subscales to show DRAOR predicted recidivism after accounting for RoC*RoI scores, and Smeth

Table 3.2

Associations Between DRAOR Scores and Recidivism Outcomes (Point Biserial Correlations and Area Under the Curve) Reported in Prior Publications

Source	Outcome	Stable		Acute		Protect	
		<i>r</i>	AUC [95% CI]	<i>r</i>	AUC [95% CI]	<i>r</i>	AUC [95% CI]
Serin & Prell (2012)	Technical violations	.23	.60 [-, -]	.25	.65 [-, -]	-.28	.67 [-, -]
	New offences only	.12	-	.11	-	-.15	-
Hanby (2013)	Technical violations or new offence	.29	.66 [.65, .68]	.38	.72 [.70, .74]	-.29	.67 [.65, .68]
	New offences only	.21	.62 [.60, .64]	.29	.67 [.65, .69]	-.21	.62 [.60, .64]
	Maori offenders:						
Smeth (2013)	Technical violations or new offence	.26	.65 [.62, .67]	.35	.70 [.68, .72]	-.26	.65 [.62, .67]
	Technical violations	.33	.69 [.62, .78]	.27	.65 [.57, .74]	-.18	.61 [.52, .69]
Chadwick (2014)	Sexual recidivism	.01	.51 [.26, .75]	.05	.46 [.30, .75]	.01	.46 [.21, .71]
	Technical violations or new offence	-	.62 [.56, .67]	-	.59 [.54, .65]	-	.58 [.52, .64]
	New offences only	-	.52 [.42, .62]	-	.53 [.43, .63]	-	.56 [.47, .66]
	Technical violations	-	.53 [-, -]	-	.53 [-, -]	-	.58 [-, -]
Yesberg & Polaschek (2014)	New conviction	-	.61 [-, -]	-	.57 [-, -]	-	.60 [-, -]

Note. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale. *r* = point-biserial correlation. AUC = area under the curve. CI = confidence interval. “-” = results not reported.

(2013) showed that each DRAOR subscale predicted parole violations above Static-99R scores (Hanson & Thornton, 2000; Helmus, Thornton, Hanson, & Babchishin, 2012). By contrast, among a sample of female offenders, only the *acute* subscale predicted recidivism; however, this subscale also showed incremental prediction above static scores (Yesberg et al., 2015). Yet, in this same study, no DRAOR scores predicted recidivism among a sample of male offenders selected to match the female offenders. However, oddly, in another study that also matched female and male offenders, all DRAOR scores predicted recidivism in both groups, except the *acute* scale among female offenders (using a sample from New Zealand; Scanlan, Fortune, Polaschek, & Yesberg, 2015). Thus, the predictive validity of DRAOR scores for general recidivism has been largely consistently significant, with some unstable findings within some offender subgroups. There has also been consistently non-significant prediction when examining sexually violent outcomes only, and in addition, DRAOR did not predict general reoffending among a subsample of sex offenders with child victims (Smeth, Serin, & Hanby, 2015).

Detecting change across time.

Items were chosen with the intention that DRAOR would be sensitive to detecting fluctuations across time, and the available research supports that DRAOR can be used to observe change. Specifically, Hanby (2013) used multilevel modeling to demonstrate that, on average, individual-level changes in *stable*, *acute*, and *protect* subscale scores occurred in the expected direction. In other words, as offenders remained crime-free in the community, risk scores showed a significant average decrease, and protective factor scores showed a significant average increase.

Similarly, among both female and male offenders, scores on each DRAOR subscale were observed to significantly change between two assessments (Scanlan et al., 2015). Change was also observed across time using total combined subscale scores in another sample (male offenders in Iowa; Chadwick, Serin, & Lloyd, 2015), but only a small percentage (9.6%) of change scores exceeded the threshold set by the *reliable change index* (see Jacobson & Truax, 1991).

There is additional evidence that changes in DRAOR scores are associated with the likelihood of recidivism. Changes in *stable* and *protect* subscale scores that demonstrated a pattern of increasing risk significantly predicted recidivism in univariate Cox regression models (Scanlan et al., 2015). Similarly, change in DRAOR assessments incrementally added to prediction models beyond first assessment scores (Chadwick et al., 2015). Both of these studies used a two-assessment design.

Summary.

Overall, the results from past research supports conceptualizing DRAOR as a dynamic risk assessment instrument. First, DRAOR is reliably associated with future recidivism (especially when recidivism is defined most broadly to include technical violations), establishing that DRAOR has utility as a risk instrument. Second, DRAOR re-assessments were observed to change during the re-entry time period, establishing that DRAOR can be considered to be a dynamic instrument. Finally, the factor structure of DRAOR appears to remain consistent across time, supporting the use of re-assessment scores in statistical models, without concern that observed change may be resulting from DRAOR assessing different latent constructs at different points in time.

Structuring the Data for Analysis

The narrative information within this section is also displayed visually as a flowchart in Figure 3.1.

Defining the unit of analysis.

Prior to preparing the data for analysis, the dataset included 3421 offenders with dynamic assessments that could be matched to a date of release. During the two-year period, some of the 3421 offenders reoffended, were returned to incarceration, and then were re-released on community supervision. Their multiple release dates resulted in multiple assessment sequences. Specifically, two sequences were recorded for 207 offenders, and three sequences were recorded for 13 offenders. These subsequent re-entry trajectories were retained as independent re-entry sequences alongside the primary sequences (compare with Howard & Dixon, 2013, who also used this strategy with similar data). This resulted in 3654 *sequences* for analysis.

This strategy was chosen for two reasons. First, each re-entry period is a theoretically independent unit of analysis, when baseline is defined by an individual's release from incarceration. Each re-entry experience can be conceptually considered a unique occasion, even if the same offender undergoes multiple releases. For individuals experiencing multiple re-entry occasions, the most critical difference between release dates is the number of prior incarcerations associated with the release. This is formally accounted for within the static risk scores used in this project's analyses, such that number of prior offences and incarcerations were items used to calculate static risk. Thus, static scores in this

Figure 3.1. Participant flowchart describing inclusion criteria for analysis.

dataset are dynamic in that individuals may have risk scores upgraded after a new recidivism event. (Static risk may also be downgraded in situations when the most recent offence is less serious than the one prior; see description of RoC*RoI in Measures. In this sample, change in static risk score was observed within $n = 533$ individuals, or 15.6% of the sample, $n = 12$ missing. Static scores were also updated in “real time” in the system during periods of supervision, such as when changes were observed in current age and recent criminal activity. Thus, examining assessment sequences, change in static risk score was observed within $n = 474$ sequences, representing 13.0% of sequences, $n = 14$ missing.)

Second, although it is typically considered problematic to assume data are independent when they are non-independent (two pieces of data were taken from the same individual), this is less of a concern when utilizing within-individual methodologies. In particular, repeated assessments are non-independent in terms of the offender’s life circumstances, stable traits, living situation, etc., and this is also true for two assessment sequences drawn from the same individual. However, when analyses control for within-individual assessment patterns, it is unlikely problematic to treat two within-individual sequences from the same individual as independent. This is especially true when considering the first point above, given that static and dynamic assessments may have different starting values for each sequence, even if two sequences were drawn from the same individual on separate occasions.

However, to ensure this decision did not bias results, all analyses presented in this dissertation were compared to analyses that excluded second and third

sequences, restricting the sample only to first sequences (i.e., 3421 offenders). Including these multiple sequences did not substantively alter the results or conclusions.

Defining units of discrete time.

As described in Chapter 2, the analytic approach adopted for this project requires establishing meaningful discrete units of time, and then assigning assessments and recidivism events to their appropriate discrete time periods. Within the dataset, the precise date was available for each assessment and recidivism event. Modeling events using days as the unit of time would closely approximate continuous time, but, when using the analytic approach adopted in this project, analyzing continuous time quickly becomes computationally unwieldy. The other important consideration is whether results can be interpreted in a meaningful way for practical decisions. Because supervision officers are unlikely to meet their clients on a daily schedule, the practical implications are less straightforward when using continuous time within analyses. As a result, it was necessary to group the data into larger units of time, for ease of modeling, and for practical conclusions.

Weeks were chosen as a meaningful unit of time, since offenders are often required to have contact with their supervision officers on a weekly, or fortnightly basis. As a result, supervision officers would benefit from research that informs whether assessment results may indicate that a client will be at increased risk during the upcoming week, prior to the next scheduled meeting. This choice required grouping the assessment data into weeks, specified as seven-day periods.

Defining baseline.

For all offenders, Day 1 was defined as the date of release, and Day 8 was the first day of the second discrete period, or the second week. For many sequences, assessments occurred on the date of release ($n = 1080$; 28.9%); for these sequences, the assessment designated as the baseline assessment was precisely meaningful. For some sequences, pre-release assessments occurred in anticipation of the return to the community. However, when pre-release assessments were recorded more than a month prior to release, it was unclear if these assessments were taken in anticipation of release dates that were ultimately canceled, or if these were community based assessments from an earlier period of release. Due to this ambiguity, only pre-release assessments that occurred within 30 days of the release date were used as baseline assessments ($n = 143$). Within this 30-day period, pre-release assessments typically occurred shortly before release ($M = 8.0$, $SD = 8.1$, Median = 5 days). Note that if a sequence included both a pre-release assessment and a Day 1 assessment, the latter assessment was retained. After choosing an assessment to represent baseline, other pre-release assessments were removed from analysis.

Thus, 1223 sequences (33.5%) included baseline assessments that were either recorded on the day of release, or during the month before release. The remaining sequences did not have assessments taken at Day 1, or pre-release. For these sequences, baseline was defined as assessments recorded within at least four weeks of release. As stated above, sequences were removed from analysis if the first assessment occurred later than Day 28 (end of Week 4; $n = 59$ offenders, $n = 78$

sequences), so that all baseline assessments were taken within a month of release date. For these removed sequences, the gap between date of release and first assessment ranged from 5-69 weeks ($M = 17.6$, $SD = 15.9$, Median = 11 weeks). For the remaining sequences, a subsequent assessment was carried backward to fill in the missing baseline. The most distal imputed baseline assessment was taken from Week 4. In those cases, Week 4 assessments were also used to impute Weeks 1-3. Of the 2431 sequences where assessments were carried backwards to baseline, most first assessments occurred in Week 1 ($n = 2113$, 86.9%), with the remainder in Week 2 ($n = 233$, 9.6%), Week 3 ($n = 59$, 2.4%), and Week 4 ($n = 26$, 1.1%).

Thus, baseline assessments were defined as the assessment closest in proximity to day of release, within a month of release date (i.e., if not on Day 1, either within 30 days prior to release, or within 28 days after release). This allowed any change in assessments across time to be meaningfully interpreted as growth from a standard timeframe, defined as the time of release from incarceration (plus-or-minus one month).

Choosing which assessment would define the discrete period.

Each subsequent discrete period was defined as a seven-day period. However, some sequences had more than one assessment recorded during a seven-day period. Although it was most typical for only one assessment to occur in a week, in every week of the one-year follow-up, at least a few sequences included multiple assessments per week. In New Zealand, policies may require supervision officers to rate offender clients on DRAOR as often as twice a week, or as seldom as every other week; decision rules depend on risk level and how long the offender has been

supervised in the community (Yesberg et al., 2015). In this dataset, at most, seven assessments were recorded for a single offender in a single seven-day period (in the third week of the follow-up).

Two alternate approaches were used to choose the assessment that best represented the discrete period; in analyses, these strategies were explicitly compared to determine which resulted in the best model fit. First, the last assessment of the week was chosen to represent the week, on the basis that these ratings contained the most information (i.e., the week's circumstances were most known to the assessor at the last assessment compared to an earlier assessment). Second, a duplicate dataset was built that instead took the unweighted mean of all assessments occurring within that week, combining all information for the discrete period. This strategy allowed any within-week fluctuations to be reflected in the score that represented that week. Testing and comparing these different selection criteria can determine if model fit is appreciatively altered due to the different approaches (Singer & Willett, 2003).

Comparing these approaches is important because each reflects a different way supervision officers might conceptualize a difficult client's week. For example, if a client begins the week exhibiting high-risk behaviors, but shows improvement by the end of the week, it is unclear if the supervision officer should (a) weight the end-of-week improvements more heavily because they are most recent, or (b) take the entire pattern of behavior into account. If differences in prediction occur as a result of these alternate conceptualizations, this would inform decisions about how to manage offenders exhibiting very short-term changes.

Truncating the follow-up period.

Assessments were compiled into weeks systematically from baseline to the end of the follow-up period (Week 52, a one-year follow-up). Although some offenders remained on supervision and continued to be assessed beyond one year, a relatively small amount of sequences (15.8%) remained in the dataset at this point.

Lagging recidivism information to ensure prospective prediction.

Timing of recidivism was similarly placed into seven-day discrete periods, such that a recidivism event was coded as occurring within a week-long period, rather than on a specific day. To ensure assessments were taken prior to recidivism events, recidivism information was lagged by a week. In other words, assessments from the week prior to the week of outcome were always used to predict the outcome event. When assessments occurred following a recidivism event, sequences were truncated to end at the week before recidivism. This ensured that all prediction was prospective, and knowledge of recidivism was not used retrospectively to predict outcome.

Truncating sequences due to long gaps between assessments.

It was uncommon for assessment sequences to be fully complete (i.e., assessments were rarely available every week until recidivism or loss to follow-up). For 3364 sequences (92.1%), at least one week was missing an assessment ($M = 17.3$, $SD = 14.7$, Median = 13 weeks). Across all sequences, missing weeks represented an average of 35.8% ($SD = 18.5\%$) of possible weeks. However, it was more likely for a week with a missing assessment to occur alone rather than in conjunction with another week missing an assessment. On average, 65.7% ($SD =$

31.6%) of missing weeks were single instances. Higher risk offenders (measured by RoC*RoI scores) were slightly less likely to have frequently missing assessments, suggesting these offenders were more likely to meet with their supervision officers weekly, rather than fortnightly, $r = -.17, p < .001$.

Despite this, some long gaps in assessments were observed. Since the purpose of analyses was to examine proximity between assessments and recidivism outcomes, a decision criterion was required to define how distal an assessment could be and still be considered proximal. Also, to meet the requirement for statistical models, data needed to be imputed for missing weeks, which also logically requires some limit to how far existing data would be stretched to fill in missing weeks.

Thus, assessment sequences were truncated if a gap of more than six weeks occurred. This ensured that assessment proximity would be defined as no more than 1.5 months from the time of the assessment. When these long gaps occurred, assessments following a seven-week gap were removed, and the offender was considered lost to follow-up at the time prior to the seven-week gap. For some sequences, this meant truncating a follow-up time prior to a known recidivism event. Although this represented a loss of information about a known outcome, this strategy ensured that prediction would only be based upon assessments that were closely proximal to the recidivism event. Overall, most recidivism events were proximal to assessments (85%), and could be retained.

Imputing assessments during weeks when no assessments occurred.

Three different strategies were used to impute data during weeks where no assessments were recorded. Multiple strategies were used so that the different approaches could be compared within analyses. Note that these three strategies were each applied separately to the two alternate ways of combining multiple assessments within the same week (see above, i.e., choosing the last assessment of the week, or taking the average of the week's assessments). This resulted in six total imputation strategies, which are displayed visually in Table 3.3.

The first strategy involved imputing each missing data point with the most recent prior assessment. The second and third strategies involved imputing each missing data point with the average of the most recent two or four prior weeks' assessments, respectively.

Singer and Willett (2003) encourage structuring imputation strategies in these reasoned, conceptual ways, and then explicitly testing whether the strategies make a difference for prediction (based on model fit indices). For this project, such comparison accomplishes two goals. First, explicitly comparing strategies provides an empirical way to justify the data imputation approach, strengthening conclusions taken from the best strategy. Second, the chosen imputation strategies approximate the way supervision officers in practice would reflect on each client. If a client is currently experiencing an upswing (or downswing) in risky behavior, the officer may wonder (a) whether this most recent assessment should be weighted most heavily for current management decisions, (b) if a combination of prior weeks

Table 3.3

Logical Strategies for Choosing Among Multiple Assessments in One Discrete Time Period and Imputing Scores in Missing Time Periods (as Recommended by Singer & Willett, 2003)

Strategies for Combining Multiple Assessments in the Same Week	Strategies for Imputing Missing Weeks		
	Imputing the Most Recent Assessment	Imputing the Average of the Most Recent 2 Assessments	Imputing the Average of the Most Recent 4 Assessments
Choosing the Last Assessment of the Week	Final Assessment of the Prior Week	Average of the Final Assessments from 2 Prior Weeks	Average of the Final Assessments from the 4 Prior Weeks
Choosing the Mean of all Assessments in the Week	Mean of all Assessments from the Prior Week	Grand Mean of all Assessments from the Prior 2 Weeks	Grand Mean of all Assessments from the Prior 4 weeks

should be considered, i.e., the prior fortnight or month, or (c) if a wait-and-see approach is the only legitimate course of action.

Summary.

After the above decision criteria were applied to the dataset (compilation, truncation, and imputation), a total of 92,104 person-period weeks were available for analysis, within 3654 sequences. These weeks represent 68,774 “true” assessments compiled into 60,158 “true” person-period weeks (65.3% of total weeks). Thus, 31,946 (34.7%) person-period weeks were imputed.

Although all 3654 sequences had consistent DRAOR assessments (as defined above), among these, six sequences were missing static risk (RoC*RoI) scores in the dataset. In the analyses presented in the next chapters, descriptive information on DRAOR was calculated for the full 3654 sequences, but to maintain consistency across all prediction models, analyses were limited to the 3648 sequences with full information on all predictors. Removing these six sequences represented the loss of 159 person-period weeks (95 “true” person-period weeks), and two recidivism events.

Limitations

There are limitations to the conclusions that can be drawn from the dataset used in this dissertation, which deserve to be acknowledged and highlighted prior to presenting the results in Chapters 4 and 5. These limitations broadly reflect the challenges of implementing a new practice with front-line staff in correctional settings. In particular, there is a general uncertainty whether supervision officers rated DRAOR items with fidelity. Although confirming implementation fidelity

represents an important research question, the information provided in the dataset used for this project is largely unable to examine the tenability of this assumption, for the reasons described below.

Data nesting within supervision officers.

From an analytic, model-building point-of-view, it is problematic that DRAOR assessments in the dataset are (in actuality) nested within supervision officers, but this nesting information is omitted from the data provided by New Zealand Department of Corrections. In other words, DRAOR assessments rated on the 3421 offenders within the dataset were completed by a finite number of supervision officers, each with multiple offenders on their caseloads. It is unknown how these assessments group together within supervision officers, resulting in nested data that cannot be analyzed in a way that models these data's natural hierarchy. Nesting information is needed to separate the differential effects across supervision officers, if these effects exist.

Differences in fidelity across supervision officers.

It is highly reasonable to hypothesize differential effects across supervision officers. Officers are likely to differ in the degree to which they (a) rate DRAOR items with diligence and insight, and (b) form effective supervisory relationships with offenders on their caseloads. Each of these features are likely to impact predictive validity of DRAOR scores, but in opposite ways, such that (a) diligent officers will more accurately rate DRAOR items, increasing the potential association between DRAOR scores and recidivism, and (b) effective officers will better manage offenders' risk, reducing the association between high DRAOR scores and recidivism.

Researchers often take for granted that risk assessment tools were implemented with fidelity (Vincent, Paiva-Salisbury, Cook, Guy, & Perrault, 2012), but studies with front-line correctional staff instead suggests that the most optimistic possible conclusion is that *some* officers might *sometimes* rate an instrument correctly. There are two critical challenges when implementing new practice: (a) training officers to rate a risk tool accurately, which requires extending effort beyond a basic training program, by either providing an intensive program (Vincent, Guy, Fusco, & Gershenson, 2012; Vincent, Paiva-Salisbury et al., 2012), refresher training (Desmarais, Van Dorn, Telford, Petrila, & Coffey, 2012), or computer programs that prompt careful consideration of the assessment items (Bosker, Witteman, Hermanns, & Heij, 2014), and (b) inspiring officers to invest in the ideology and utility of the risk instrument for their practice (Haas & DeTardo-Bora, 2009; Savaya, Monnickendam, & Waysman, 2001; Viglione, Rudes, & Taxman, 2014). Results from a national survey of supervision officers suggested these two challenges are independent, such that even officers who reported completing risk tools carefully and accurately were still unlikely to implement the results of the tool into their supervision practice (Miller & Maloney, 2013).

Similarly, a body of research suggests it is highly typical in supervision practice that (a) officers do not use risk tools as intended, nor how they were specifically trained to use them (Viglione et al., 2014), and (b) officers are unlikely to use information from risk tools to make informed case management decisions (Bonta, Rugge, Scott, Bourgon, & Yessine, 2008; Miller & Maloney, 2013; Peterson-

Badali, Skilling, & Haqanee, 2014; Savaya et al., 2000; Viglione et al., 2014). Note that these results arise from officers who were trained to accurately rate risk tools.

Thus, the largest challenge appears to be setting up the conditions that encourage officers to “buy-in” to the value of using a mandated risk instrument in their practice. For example, juvenile probation officers largely reported discounting the importance of a risk assessment tool, instead viewing the assessment results as a way to support the decisions they had already made with their own professional judgment (Guy, Nelson, Fusco-Morin, & Vincent, 2014). Similarly, officers gave preference to their professional judgment when it disagreed with assessment results (Savaya et al., 2000); as expected from the broader literature on clinical prediction accuracy, research shows that accuracy degrades when front-line supervision officers overrode risk tool scores (Vaswani & Merone, 2014).

These two implementation features (training and “buy-in”) represent limitations of the present research because it is not possible to estimate the degree to which DRAOR training and investment differed across supervision officers who provided data on the offenders in the sample. As discussed above, due to lack of information about nesting in the dataset, it is also impossible to *a posteriori* estimate these differential effects by grouping information into hierarchical, multilevel analyses.

However, differences in training place a lower limit on the predictive ability of a risk tool rather than artificially inflating findings. For example, research with front-line staff rating the Level of Supervision Inventory-Revised (LSI-R) showed the relationship between LSI-R scores and recidivism was significant among formally

trained officers, $r = .21$, 95% CI [.16, .26], but null among untrained professionals, $r = .08$, 95% CI [-.02, .18] (Flores, Lowenkamp, Holsinger, & Latessa, 2006). Similar results were observed with Iowa officers who rated DRAOR. Scores gathered from formally trained officers attained significant prediction of general recidivism, AUC = .68 [.60, .76], .65 [.56, .73], and .62 [.53, .70] for the *stable*, *acute*, and *protect* subscales, respectively, whereas scores from untrained officers did not predict, AUC = .55 [.48, .63], .52 [.45, .60], and .52 [.45, .60], for the *stable*, *acute*, and *protect* subscales, respectively (Chadwick, 2014).

By contrast to the use of DRAOR in Iowa, all supervision officers who provided data for the dataset used in this dissertation were formally trained to properly rate DRAOR items. Furthermore, a pilot study with 59 New Zealand supervision officers that included in-depth interviews with seven participants suggested that officers attained a good understanding of item content (Tamatea & Wilson, 2009). As a result, training status can be assumed to be constant across all officers. On the other hand, the dataset offers no information about differences in level of “buy-in”, despite it being likely that individual differences in officers’ perceptions of the utility of DRAOR scores would have been observed.

Unknown inter-rater reliability between supervision officers.

Related to concerns regarding fidelity, it is an important property of a standardized risk tool that two raters are able to assess the same case with equivalent ratings. At present, there are no existing data to provide evidence of inter-rater reliability for DRAOR scores.

This is a common criticism of risk tools used in correctional practice. In their review of 53 studies examining risk assessment use in the United States among general offender populations, Desmarais and Singh (2013) identified only two studies that evaluated inter-rater reliability. Thus, tool developers have not traditionally focused on establishing inter-rater reliability; however, results from these two studies were highly promising because reliability was examined in the field (comparing two practitioners' ratings), and inter-rater reliability was excellent in both studies. By contrast, most existing inter-rater reliability research has not had high ecological validity (Vincent, Guy et al., 2012), either relying on laboratory settings (i.e., rating vignettes), or comparing officers' ratings to researchers' file coding. It is arguably more informative to know that two front-line officers are similar in their perceptions of the same case, but it is also rare in practice for two officers to manage the same client, making the assessment of inter-rater reliability neither easily feasible nor representative of real practice (Morgan et al., 2013).

In general, low reliability reduces the upper limit on the possible validity of a measure, and in one study, poor inter-rater agreement was specifically related to poor prediction of future violence (McNiel, Lam, & Binder, 2000). McNiel and colleagues (2000) showed that concordant ratings accurately predicted future offences (using assessments from either rater), but the predictive accuracy of discordant ratings (which were inaccurate using either rater's assessments) indicated that some feature inherent in the ambiguity of inharmonious cases degraded the assessment's predictive validity for those cases. Thus, poor inter-rater reliability may represent (a) unclear conceptual development in a tool, (b) vague

transfer of knowledge to raters, or, if the tool is generally accurate when used by that sample of raters, (c) an indication that the tool is being applied with an unidentified subgroup whose features may not appropriately match the assessment items. Thus, the inter-rater reliability of a risk tool is an important way to establish that those trained to rate the items have clear understanding of the conceptual underpinning of the items, and that these concepts can be unambiguously recognized when interviewing clients.

For DRAOR items, it may be particularly challenging to attain strong inter-rater reliability, because past research suggests raters find it more difficult to attain similar ratings on risk items defined by more subjective, complex content (i.e., dynamic, psychologically relevant items). For example, it was possible to train correctional staff to achieve good inter-rater reliability with most LSI-R items, except some of the most subjective items (specifically, marital/family relationships and financial problems; Austin et al., 2003). Similarly, Lindsay and colleagues (2004) showed poor inter-rater reliability only for the most psychologically relevant dynamic risk variable in their assessment (therapeutic alliance). Similar to McNiel and colleagues (2000), Mills and Kroner (2006) found when results from multiple risk measures were discordant (defined as the offender's relative standing on the measure, not necessarily the offender's predicted likelihood of recidivism), scores had a weaker association with recidivism. However, this occurred only for risk measures that included dynamic items, which may suggest that rating dynamic items is more problematic, due to subjective item content (Mills & Kroner, 2006; see also Gottfredson & Moriarty, 2006). It may be that raters trained on a risk tool with

strong conceptual development can still have difficulty attaining agreement on dynamic items, if they are unlikely to recognize their prototypical understanding of an item within clients' complex lives.

On the other hand, although it is logical that dynamic, subjective items are more challenging to rate compared to static, fact-based items, a number of field studies support the conclusion that practitioners can reliably rate dynamic, clinical items (Hanson et al., 2007; Morgan et al., 2013; Müller-Isberner et al., 2007; Philipse et al., 2006). Specifically, Hanson and colleagues (2007) demonstrated that trained supervision officers were able to assess dynamic information that predicted recidivism, and researchers were able to attain comparable scores when conducting an independent retrospective file review. Using a prospective design, Jones and colleagues (2010) compared relatively quick ratings assessed by supervision officers to a more exhaustive protocol assessed by researchers with the same offender clients. Their results suggested ratings between researchers and officers were highly correlated for situational triggers, moderately correlated for stable features (such as substance use, criminal associates, and level of compliance), but non-significantly or weakly correlated for cognitions and emotional states. Even when correlations were weak, ratings from supervision officers predicted recidivism to the same degree as scores from the researchers' protocols. These results suggest that supervision officers are able to accurately rank offenders on risk-relevant variables that predict recidivism, even when there is evidence that officers are not assessing the same constructs the researchers assessed. However, note that Jones and colleagues (2010) did not explicitly examine inter-rater

reliability (since the researchers and officers rated each construct with different measures).

Thus, there is evidence that trained front-line staff can accurately rate subjective items. Still, it should be considered a limitation of the present research that it is unknown whether supervision officers can rate DRAOR items similar to each other, or similar to how a researcher would rate the items. This casts doubt on whether item scores reflect the constructs they intend to measure, limiting statements about the theoretical relationships between the constructs assessed within DRAOR and recidivism outcomes. However, as demonstrated by Jones and colleagues (2010), accurate prediction may still occur, even when the underlying construct of the measurement is unknown (i.e., when there is poor evidence of construct validity, given weak correlations with other measures of the intended construct).

Summary.

Understanding the implementation of correctional practices in field settings is a worthy independent program of research. Research on DRAOR has so far focused on its utility, its psychometric properties, and predictive validity, without attention to implementation fidelity, and for this reason, conclusions using the present dataset may be limited by ambiguity regarding how DRAOR has been used in practice. In particular, supervision officers are likely different from each other in how they each implement the DRAOR into their practice. Yet, because these DRAOR data are not grouped according to supervision officers, this project is unable to estimate these differences in fidelity. Furthermore, because research has not

examined whether two officers rate the same offender with equal ratings on DRAOR items or subscales, it is unknown whether rating discordance may limit DRAOR's predictive effects. Overall, however, these limitations are more likely to result in weakening the potential to find strong prediction with DRAOR scores, rather than artificially inflating its predictive effects.

Strengths

Although the limitations discussed above temper uncritical acceptance of the results that follow in the next chapters, these limitations are also balanced by the high value of field research. Data collected from correctional staff not specifically trained for research purposes are helpful for estimating the degree of effectiveness that can be obtained in real-world practice. Since the goal of this research is to study the ability of supervision officers to measure real-time changes in risk variables, it is appropriate to examine real-world practices rather than rely on a more controlled, but artificial research study. Much existing knowledge in correctional research arose from demonstrating a new practice under ideal conditions, with subsequent research showing that the effects degrade when implemented with front-line correctional staff. Arguably, there is also value in implementing a new practice in the field, then determining what additional training or technical assistance may be required to increase effectiveness.

In addition to recognizing the general value of field research, this project is substantially unique for (a) including a jurisdiction's entire population of offenders on post-prison, parole supervision, (b) examining a contemporary cohort of offenders, (c) incorporating multiple (more than two) within-individual

assessments, (d) including a measure that assesses multiple constructs identified as important by different theoretical traditions, each which also has psychological relevance for understanding offending, and (e) using assessments from a community setting, taken proximal to recidivism events, which allows the approximation of “real-time” risk assessment.

Chapter 4: Basic Model Building

The purpose of this chapter is to present initial results that lay the foundation and provide justification for the more complex models presented in Chapter 5. This chapter presents (a) descriptive information, (b) tests of model assumptions, and (c) simple prediction models using time-varying predictors. This chapter may specifically appeal to readers who seek to become more familiar with the logic of preparing longitudinal re-assessment data for analysis with Cox regression models, and may help such readers determine whether this statistical approach is appropriate for their data. Readers who are already familiar with Cox regression with time-varying predictors may prefer Chapter 5, which presents the results of tests examining the main hypothesis of this dissertation, by examining whether models using proximal assessments are more closely associated with recidivism outcomes compared to models using only relatively distal assessments.

All analyses (in this chapter, and Chapter 5) were conducted using either SPSS (version 22) or *R* (version 3.1.3), and for many analyses, both programs were used.

Descriptive Statistics

The descriptive statistics displayed in this section outline a general overview of the data, which provides important context for interpreting the results of the complex statistical models presented later. In particular, this section describes (a) sample recidivism rates, and how they changed across different periods during the follow-up, (b) decisions regarding how DRAOR scores were combined and imputed, (c) shared variance among DRAOR subscales, (d) baseline DRAOR scores, and how scores changed at re-assessment across the follow-up, on average, and (e) simple explorations of individual differences in DRAOR score changes across time.

Life Table

Table 4.1 presents a *life table* of the dataset. A life table describes participant outcomes, by ordering observed recidivism events as they occurred across the time frame of the study follow-up. Note that the one-year follow-up has been divided into 52 discrete time periods, with each period defined as one week. The nomenclature in the first column of Table 4.1, i.e., “[1,2]”, indicates that participants listed in that row were observed during Week 1, inclusively, and events listed in that row occurred prior to the beginning of Week 2. A participant observed within Week 1 may have (a) experienced a recidivism event, (b) been censored, meaning they were lost to further follow-up without experiencing a recidivism event, or (c) did not experience a recidivism event, but were again observed in Week 2.

The second column in Table 4.1 provides the number of re-entry sequences that were observed in each week, which were only included if neither recidivism nor

Table 4.1

Life Table Describing the Number of Weeks Offenders Remained in the Community Following Release From Incarceration During a One-year Follow-up Period

Week	Entered week	Recidivated	Censored	% Recidivated	% Remaining
0	[0,1)	3654	-	-	1.000
1	[1,2)	3654	7	0.018	0.982
2	[2,3)	3580	37	0.018	0.964
3	[3,4)	3476	57	0.016	0.949
4	[4,5)	3363	40	0.019	0.931
5	[5,6)	3258	36	0.016	0.916
6	[6,7)	3170	41	0.019	0.899
7	[7,8)	3070	36	0.019	0.882
8	[8,9)	2976	41	0.020	0.864
9	[9,10)	2875	45	0.017	0.849
10	[10,11)	2782	38	0.023	0.829
11	[11,12)	2681	24	0.014	0.817
12	[12,13)	2619	40	0.013	0.806
13	[13,14)	2544	16	0.018	0.791
14	[14,15)	2481	10	0.017	0.778

15	[15,16)	2428	46	28	0.019	0.763
16	[16,17)	2354	35	54	0.015	0.752
17	[17,18)	2265	41	46	0.018	0.738
18	[18,19)	2178	35	36	0.016	0.726
19	[19,20)	2107	29	28	0.014	0.716
20	[20,21)	2050	34	37	0.017	0.704
21	[21,22)	1979	23	26	0.012	0.696
22	[22,23)	1930	26	33	0.013	0.687
23	[23,24)	1871	29	40	0.015	0.677
24	[24,25)	1802	18	39	0.010	0.670
25	[25,26)	1745	25	51	0.014	0.661
26	[26,27)	1669	27	89	0.016	0.650
27	[27,28)	1553	28	119	0.018	0.638
28	[28,29)	1406	18	74	0.013	0.630
29	[29,30)	1314	9	25	0.007	0.626
30	[30,31)	1280	15	30	0.012	0.618
31	[31,32)	1235	13	23	0.011	0.611
32	[32,33)	1199	14	27	0.012	0.604
33	[33,34)	1158	7	32	0.006	0.600

34	[34,35)	1119	8	28	0.007	0.596
35	[35,36)	1083	6	22	0.006	0.592
36	[36,37)	1055	7	26	0.007	0.588
37	[37,38)	1022	5	26	0.005	0.585
38	[38,39)	991	12	23	0.012	0.578
39	[39,40)	956	9	25	0.009	0.573
40	[40,41)	922	10	24	0.011	0.567
41	[41,42)	888	11	22	0.012	0.560
42	[42,43)	855	5	21	0.006	0.557
43	[43,44)	829	4	24	0.005	0.554
44	[44,45)	801	5	20	0.006	0.551
45	[45,46)	776	5	18	0.006	0.547
46	[46,47)	753	1	18	0.001	0.547
47	[47,48)	734	7	17	0.010	0.541
48	[48,49)	710	4	26	0.006	0.538
49	[49,50)	680	4	25	0.006	0.535
50	[50,51)	651	6	21	0.009	0.530
51	[51,52)	624	3	18	0.005	0.527
52	[52,53)	603	2	23	0.003	0.526

Note. $N = 3654$ sequences. Recidivism = any new crime or technical violation of supervision

order. Sequences with a recidivism event after Week 52 were considered censored at Week 52.

censoring had occurred in a prior week. The next two columns list the number of recidivism events and censored sequences that occurred in that week.

The final two columns provide (a) the proportion of recidivism events observed in each week, and (b) the cumulative proportion of re-entry sequences in which participants continued to remain in the community without sanction. The first of these two columns displays the hazard function. These values are graphed in Figure 4.1, along with a regression line of best fit. Note that the hazard (observed recidivism rates) remains small throughout the follow-up period, and declines at a generally steady rate across the 52 weeks. Since the discrete time periods were brief, hazard rates were low.

The second of the final two columns displays the survivor function (graphed in Figure 4.2). Note that at the end of the one-year follow-up, approximately half (52.6%) of the sample had not experienced a recidivism event, and half had recidivated. At three months, the estimated survival rate (the rate of remaining crime-free) was 79.1%, at six months, the rate was 65.0%, and at nine months, the rate was 57.3%.

Finally, also note that at least one recidivism event was observed during each of the 52 weeks of the follow-up period. As discussed in Chapter 2, the most straightforward way to set up data for analysis using a Cox regression model with time-varying predictors requires that researchers (a) divide the follow-up into discrete time periods, and (b) choose discrete periods such that at least one recidivism event was observed in each time period.

Figure 4.1. Observed hazard rate at each week of the follow up (with values presented as *circles*), with a regression line of best fit (*squares*) superimposed onto the graph.

Figure 4.2. Observed cumulative proportion of non-recidivists across the one-year follow-up period.

Central Tendency and Imputation Strategies

Table 4.2 displays the means and standard deviations of DRAOR subscales within each week of the follow-up. Recall from Chapter 3 that DRAOR assessments described in Table 4.2 and the recidivism events displayed in Table 4.1 were staggered, meaning that decision rules required participants to have a recorded DRAOR assessment in the week *prior* to the week of any observed recidivism. Since supervision officers did not see all participants on a weekly schedule, the final column of Table 4.2 also indicates the proportion of observations within each week that represent truly observed scores (i.e., not imputed).

As previously reported in Chapter 3, note that approximately one-third of the re-entry sequences included a DRAOR observation at baseline (on the day of release, or within 30 days prior to release). However, for two-thirds of the sample, it was necessary to impute baseline scores by using the first observed DRAOR scores of the re-entry sequence, as long as these assessments occurred within 30 days of release.

Score imputation was also required during the subsequent weeks of follow-up. Recall from Chapter 2 that Cox regression models with time-varying predictors evaluate whether recidivists' scores at the time of recidivism differ systematically from the scores of non-recidivists. When using a discrete-time model, the implication is that every non-censored participant must have a score within each discrete time period that occurs prior to censoring or recidivism. The purpose of this project was to examine assessment proximity, so a critical consideration was whether assessments were proximal to recidivism events. As discussed in Chapter 3, a participant who was not observed for a period lasting more than six weeks

Table 4.2

Descriptive Statistics (Means and Standard Deviations) for DRAOR Subscales Displayed for Each Week of the One-year Follow-up Period, With Percentage of Non-imputed Values

Week	<i>n</i>	Stable	Acute	Protect	% of weeks with observed values
		<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>	
1	3654	6.39 (2.54)	5.94 (2.42)	6.02 (2.40)	33.5%
2	3580	6.29 (2.57)	5.77 (2.37)	6.13 (2.42)	84.9%
3	3476	6.14 (2.68)	5.51 (2.32)	6.31 (2.47)	85.3%
4	3363	6.00 (2.72)	5.31 (2.31)	6.47 (2.47)	85.8%
5	3258	5.86 (2.74)	5.16 (2.29)	6.61 (2.49)	84.3%
6	3170	5.74 (2.78)	5.06 (2.31)	6.74 (2.51)	81.1%
7	3070	5.64 (2.81)	4.95 (2.30)	6.85 (2.51)	78.9%
8	2976	5.55 (2.82)	4.83 (2.28)	6.94 (2.51)	77.3%
9	2875	5.44 (2.82)	4.75 (2.27)	7.03 (2.49)	74.6%
10	2782	5.37 (2.83)	4.69 (2.27)	7.08 (2.49)	73.9%
11	2681	5.32 (2.84)	4.65 (2.31)	7.13 (2.49)	70.7%
12	2619	5.26 (2.82)	4.58 (2.31)	7.22 (2.48)	72.4%
13	2544	5.20 (2.83)	4.56 (2.32)	7.28 (2.50)	67.3%
14	2481	5.14 (2.84)	4.50 (2.31)	7.36 (2.51)	66.5%
15	2428	5.07 (2.84)	4.43 (2.28)	7.44 (2.48)	65.7%
16	2354	5.00 (2.83)	4.35 (2.26)	7.51 (2.50)	66.4%
17	2265	5.00 (2.82)	4.34 (2.26)	7.56 (2.51)	63.4%
18	2178	4.94 (2.82)	4.28 (2.26)	7.62 (2.52)	64.3%
19	2107	4.87 (2.83)	4.21 (2.23)	7.68 (2.52)	60.3%
20	2050	4.82 (2.83)	4.19 (2.25)	7.73 (2.51)	62.0%
21	1979	4.78 (2.82)	4.14 (2.25)	7.77 (2.52)	60.4%
22	1930	4.69 (2.80)	4.11 (2.21)	7.85 (2.49)	60.5%

23	1871	4.66 (2.79)	4.06 (2.18)	7.89 (2.47)	59.7%
24	1802	4.62 (2.79)	3.98 (2.18)	7.95 (2.45)	59.7%
25	1745	4.58 (2.80)	3.92 (2.21)	7.99 (2.44)	57.4%
26	1669	4.52 (2.77)	3.84 (2.16)	8.06 (2.42)	59.9%
27	1553	4.43 (2.73)	3.77 (2.13)	8.12 (2.40)	58.8%
28	1406	4.24 (2.66)	3.68 (2.09)	8.28 (2.32)	57.3%
29	1314	4.15 (2.65)	3.62 (2.12)	8.35 (2.31)	55.8%
30	1280	4.11 (2.65)	3.56 (2.08)	8.37 (2.32)	55.2%
31	1235	4.03 (2.59)	3.52 (2.05)	8.43 (2.29)	55.4%
32	1199	3.99 (2.59)	3.49 (2.06)	8.46 (2.28)	54.1%
33	1158	3.96 (2.59)	3.44 (2.00)	8.52 (2.27)	55.8%
34	1119	3.93 (2.57)	3.41 (2.03)	8.56 (2.26)	55.3%
35	1083	3.92 (2.57)	3.40 (2.05)	8.58 (2.23)	54.0%
36	1055	3.89 (2.58)	3.39 (2.07)	8.62 (2.25)	50.4%
37	1022	3.86 (2.60)	3.35 (2.02)	8.64 (2.26)	55.8%
38	991	3.85 (2.62)	3.35 (2.06)	8.64 (2.28)	51.2%
39	956	3.81 (2.60)	3.33 (2.10)	8.70 (2.26)	54.4%
40	922	3.74 (2.57)	3.32 (2.08)	8.75 (2.27)	49.9%
41	888	3.72 (2.56)	3.26 (2.03)	8.75 (2.25)	52.7%
42	855	3.67 (2.54)	3.21 (1.99)	8.78 (2.21)	52.5%
43	829	3.65 (2.54)	3.20 (2.02)	8.80 (2.22)	51.0%
44	801	3.62 (2.55)	3.17 (2.00)	8.82 (2.23)	51.8%
45	776	3.60 (2.53)	3.15 (2.04)	8.82 (2.25)	51.7%
46	753	3.57 (2.48)	3.14 (2.06)	8.83 (2.24)	51.1%
47	734	3.52 (2.44)	3.09 (2.02)	8.87 (2.24)	52.0%
48	710	3.50 (2.42)	3.11 (2.04)	8.90 (2.23)	52.0%
49	680	3.52 (2.45)	3.10 (2.05)	8.90 (2.22)	52.2%
50	651	3.49 (2.47)	3.09 (2.09)	8.90 (2.25)	52.8%
51	624	3.48 (2.47)	3.05 (2.03)	8.91 (2.30)	52.2%

52	603	3.47 (2.47)	3.05 (2.00)	8.93 (2.28)	51.7%
----	-----	-------------	-------------	-------------	-------

Note. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale.

during the follow-up year was considered censored at the last observation prior to that gap in time. Thus, this project defines *proximity* as “assessments occurring within at least six weeks”.

After implementing this decision rule, as shown in Table 4.2, imputation was required for between 15-50% of re-entry sequences within each week, starting the week after baseline. Compared to early in the follow-up period, as the follow-up year proceeded, assessments were less consistently recorded on a weekly basis, and relatively more imputation was required later in the year.

Recall from Chapter 3 that that dataset used for analysis was built in six different ways with six different imputation strategies, in order to explicitly compare the strategies, and use a data-driven approach to decide which imputation approach should be used for complex model-building (as recommended by Singer & Willett, 2003). The conceptual differences between these six choices were listed previously in Table 3.1. The results from comparing model fit for these six strategies are listed in Table 4.3, which is structured to mirror Table 3.1. In Table 4.3, bolded values indicate the model of best fit, as indicated by the smallest AIC (Akaike Information Criterion) value.

As displayed in Table 4.3, when more than one DRAOR assessment occurred within the same week, model fit was always better when the last assessment of the week was preferred over taking the mean of all the assessments within that week. In other words, choosing the DRAOR assessment that occurred on the seventh day of a one-week period is preferred over averaging that assessment with an assessment from the second day of that same week. Next, note that for two DRAOR

Table 4.3

Model Fit Indices For Cox Regression Time-Varying Predictor Models Using DRAOR Subscales to Predict Time to Recidivism, Comparing Six Dataset Building Strategies

Strategies for Combining Multiple Assessments in the Same Week	Strategies for Imputing Missing Weeks		
	Imputing the Most Recent Assessment	Imputing the Average of the Most Recent 2 Assessments	Imputing the Average of the Most Recent 4 Assessments
Choosing the Last Assessment of the Week	Stable AIC = 13,580.13	Stable AIC = 13,583.44	Stable AIC = 13,582.64
	Acute AIC = 13,512.21	Acute AIC = 13,508.42	Acute AIC = 13,507.91
	Protect AIC = 13,579.83	Protect AIC = 13,581.22	Protect AIC = 13,580.54
Choosing the Mean of all Assessments in the Week	Stable AIC = 13,584.46	Stable AIC = 13,587.54	Stable AIC = 13,587.07
	Acute AIC = 13,517.07	Acute AIC = 13,515.37	Acute AIC = 13,515.30
	Protect AIC = 13,583.89	Protect AIC = 13,586.19	Protect AIC = 13,586.44

Note. $n = 92,104$ sequences, recidivism $n = 1344$ events. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). AIC = Akaike Information Criterion. **Bolded values** indicate the best model fit for that subscale. Recidivism = any new crime or technical violation of supervision order.

subscales (*stable* and *protect*), model fit was at its best when a week without a DRAOR observation was imputed with only the most recent assessment (i.e., the DRAOR assessment from the week prior). By contrast, for the *acute* subscale, model fit was slightly improved when the average of the most recent four weeks was used to impute a missing week.

Thus, for two of the three DRAOR subscales, results displayed in Table 4.3 suggest that analyses should proceed with the dataset using a *carry-the-prior-week-forward* strategy for imputation. In order to maintain consistency in interpretation, this imputation strategy was also used for analyses that included the *acute* DRAOR subscale. Thus, all analyses presented in this dissertation resulted from the dataset using these selection and imputation strategies (i.e., selecting the last assessment of the 7-day period when multiple assessments per week were recorded, and imputing missing weeks with the scores from the most recent week). When analyses were conducted using any of the other five dataset-building strategies, results did not differ substantially, and conclusions did not differ.

Thus, the mean subscale scores displayed across the 52 weeks of follow-up in Table 4.2 also include imputed values. On average, DRAOR risk scores (*stable* and *acute*) declined steadily throughout the follow-up period. Similarly, *protect* scores steadily changed in the direction of decreasing risk, by increasing across time, on average.

Subscale Correlations and Internal Consistency

The descriptive information displayed in Table 4.4 shows subscale internal consistencies (Cronbach's alpha), inter-subscale correlations, and correlations

Table 4.4

Psychometric Properties (Internal Consistency, Inter-subscale Correlations, and Correlations with Static Risk) of DRAOR Scores, Displayed for Each Week of the One-year Follow-up Period

Week	<i>n</i>	Cronbach's α			Pearson's <i>r</i>			Pearson's <i>r</i>		
		Stable	Acute	Protect	Stable & Acute	Stable & Protect	Acute & Protect	Stable & RoC*RoI	Acute & RoC*RoI	Protect & RoC*RoI
1	3654	0.80	0.63	0.84	.54	-.59	-.46	.31	.23	-.24
2	3580	0.80	0.61	0.84	.53	-.60	-.44	.31	.24	-.24
3	3476	0.81	0.59	0.83	.54	-.62	-.46	.31	.22	-.24
4	3363	0.81	0.59	0.83	.54	-.62	-.45	.30	.21	-.24
5	3258	0.81	0.58	0.83	.54	-.64	-.46	.30	.20	-.24
6	3170	0.81	0.59	0.83	.55	-.63	-.46	.29	.20	-.24
7	3070	0.81	0.59	0.82	.54	-.63	-.45	.28	.19	-.24
8	2976	0.81	0.57	0.82	.53	-.64	-.46	.27	.19	-.23
9	2875	0.81	0.57	0.81	.53	-.63	-.46	.26	.19	-.23
10	2782	0.81	0.57	0.81	.53	-.62	-.46	.26	.18	-.22
11	2681	0.81	0.59	0.81	.54	-.63	-.47	.26	.18	-.21
12	2619	0.81	0.59	0.81	.54	-.63	-.46	.27	.18	-.22

13	2544	0.81	0.59	0.81	.54	-.64	-.47	.26	.18	-.22
14	2481	0.81	0.59	0.81	.54	-.64	-.48	.26	.19	-.22
15	2428	0.81	0.59	0.81	.53	-.65	-.48	.26	.19	-.22
16	2354	0.81	0.58	0.81	.54	-.66	-.50	.25	.18	-.21
17	2265	0.81	0.58	0.81	.54	-.66	-.50	.25	.18	-.21
18	2178	0.81	0.58	0.81	.54	-.66	-.49	.25	.18	-.21
19	2107	0.81	0.57	0.81	.54	-.67	-.50	.25	.19	-.21
20	2050	0.81	0.58	0.81	.54	-.67	-.50	.24	.19	-.21
21	1979	0.81	0.58	0.81	.54	-.66	-.49	.24	.20	-.20
22	1930	0.80	0.56	0.80	.53	-.65	-.50	.24	.20	-.19
23	1871	0.80	0.55	0.80	.53	-.66	-.50	.23	.19	-.19
24	1802	0.80	0.55	0.79	.51	-.66	-.49	.23	.19	-.18
25	1745	0.80	0.56	0.79	.52	-.66	-.49	.23	.18	-.18
26	1669	0.80	0.54	0.79	.52	-.65	-.49	.24	.17	-.17
27	1553	0.79	0.54	0.79	.51	-.63	-.49	.23	.16	-.16
28	1406	0.78	0.52	0.77	.51	-.62	-.47	.20	.15	-.14
29	1314	0.78	0.53	0.77	.52	-.62	-.47	.22	.15	-.14
30	1280	0.78	0.52	0.77	.51	-.61	-.46	.21	.14	-.13
31	1235	0.77	0.51	0.76	.49	-.60	-.45	.20	.13	-.12

32	1199	0.77	0.52	0.76	.50	-.61	-.46	.20	.14	-.12
33	1158	0.77	0.49	0.76	.49	-.61	-.44	.19	.14	-.11
34	1119	0.76	0.51	0.76	.51	-.61	-.45	.21	.15	-.12
35	1083	0.77	0.51	0.75	.50	-.61	-.45	.20	.15	-.11
36	1055	0.77	0.53	0.76	.51	-.62	-.47	.20	.15	-.11
37	1022	0.77	0.50	0.76	.50	-.62	-.46	.20	.16	-.11
38	991	0.78	0.53	0.76	.50	-.62	-.46	.21	.15	-.09
39	956	0.77	0.54	0.76	.51	-.62	-.47	.20	.14	-.09
40	922	0.77	0.53	0.76	.49	-.62	-.46	.19	.13	-.09
41	888	0.77	0.52	0.76	.48	-.61	-.44	.19	.13	-.08
42	855	0.77	0.50	0.75	.47	-.60	-.43	.17	.11	-.07
43	829	0.77	0.50	0.75	.46	-.61	-.43	.17	.11	-.08
44	801	0.77	0.51	0.76	.48	-.60	-.45	.17	.11	-.07
45	776	0.77	0.53	0.76	.46	-.60	-.43	.16	.12	-.07
46	753	0.76	0.54	0.76	.46	-.59	-.42	.17	.11	-.07
47	734	0.75	0.53	0.76	.44	-.58	-.41	.18	.10	-.07
48	710	0.75	0.53	0.76	.44	-.58	-.41	.17	.10	-.08
49	680	0.75	0.53	0.76	.45	-.58	-.42	.17	.10	-.08
50	651	0.75	0.56	0.77	.44	-.59	-.42	.17	.08	-.09

51	624	0.75	0.53	0.78	.43	-.59	-.43	.17	.08	-.10
52	603	0.76	0.52	0.77	.44	-.60	-.45	.18	.09	-.11

Note. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale. RoC*RoI = Risk of Reconviction* Risk of Re-imprisonment (Bakker et al., 1999).

between subscales and static risk (RoC*RoI) scores. At baseline, internal consistencies were similar to Cronbach's alpha estimates using DRAOR data drawn from another jurisdiction (i.e., Iowa; see Chadwick, 2014, or refer back to Table 3.2). Across the follow-up period, internal consistencies degraded, but remained in the acceptable to good range for the *stable* and *protect* subscales, and were always in the weak range for the *acute* subscale.

Observed correlations between subscale scores were all in the moderate range. The magnitude of correlations remained largely consistent throughout the follow-up period, with some fluctuation. Also, correlations between static risk (RoC*RoI) and DRAOR subscales began at baseline in the low-to-moderate range, but weakened over the follow-up period.

Change Across Time, Intraclass Correlations, and Stability Coefficients

As mentioned above, as participants remained in the community, DRAOR risk scores decreased, on average. However, there were individual differences in the amount and direction of change. Table 4.5 displays the proportion of week-to-week changes in scores for each DRAOR subscale. In particular, *stable* and *protect* scores remained generally unchanging throughout the follow-up period, with about 85-95% individual-level scores typically remaining constant when comparing any week to the next. Scores on the *acute* subscale demonstrated slightly more change, with about 65-90% of individual-level scores remaining constant between any two weeks.

Among the relatively small proportion of participants demonstrating individual-level change between any two weeks, risk scores more typically

Table 4.5

Proportion of Change Across DRAOR Subscale Assessments, Displayed For Each Week-to-Week Transition During the One-year Follow-up Period

Week	Stable			Acute			Protect		
	Increase	Decrease	No Change	Increase	Decrease	No Change	Increase	Decrease	No Change
1-2	2.0%	4.9%	93.1%	3.7%	10.3%	86.0%	5.4%	1.7%	92.9%
2-3	6.0%	11.9%	82.1%	10.1%	22.0%	67.9%	13.0%	4.7%	82.3%
3-4	5.4%	10.3%	84.3%	10.5%	20.2%	69.3%	11.7%	4.6%	83.7%
4-5	4.5%	10.4%	85.1%	10.5%	18.1%	71.4%	10.0%	4.3%	85.7%
5-6	4.6%	8.8%	86.6%	10.6%	16.1%	73.3%	9.3%	4.0%	86.7%
6-7	3.9%	8.1%	88.0%	9.8%	15.1%	75.1%	9.1%	4.1%	86.8%
7-8	4.1%	7.2%	88.7%	9.5%	15.0%	75.5%	8.0%	3.8%	88.2%
8-9	3.4%	7.4%	89.2%	9.7%	14.1%	76.2%	7.3%	4.3%	88.3%
9-10	3.6%	6.7%	89.7%	9.6%	13.8%	76.6%	6.6%	3.6%	89.9%
10-11	4.2%	5.1%	90.7%	9.8%	12.0%	78.2%	6.8%	4.0%	89.1%
11-12	3.3%	6.2%	90.5%	9.4%	13.8%	76.8%	6.9%	2.9%	90.2%
12-13	3.8%	5.4%	90.8%	9.3%	10.5%	80.2%	5.7%	3.4%	90.9%
13-14	3.3%	5.8%	90.9%	8.4%	12.0%	79.6%	6.1%	2.7%	91.2%
14-15	2.6%	5.2%	92.2%	8.0%	11.2%	80.8%	5.3%	3.3%	91.4%

15-16	2.9%	5.2%	91.9%	7.5%	11.6%	80.9%	5.5%	2.9%	91.6%
16-17	3.5%	4.3%	92.2%	9.3%	10.3%	80.4%	4.9%	2.8%	92.3%
17-18	2.3%	4.4%	93.3%	7.8%	10.1%	82.1%	4.2%	2.8%	93.1%
18-19	2.3%	3.9%	93.8%	7.2%	9.5%	83.3%	4.3%	2.4%	93.3%
19-20	2.9%	4.1%	93.0%	8.1%	9.9%	82.0%	4.2%	2.6%	93.2%
20-21	2.4%	4.5%	93.1%	6.8%	9.1%	84.1%	4.0%	2.0%	93.9%
21-22	2.2%	4.5%	93.3%	7.0%	8.3%	84.7%	4.5%	2.1%	93.4%
22-23	2.1%	2.9%	95.0%	7.0%	9.3%	83.7%	2.9%	2.3%	94.8%
23-24	2.1%	2.8%	95.1%	5.3%	10.0%	84.7%	3.7%	1.8%	94.5%
24-25	1.8%	3.1%	95.1%	5.4%	8.8%	85.8%	3.0%	1.7%	95.2%
25-26	1.9%	3.0%	95.1%	6.1%	10.1%	83.8%	4.0%	2.2%	93.8%
26-27	1.8%	3.5%	94.7%	5.9%	8.0%	86.1%	2.8%	1.8%	95.4%
27-28	1.5%	3.8%	94.7%	5.6%	8.3%	86.1%	3.1%	1.3%	95.6%
28-29	1.4%	3.1%	95.5%	5.8%	7.4%	86.8%	2.5%	1.5%	96.0%
29-30	1.4%	2.1%	96.5%	5.2%	7.0%	87.8%	1.6%	1.1%	97.3%
30-31	1.6%	3.3%	95.1%	6.8%	7.1%	86.1%	2.3%	1.2%	96.5%
31-32	1.5%	2.6%	95.9%	5.1%	7.6%	87.3%	2.4%	1.8%	95.7%
32-33	1.7%	2.4%	95.9%	5.3%	7.5%	87.2%	1.7%	1.5%	96.8%
33-34	1.5%	2.2%	96.3%	6.2%	8.1%	85.7%	1.8%	1.0%	97.2%

34-35	1.5%	1.6%	96.9%	4.6%	5.8%	89.6%	2.1%	0.6%	97.2%
35-36	1.0%	1.8%	97.2%	5.5%	7.7%	86.8%	2.2%	0.7%	97.2%
36-37	1.2%	2.3%	96.5%	4.4%	7.6%	88.0%	1.9%	1.0%	97.2%
37-38	1.2%	1.6%	97.2%	4.1%	5.3%	90.6%	1.6%	0.8%	97.6%
38-39	1.5%	2.5%	96.0%	4.5%	5.8%	89.7%	2.5%	0.9%	96.5%
39-40	1.0%	2.0%	97.0%	4.8%	5.1%	90.1%	2.1%	0.3%	97.6%
40-41	1.8%	2.5%	95.7%	4.8%	7.8%	87.4%	1.4%	0.9%	97.7%
41-42	1.4%	1.8%	96.8%	4.1%	6.2%	89.7%	1.5%	0.8%	97.7%
42-43	1.4%	1.0%	97.6%	3.7%	3.8%	92.5%	1.2%	0.2%	98.6%
43-44	0.9%	2.6%	96.5%	4.5%	6.1%	89.4%	1.7%	0.7%	97.5%
44-45	0.9%	1.5%	97.6%	3.9%	5.6%	90.5%	1.4%	0.8%	97.8%
45-46	1.1%	2.0%	96.9%	5.0%	5.5%	89.5%	1.9%	1.1%	97.1%
46-47	1.4%	0.9%	97.7%	4.0%	5.0%	91.0%	1.2%	1.1%	97.7%
47-48	0.7%	1.7%	97.6%	4.2%	4.4%	91.4%	1.4%	0.1%	98.5%
48-49	1.3%	2.1%	96.6%	4.0%	4.6%	91.5%	1.2%	1.0%	97.8%
49-50	1.7%	2.0%	96.3%	4.5%	6.6%	88.9%	0.9%	1.1%	98.0%
50-51	1.3%	1.9%	96.8%	5.1%	5.6%	89.3%	1.4%	1.4%	97.1%
51-52	0.8%	1.5%	97.7%	4.5%	4.6%	90.9%	1.3%	0.8%	97.8%

Note. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale. Totals may not add exactly to 100% due to rounding.

decreased rather than increased, with risk reduction occurring at approximately twice the rate of risk increases. Still, DRAOR risk scores were observed to decline on average in the full sample; this was also the case when examining either recidivists or non-recidivists (defined cumulatively, i.e., non-recidivists were participants who remained crime-free throughout the entirety of the follow-up period). These results are displayed in the fifth row of Table 4.6.

First, however, note in Table 4.6 that at baseline, non-recidivists were rated lower in risk, and higher in protective factor scores, compared to eventual recidivists. Thus, at release, future recidivists were rated with scores approximately one point higher in risk on DRAOR subscales, compared to future non-recidivists. By definition, recidivists exited the follow-up period earlier than non-recidivists, and also demonstrated less mean change, when comparing first assessments in re-entry sequences to final assessments. Note that the amount of time between first and last assessments varied, depending on how long each individual remained in the dataset without experiencing recidivism, or being lost to follow-up. Comparing the two groups on mean change, non-recidivists were rated as having approximately twice as much decline in risk, compared to recidivists. Thus, although both groups, on average, demonstrated decreases in risk scores across time, from baseline to censoring, non-recidivists demonstrated more substantial decreases in risk (i.e., more than one point in each DRAOR subscale, compared to approximately half a point).

Also shown in Table 4.6, mean net change quantifies the average amount of DRAOR subscale change from week-to-week. Here the pattern is reversed, such that

Table 4.6

Descriptive Statistics (Observed Range, Baseline Mean, Mean Change, and Mean Net Change) Across DRAOR Assessments During the One-year Follow-up Period

Statistic	Full Sample			Non-Recidivists			Recidivists		
	Stable	Acute	Protect	Stable	Acute	Protect	Stable	Acute	Protect
Maximum subscale points	12	14	12	12	14	12	12	14	12
Observed range at initial assessment	0-12	0-14	0-12	0-12	0-14	0-12	0-12	0-14	0-12
Mean (SD) at initial assessment	6.39 (2.54)	5.94 (2.42)	6.02 (2.40)	5.95 (2.50)	5.51 (2.32)	6.40 (2.34)	7.14 (2.43)	6.67 (2.41)	5.38 (2.37)
Mean total change (SD; % of initial range)	-0.96 (2.31; 8.0%)	-1.31 (2.47; 9.4%)	1.10 (2.26; 9.2%)	-1.20 (2.37; 10.0%)	-1.59 (2.46; 11.4%)	1.35 (2.32; 11.3%)	-0.55 (2.15; 4.5%)	-0.81 (2.41; 5.8%)	0.67 (2.09; 5.6%)
Mean net change [‡] (SD; % of initial range)	0.16 (0.28; 1.3%)	0.34 (0.37; 2.4%)	0.17 (0.31; 1.4%)	0.13 (0.21; 1.1%)	0.28 (0.28; 2.0%)	0.14 (0.23; 1.2%)	0.22 (0.36; 1.8%)	0.45 (0.48; 3.2%)	0.23 (0.41; 3.4%)
% participants with any change	65.7%	86.6%	65.7%	70.0%	90.3%	69.9%	58.4%	80.3%	58.4%

Note. *N* = 3654 sequences; *n* = 2310 non-recidivists; *n* = 1334 recidivism events. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale. [‡]Average absolute change observed from week-to-week.

on average, recidivists showed about twice the amount of week-to-week change, compared to non-recidivists. Taken together, these patterns suggest that non-recidivists experienced greater overall declines in risk scores, whereas from week-to-week, recidivists experienced more instability in risk scores, albeit without demonstrating a general pattern of decline. Thus, on average, future recidivists maintained higher, but erratic scores, whereas non-recidivists showed smaller weekly changes that nevertheless indicated a systematic pattern of decrease in risk.

Taking into account the whole pattern of change across all assessments, the differences between recidivists and non-recidivists can also be displayed as differences between average regression coefficients, gathered from regression lines fit individually to each sequence. For *stable*, recidivists, $B = -0.93$ ($SD = 4.81$), showed less decline across assessments than non-recidivists, $B = -2.78$ ($SD = 7.34$). Also, smaller decreases in risk were observed for *acute*, $B = -1.08$ ($SD = 3.54$) versus $B = -3.04$ ($SD = 6.13$), and *protect*, $B = 1.31$ ($SD = 4.65$) versus $B = 3.96$ ($SD = 7.33$).

Note that the magnitude of change is presented in Table 4.6 as percentages of the total possible scores a participant can achieve on each DRAOR subscale, yet the substantive meaning of these scores is unknown. Although it is a helpful descriptive exercise to examine the degree of change in DRAOR scores, this strategy is limited because these subscales represent arbitrary metrics (see discussion in Chapter 2). In other words, it is unknown if the difference in one score on a DRAOR subscale is in actuality related to an observable difference in recidivism outcome rates, and this could only be fully established using large-scale, multi-sample research.

There are two final statistics that help quantify the nature of DRAOR change across time. First, the intraclass correlation (ICC) is the average correlation between two randomly chosen assessments measured on the same individual. Using a one-way random, average measures ICC, results for the entire sample show that two assessments from the same participant were highly related to each other. For the *stable* subscale, $ICC_1 = 0.989$, 95% CI [0.985, 0.992], for *acute*, $ICC_1 = 0.977$, 95% CI [0.968, 0.983], and for *protect*, $ICC_1 = 0.985$, 95% CI [0.980, 0.990]. As a measure of test-retest reliability, these statistics suggest that there is high consistency in DRAOR ratings across measurement occasions.

Comparing intraclass correlations for recidivists and non-recidivists demonstrated slightly more variability in DRAOR scores among recidivists compared to non-recidivists. Specifically, for eventual recidivists, intraclass correlations were slightly lower for *stable*, $ICC_1 = 0.983$, 95% CI [0.971, 0.990], *acute*, $ICC_1 = 0.959$, 95% CI [0.934, 0.979], and *protect* scores, $ICC_1 = 0.977$, 95% CI [0.962, 0.987]. For non-recidivists, there was slightly more intra-individual consistency, on *stable*, $ICC_1 = 0.991$, 95% CI [0.986, 0.994], *acute*, $ICC_1 = 0.980$, 95% CI [0.971, 0.987], and *protect*, $ICC_1 = 0.987$, 95% CI [0.982, 0.992].

Second, the stability coefficient is the average Pearson correlation within the correlation matrix comparing all DRAOR assessments across all weeks of the follow-up. The stability coefficient indicates the degree of stability in rank order across time, and high correlations suggest that (even when intra-individual change is occurring), participants remain consistently ranked, relative to each other. The stability coefficients for DRAOR subscales were high, for *stable*, $r = .80$, for *acute*, r

= .65, and for *protect*, $r = .78$. These coefficients suggest that there is some crossover in rank ordering over time, but that high-scoring individuals tend to maintain higher DRAOR scores relative to others in the sample.

Testing Model Assumptions

This section presents tests to determine if the variables in the dataset meet the assumptions of the statistical model, Cox regression with time-varying predictors. The section is divided into two sections: (a) general assumptions underlying all regression models, and (b) assumptions more specific to Cox regression models.

Assumptions of General Regression Models

Cox regression does not share many of the assumptions of general regression models; for example, predictors do not need to be characterized by a normal distribution, homoscedasticity, or linearity (Tabachnick & Fidell, 2007). However, Cox models with predictors that meet these assumptions demonstrate greater power to detect effects. Thus, it is worthwhile to examine these assumptions for descriptive purposes, and to determine if statistical power may be limited by non-linear relationships, heteroscedasticity, or departures from normality. On the other hand, assumptions regarding multivariate outliers and multicollinearity are inherent within both Cox models and general regression models, as described below.

Normality.

For each DRAOR subscale at each discrete time period, Table 4.7 displays skewness and kurtosis ratios, and the number of extreme univariate outliers (defined by a z-score of ± 3.29 and above). Bolded values within Table 4.7 indicate significant departures from normality. Throughout the follow-up, the distributions of the two risk subscales (*stable* and *acute*) showed significant positive skew. Positive skew generally became stronger as the follow-up year continued, as participant scores on average tended to decline in risk, and the few participants who maintained high DRAOR risk scores became outliers within the distribution. For the *acute* subscale, at least a few extreme outliers were observed in every time period. Scores on the *protect* subscale showed an identical pattern, except demonstrated negative skew due to items being scored in reverse to risk on that subscale.

Applying square-root transformations to the *stable* and *protect* subscales reduced skew and kurtosis, but did not fully eliminate these departures from normality. All results presented in this dissertation refer to analyses using the original, non-transformed versions of these subscales, but these models were compared to models that instead used transformed versions of these subscales. Results were essentially identical, but model fit was slightly worse when using the transformed versions of *stable* and *protect*.

By contrast, for the *acute* subscale, transformation of scores brought all outliers into range, and rendered skewness and kurtosis ratios statistically non-significant. Models using transformed and un-transformed versions of the *acute* subscale were also compared. As above, results were essentially identical, but in

Table 4.7

Statistics for Diagnosing Distribution Normality of DRAOR Subscale Scores (Skewness Ratios, Kurtosis Ratios, and Number of Extreme Outliers), Displayed for Each Week of the One-year Follow-up Period

Week	Stable			Acute			Protect		
	Skew	Kurtosis	<i>n</i> extreme outliers	Skew	Kurtosis	<i>n</i> extreme outliers	Skew	Kurtosis	<i>n</i> extreme outliers
1	0.71	-3.51	0	7.37	-1.89	4	-4.73	5.64	0
2	1.46	-4.18	0	8.32	-1.29	14	-4.51	4.77	0
3	1.48	-5.99	0	8.90	-1.30	8	-4.38	2.75	0
4	1.92	-6.57	0	9.33	-1.45	5	-4.67	1.70	0
5	2.74	-6.49	0	9.91	0.02	8	-4.02	0.56	0
6	3.40	-6.87	0	11.27	0.64	17	-4.63	0.07	0
7	3.93	-7.18	0	10.82	0.28	15	-4.57	-0.35	0
8	4.62	-6.98	0	10.58	1.44	15	-4.36	-0.90	0
9	5.09	-6.60	0	10.70	1.49	14	-4.50	-0.91	0
10	5.57	-6.53	0	10.57	1.25	14	-4.91	-0.77	0
11	5.45	-6.46	0	12.11	2.19	14	-5.32	-0.91	0
12	5.65	-6.26	0	11.60	2.06	15	-5.75	-0.79	0
13	6.10	-6.22	0	11.43	1.48	11	-6.08	-1.38	0

14	6.41	-6.10	0	11.24	1.72	9	-6.37	-1.00	0
15	6.70	-5.84	0	11.40	3.13	12	-5.44	-2.25	0
16	7.44	-5.33	0	11.34	2.69	6	-6.00	-2.19	9
17	7.04	-5.47	0	11.84	3.83	10	-6.67	-2.05	9
18	7.40	-5.15	0	12.44	5.59	10	-7.15	-1.60	9
19	7.38	-4.97	0	11.70	3.60	21	-7.09	-1.66	8
20	7.28	-4.96	0	11.46	2.95	23	-7.13	-1.57	8
21	7.35	-4.88	0	11.96	3.95	23	-7.25	-1.36	9
22	7.61	-4.50	0	11.09	3.24	20	-7.43	-1.05	8
23	7.49	-4.30	0	11.47	4.98	18	-6.72	-1.69	5
24	7.79	-4.03	0	12.12	4.97	17	-6.83	-1.43	4
25	7.90	-3.87	0	13.12	6.64	17	-6.92	-0.93	4
26	7.90	-3.47	0	12.20	5.40	10	-7.02	-0.43	5
27	7.89	-2.78	0	11.21	5.17	10	-6.94	-0.32	4
28	8.85	-1.18	0	11.95	7.65	19	-6.94	0.32	7
29	8.99	-0.90	0	11.42	6.70	18	-6.61	-0.24	6
30	9.28	-0.58	0	11.96	9.06	15	-6.65	-0.45	5
31	9.07	-0.28	9	10.34	6.64	13	-6.31	-0.58	4
32	8.63	-0.54	8	10.42	7.14	12	-6.13	-0.54	4

33	8.64	-0.32	9	9.10	3.76	8	-5.69	-1.49	2
34	8.37	-0.38	8	10.23	5.58	10	-5.84	-1.51	2
35	8.11	-0.81	7	10.35	5.03	10	-5.26	-2.24	1
36	8.05	-0.71	7	10.88	6.49	10	-5.48	-2.17	1
37	8.01	-0.86	7	10.39	6.28	10	-5.66	-2.12	1
38	8.09	-0.73	7	11.36	7.75	11	-6.09	-1.13	2
39	8.16	-0.66	6	12.75	10.78	12	-6.57	-0.63	2
40	7.83	-0.85	4	11.88	8.84	10	-6.81	-0.002	2
41	8.04	-0.49	4	12.05	10.02	9	-6.84	0.34	2
42	7.68	-0.63	3	10.60	7.43	8	-6.07	-0.51	3
43	7.73	-0.48	3	9.62	5.12	8	-5.85	-1.00	3
44	7.86	-0.18	3	9.50	4.76	6	-6.34	-0.10	4
45	7.67	-0.26	3	9.68	4.81	6	-6.92	1.05	5
46	7.28	-0.63	5	11.16	8.03	8	-6.90	1.14	5
47	7.53	0.24	5	11.21	9.70	7	-7.63	2.21	7
48	7.07	-0.13	4	11.23	9.21	6	-7.47	2.33	6
49	6.73	-0.50	4	11.76	10.29	7	-7.06	2.00	6
50	7.18	0.03	5	11.43	9.77	7	-7.08	1.41	6
51	6.77	-0.48	4	10.77	9.33	6	-7.48	1.70	7

52	6.73	-0.43	3	9.22	6.53	5	-6.90	1.14	5
----	-------------	-------	----------	-------------	-------------	----------	--------------	------	----------

Note. *N* = 3654 sequences. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale. **Bolded values** indicate substantial departures from normality. Outliers = univariate outliers.

this case, model fit was slightly improved when using the transformed version of the *acute* subscale. However, for consistency, all results presented in this dissertation are from models using the original scale (without transformations).

Since this is a large sample that represents all parolees in the jurisdiction's population of supervised offenders, it is expected that risk distributions would demonstrate some skewness. If the sample had been pre-selected to include offenders with a more homogenous range of risk, normal distributions may have been observed. Also note that the large sample size makes it unlikely that these skewed distributions would strongly impact model results.

Homoscedasticity and linearity.

There were no clear departures from homoscedasticity when visually inspecting scatterplots between predictors, and scatterplots between predictors and outcome. Cox regression does not require that predictors share a linear relationship with each other, but the model does assume there are linear relationships between continuous predictors and the outcome (specifically, the logit transformed version of the outcome; see Tabachnick & Fidell, 2007). Using the Box-Tidwell approach, results suggested all predictors shared linear relationships with the logit-transformed outcome variable. A later section in this chapter presents a second approach for testing the potential of non-linear effects; tests using this second approach occurred after first building the basic prediction models (see below).

Multivariate outliers.

Multivariate outliers were examined among all predictors (RoC*RoI, *stable*, *acute*, and *protect*) by generating Mahalanobis distances (see Tabachnick & Fidell,

2007). Mahalanobis distances identified 59 sequences as containing multivariate outliers. In 51 of these sequences, visual inspection of the patterns revealed conceptual inconsistencies between subscale ratings, for example, patterns where a *stable* risk score was particularly high, while at the same time the *acute* risk score was particularly low.

More specifically, in 20 of these 51 sequences, subscale inconsistencies occurred within the sequence immediately following release, but subscale scores later became more conceptually aligned with each other as more assessments were rated. This suggests that initial ratings were incorrect, but scoring improved after a trial period. Most typically in these 20 cases, no protective factors were identified in the weeks following release, but protective factors were rated higher a few weeks following.

For 15 of the 51 sequences, inconsistencies between DRAOR subscales likely resulted from data entry error. For these sequences, there was a short break (usually only one week) in the typical pattern of that sequence's scores. This is speculation, but the break in pattern seemed consistent with a rater's manual error.

For 10 of the 51 sequences, a short, dramatic increase (spike) in *acute* scores occurred that broke with the general pattern of the other two subscales (*stable* and *protect*). Most typically, this occurred only briefly in the sequence (a week or two) before the *acute* scores returned to their previous level. It is unknown if, but possible that, these spikes in *acute* scores represented periods where these participants maintained low risk on stable factors, high protection on protective

factors, but encountered a temporary increase in difficult situations characterized by acute factors.

Finally, for six of the 51 sequences, subscale inconsistencies were observed at the end of long assessment sequences among individuals who remained in the community for many months. These individuals were rated to simultaneously have high static RoC*RoI, *stable*, and *protect* DRAOR scores, but low *acute* scores. Conceptually, these six individuals may represent *successful high-risk individuals*, or offenders who despite being exposed to high long-term risk factors, have nevertheless been able to build enough protective factors to demonstrate low immediate (acute) risk. Potentially supporting this explanation, despite having high RoC*RoI scores, five of these individuals did not recidivate during the one-year follow-up, whereas one individual did.

Although the 51 sequences described above were characterized by conceptually inconsistent scores among the predictors, there were an additional eight sequences identified as multivariate outliers that had unusual, but consistent score patterns. These individuals were rated as having very high static risk, high stable risk, high acute risk, and few or no protective factors. In addition, all of these individuals became recidivists quickly after release. Mahalanobis distances identified this combination of scores as unusual, but because there was no evidence that these cases evidenced an unexplainable pattern, they were retained in all analyses.

Models that either included or excluded these 51 sequences with multivariate outliers were compared. Removing these sequences essentially had no

effect on results. Results presented in this dissertation were from analyses that included the multivariate outliers.

Multicollinearity.

Consistent with the correlations displayed above in Table 4.4, there was no evidence of multicollinearity among the predictors (RoC*RoI, *stable*, *acute*, and *protect*). Specifically, all tolerance values were well above 0.2.

Assumptions and Decisions Specific to Cox Regression Models

Equivalence between censored and remaining cases.

As discussed in Chapter 2, unlike other longitudinal models, the critical advantage of Cox regression models is that missing data points may be treated as meaningfully missing, not randomly missing. However, the logic underlying Cox regression models assumes that those individuals who stopped being observed within the dataset but did *not* experience recidivism (i.e., are lost to follow-up) have *not* been censored for any systematic, meaningful reason. It is apparent from the life table shown in Table 4.1 that the estimated hazard rates would become increasingly inaccurate across time if a researcher cannot extrapolate that censored individuals were similar to observed individuals who remained crime-free further into the follow-up.

Table 4.8 displays a strategy for testing this assumption, by using independent sample *t*-tests to compare censored individuals to individuals remaining in the dataset. Specifically, at each week, static and DRAOR risk scores were compared for (a) individuals who were censored during that week and (b) individuals who had not yet recidivated at that point in time. Note that this

Table 4.8

Independent Samples t-Tests Comparing the Static Risk and DRAOR Scores of Censored (During a Prior Week) and Uncensored Non-recidivists (From a Subsequent Week), Displayed for Each Week-to-Week Comparison During the One-year Follow-up Period

Week	RoC*RoI		Stable		Acute		Protect	
	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
1-2	1.82	.07	0.68	.50	-0.21	.84	1.01	.31
2-3	0.62	.54	0.78	.43	-0.72	.47	1.89	.07
3-4	-1.34	.18	2.28	.02	0.56	.58	1.73	.08
4-5	0.05	.96	0.92	.36	1.02	.31	1.17	.24
5-6	0.74	.46	1.29	.20	-0.60	.55	0.42	.68
6-7	-0.28	.78	1.52	.13	-0.56	.58	0.96	.34
7-8	0.58	.56	0.90	.37	-0.39	.70	0.22	.83
8-9	2.23	.03	0.83	.40	0.28	.78	2.08	.04
9-10	0.32	.75	0.79	.43	-0.94	.35	0.18	.86
10-11	0.75	.45	1.46	.15	0.19	.85	0.01	.99
11-12	2.51	.02	0.63	.53	0.27	.79	1.02	.31
12-13	-0.62	.54	0.76	.45	-1.14	.26	2.14	.03
13-14	-0.10	.92	0.18	.86	-1.26	.21	2.16	.05
14-15	1.42	.16	1.43	.15	1.81	.07	2.65	.008
15-16	0.67	.50	-0.14	.89	-1.21	.23	-0.38	.71
16-17	0.80	.42	-1.73	.08	-2.14	.03	0.47	.64
17-18	0.89	.38	2.00	.05	1.54	.13	1.42	.16
18-19	1.45	.15	0.96	.34	1.73	.09	1.55	.13
19-20	0.93	.35	2.84	.005	0.62	.53	1.63	.10
20-21	0.41	.69	0.49	.63	1.26	.21	0.91	.36
21-22	2.31	.02	3.07	.002	0.33	.74	3.11	.002
22-23	-1.25	.21	-0.06	.95	-0.07	.95	0.97	.33
23-24	1.38	.17	-0.19	.85	0.17	.86	0.03	.98

24-25	1.74	.08	1.81	.07	2.26	.02	1.48	.14
25-26	2.86	.004	2.64	.008	2.20	.03	2.65	.008
26-27	1.36	.18	3.77	< .001	2.26	.02	2.88	.004
27-28	4.44	< .001	5.37	< .001	2.79	.006	5.15	< .001
28-29	1.86	.06	2.83	.005	1.34	.18	2.89	.004
29-30	-0.69	.50	2.65	.008	2.77	.006	2.01	.05
30-31	0.49	.63	2.41	.02	1.92	.07	2.30	.02
31-32	0.40	.69	0.13	.90	-1.32	.19	-0.09	.93
32-33	-0.62	.53	1.34	.19	-0.05	.96	2.60	.009
33-34	1.13	.26	0.36	.72	0.47	.64	0.88	.38
34-35	0.15	.88	0.15	.88	-0.31	.76	0.64	.52
35-36	-0.44	.66	-0.25	.81	0.02	.98	-0.08	.94
36-37	1.64	.10	0.70	.49	-1.63	.10	0.47	.64
37-38	-0.14	.89	0.30	.77	0.68	.50	-0.63	.53
38-39	1.22	.22	0.79	.43	0.72	.47	2.11	.04
39-40	3.07	.002	2.38	.02	0.51	.61	2.04	.04
40-41	1.81	.07	0.37	.71	-0.50	.62	-1.06	.29
41-42	2.09	.04	1.43	.15	0.19	.84	0.57	.57
42-43	1.09	.27	1.32	.19	0.21	.83	0.05	.96
43-44	0.79	.43	0.27	.79	0.22	.83	0.31	.76
44-45	-2.11	.04	-0.18	.86	-0.83	.41	0.25	.80
45-46	0.62	.53	0.70	.48	0.64	.52	-0.87	.39
46-47	1.65	.10	2.71	.02	2.56	.02	2.66	.008
47-48	-0.68	.50	0.11	.91	-0.68	.50	0.74	.46
48-49	-0.86	.39	-1.24	.21	-0.42	.68	0.29	.78
49-50	1.31	.19	1.35	.18	0.86	.39	0.51	.61
50-51	0.01	.99	0.01	.99	0.79	.43	0.12	.90
51-52	0.84	.40	-0.25	.81	0.71	.50	0.35	.73

Note. $N = 3654$ sequences. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). RoC*RoI = Risk of Reconviction*Risk of Re-imprisonment (Bakker et al., 1999). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR

protect subscale. **Bolded values** indicate significant group differences between individuals lost to follow-up and non-recidivists remaining in the dataset. Note that a non-recidivist was defined as participants who had not recidivated *at that point in time*, rather than participants who did not recidivate at any point during the follow-up period.

approach uses an incident definition of non-recidivism, rather than a cumulative definition (i.e., a non-recidivist may later become a recidivist). In Table 4.8, bolded values indicate significant group differences, such that censored individuals were rated to have higher risk scores (or lower protect scores) compared to individuals remaining in the dataset.

Although these results suggest that the equivalence assumption does not always hold, the bolded values in Table 4.8 do not show a strongly systematic pattern of non-equivalence. The one potential exception to this statement occurs approximately halfway through the follow-up. This is largely expected, due to New Zealand's policies regarding supervision length. Recall from Chapter 3 that all incarcerated offenders are required to be supervised in the community for *at least* six months. This policy also allows early release, such that generally lower risk offenders will spend some of their sentence in the community, resulting in supervision that lasts longer than six months. Thus, it is consistent with New Zealand's policy that there was a systematic censoring of higher risk individuals at the half-year time point. During the weeks at the six-month mark, higher-risk offenders who were incarcerated for their entire sentence were legally freed from their supervision orders.

Thus, there is evidence of some departures from the assumption of equivalence between censored and remaining cases. These departures are generally not systematic, but the presence of some statistically significant comparisons suggests exercising some caution when interpreting model results.

Choosing a representation of time.

As discussed in Chapter 2, an important trade-off when choosing a general discrete-time hazard model versus a true Cox model is that the researcher must explicitly model a variable that represents *time*. Table 4.9 displays the logical process used to choose how time was represented within models presented in this dissertation.

In Table 4.9, bolded values indicate a statistically significant difference compared to models using other specifications of time. As seen in Column 4 of Table 4.9, a linear representation of *time* significantly improved upon a constant-only model. In addition, a linear representation did not result in a significantly poorer fit compared to a completely general representation of *time* (see bottom row, where *time* is represented using 52 parameters, one for each of the 52 time periods). However, a quadratic representation significantly improved upon a linear representation. Comparing AIC values across models suggests that the quadratic representation is the best approach to modeling the effect of *time* on hazard during the follow-up period, whereas BIC values marginally support a linear representation.

Figure 4.3 displays the linear and quadratic lines of best fit within a graphical display of the hazard probabilities (logit transformed). Note that this graph is identical to Figure 4.1, with the exception of mathematical transformation. The quadratic line of best fit is a slightly sloping concave curve, reaching its highest peak in the early time periods of the follow-up. The linear and quadratic lines show the most departure from each other at the end of the one-year follow-up period, when hazard is particularly low. Thus, including a quadratic representation of time when

Table 4.9

Comparison of Alternative Smooth Polynomial Representations for the Main Effect of Time in a Baseline Discrete-time Hazard Model

Representation for <i>time</i>		<i>n</i> parameters	Deviance	Δ Previous Model	χ^2 (df)	Δ General Model	χ^2 (df)	AIC	BIC
Constant		1	14031.19	-		166.89**	(51)	14033.19	14042.62
Linear		2	13922.64		108.55**	(1)	58.34	(50)	13926.64
Quadratic		3	13912.88		9.76**	(1)	48.58	(49)	13918.88
Cubic		4	13911.64		1.24	(1)	47.34	(48)	13919.64
General		52	13864.30	-				13968.30	14458.70

Note. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria. BIC was calculated using the person-period sample size ($N = 92,104$), rather than the person sample size ($N = 3654$).

** $p < .001$; **Bolded values** indicate statistically significant comparisons between models.

Figure 4.3. Logit-transformed observed hazard rates at each week of the one-year follow-up period (with logit-transformed values presented as xs), with linear (*squares*) and quadratic (*circles*) regression lines of best fit superimposed onto the graph.

modeling these data helps more accurately model discrete periods at the end of the follow-up, when likelihood of recidivism is low.

Proportionality of hazards.

A set of inter-connected assumptions common to all Cox regression models (whether including time-varying or time-invariant predictors) is that, within each time period, (a) at every value of a predictor (i.e., *DRAOR stable* has thirteen values on its scale of 0-12), hazard rates differ from each other (i.e., each subsequent score should be associated with a slightly higher likelihood of recidivism), (b) hazard functions associated with each predictor value follow the same shape across time periods (when these functions are logit transformed), and (c) the distance between these shapes remain parallel to each other throughout the follow-up period. One helpful way to evaluate whether data likely meet this assumption is to graphically display hazard functions characterized by different scores on each predictor.

For example, Figure 4.4 displays two hazard functions, dividing the sample using a median split on static risk (RoC*RoI) scores. Note in Figure 4.4 that (a) higher RoC*RoI scores are associated with a higher (logit) hazard rate, (b) the fitted lines and quadratic curves of both groups show a similar downward slope, and most crucially, (c) these fitted slopes remain essentially parallel to each other across follow-up weeks. Although the two slopes displayed in Figure 4.4 remain remarkably equidistant across time, some variation would not automatically indicate a violation of the proportional hazards assumption, whereas slopes that meet each other, or cross each other would certainly require re-thinking the use of a Cox regression model, or adapting the model to account for this violation.

Figure 4.4. Logit-transformed observed hazard rates at each week of the one-year follow-up period, separating the sample by a median split on RoC*RoI scores. Logit-transformed values presented as xs, with linear (*squares*) and quadratic (*circles*) regression lines of best fit superimposed onto the graph. $N = 3648$ ($N = 91,945$ person periods, $n = 1342$ recidivism events). RoC*RoI = Risk of Reconviction*Risk of Re-imprisonment (Bakker et al., 1999).

Although plotting hazard ratios gives an intuitive sense of a predictor's effects across time, plotting residuals is a more rigorous way to examine the proportionality assumption. When using categorical predictors or a Cox model with time-invariant predictors (i.e., a single measurement occasion design), it is appropriate to examine *log minus log plots*. By contrast, assumption checking for time-varying, continuous predictors is best examined through plots of Schoenfeld residuals graphed against time.

Figure 4.5 displays the Schoenfeld residuals for static risk (RoC*RoI) scores. Note in Figure 4.5 that the superimposed line of best fit (a) sits above 0, but (b) is essentially flat. This indicates that (a) at the time of recidivism, recidivists were likely to have higher RoC*RoI scores compared to others remaining in the sample (i.e., note that Schoenfeld residuals represent the observed minus the expected covariate values at each event time), and (b) this effect is consistent across time, suggesting the predictive effect of RoC*RoI does not change across the follow-up period. A statistical test confirmed that this slope does not differ from zero, $\rho = -0.01$, $\chi^2 = 0.25$, $p = 0.62$.

The proportionality assumption was assessed in the same ways with DRAOR subscale predictors. Figures 4.6, 4.7, and 4.8 each display graphs of two hazard functions, defined by selecting groups with dynamic DRAOR scores at approximately ± 1 SD from the mean of each subscale. Note in these graphs, that although the linear and quadratic lines of best fit are not strictly parallel, the general sense is that the size of the effect between the two scores is relatively invariant across all time periods. If these graphs suggest any violations of proportionality, they suggest that

Figure 4.5. Scatterplot of Schoenfeld residuals resulting from a Cox regression model with static risk (RoC*RoI) as the single predictor, with line of best fit superimposed onto graph. $N = 3648$ ($N = 91,945$ person periods, $n = 1342$ recidivism events). RoC*RoI = Risk of Reconviction*Risk of Re-imprisonment (Bakker et al., 1999).

Figure 4.6. Logit-transformed observed hazard rates at each week of the one-year follow-up period for scores that represent ± 1 SD on DRAOR *stable* subscale. Linear and quadratic regression lines of best fit are superimposed onto the graph for each subgroup. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Figure 4.7. Logit-transformed observed hazard rates at each week of the one-year follow-up period for scores that represent ± 1 SD on DRAOR *acute* subscale. Linear and quadratic regression lines of best fit are superimposed onto the graph for each subgroup. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Figure 4.8. Logit-transformed observed hazard rates at each week of the one-year follow-up period for scores that represent ± 1 *SD* on DRAOR *protect* subscale. Linear and quadratic regression lines of best fit are superimposed onto the graph for each subgroup. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

the predictive effects of DRAOR subscales increase at the end of the follow-up period. In contrast to RoC*RoI, which remains a steady predictor of recidivism whether participants have been in the community for a few months to several months, DRAOR subscales may become stronger predictors among participants who are able to survive crime-free in the community for longer periods of time.

More rigorous tests somewhat support this conclusion; however, if meaningfully present, this effect is non-linear, and small. First, Figure 4.9 displays the Schoenfeld residuals for the *stable* subscale. The relationship between *stable* scores and recidivism appears to be slightly curvilinear across time; this relationship is significantly different from zero, $\rho = -0.07$, $\chi^2 = 6.02$, $p = 0.01$, even after applying a Bonferroni correction for family-wise error. Yet, this correlation is small in magnitude. In order to account for non-proportional hazards, Cox models can be adapted to include interactions between predictors and time, as displayed in Table 4.10. As displayed in Column 2 of Table 4.10, the bolded value indicates that there is a statistically significant interaction between *stable* scores and quadratic time. Overall, these results suggest a convex relationship between risk scores and time, such that *stable* scores are at a peak in predictive ability at the beginning of the follow-up period, decline in predictive ability slightly over time, and then rise again by the end of the follow-up.

Conclusions were similar, but a more mixed, when examining proportionality with the *acute* and *protect* subscales. Figures 4.10 and 4.11 suggest slight departures from null-slope lines, with statistically significant tests for *acute*, $\rho = -0.06$, $\chi^2 = 4.36$, $p = 0.04$, and for *protect*, $\rho = 0.09$, $\chi^2 = 9.81$, $p = 0.002$. In Table 4.10,

Figure 4.9. Scatterplot of Schoenfeld residuals resulting from a Cox regression model with DRAOR *stable* subscale as the single predictor, with line of best fit superimposed onto graph. $N = 3648$ ($N = 91,945$ person periods, $n = 1342$ recidivism events). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Table 4.10

*Results for Fitting Three Discrete-time Hazard Models Predicting Criminal Recidivism Using DRAOR Subscales, Including Interaction Terms Representing Score*Time*

Parameters	Stable	Acute	Protect
	<i>B (SE)</i>	<i>B (SE)</i>	<i>B (SE)</i>
Constant	-5.55*** (0.17)	-5.61*** (0.16)	-2.64*** (0.13)
Week	0.06** (0.02)	0.05** (0.02)	-0.004 (0.02)
Week ²	-0.002*** (0.0004)	-0.001*** (0.0004)	-0.0005 (0.0005)
Time-varying Score	0.23*** (0.02)	0.26*** (0.02)	-0.24*** (0.02)
Score x Week	-0.007** (0.003)	-0.005* (0.003)	0.004 (0.003)
Score x Week ²	0.0001* (6.1 x 10⁻⁵)	0.0001* (6.5 x 10⁻⁵)	-0.00006 (6.5 x 10 ⁻⁵)
Deviance	13542.65	13476.06	13541.77
AIC	13554.65	13488.06	13553.77
BIC	13611.22	13544.63	13610.34
$\chi^2 \Delta$ from model with no interaction terms	7.06* (2)	3.92 (2)	5.94 (2)

Note. $N = 91,945$, n recidivism events = 1342. BIC was calculated using the person-period sample size ($N = 91,945$), rather than the person sample size ($N = 3648$). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

* $p < .05$, ** $p < .01$, *** $p < .001$; **Bolded values** indicate the most relevant statistically significant results.

Figure 4.10. Scatterplot of Schoenfeld residuals resulting from a Cox regression model with DRAOR *acute* subscale as the single predictor, with line of best fit superimposed onto graph. $N = 3648$ ($N = 91,945$ person periods, $n = 1342$ recidivism events). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Figure 4.11. Scatterplot of Schoenfeld residuals resulting from a Cox regression model with DRAOR *protect* subscale as the single predictor, with line of best fit superimposed onto graph. $N = 3648$ ($N = 91,945$ person periods, $n = 1342$ recidivism events). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Column 3 shows that the interaction terms were significant within the *acute* model, yet the test examining whether adding the interaction terms improved model fit was not statistically significant, $\chi^2(2) = 3.92, p = 0.14$. For *protect*, Column 4 displays statistically non-significant results for the interaction terms, and a statistically non-significant chi-square test.

Thus, overall, these results do not provide compelling reasons to suggest models using the DRAOR subscales substantially violate the assumption of proportional hazards across time. In particular, there were small correlations between Schoenfeld residuals and time, and, visually, the slopes appear largely flat in the graphs. It is unlikely these small correlations would have been statistically significant if the sample size were smaller, as high statistical power is able to detect legitimate, but relatively miniscule and inconsequential effects. In particular, little effort should be put into interpreting the curvilinear shape, since the upswing at the end of follow-up period is influenced by a relatively small number of participants remaining in the sample, and relatively few recidivists.

However, all models presented in this dissertation were compared to models that also included predictor-time interaction effects. Including these interaction terms had essentially no impact on results, and did not change the substantive conclusions.

Model Building

Prior to testing the core hypothesis of this dissertation in Chapter 5, this section presents single-predictor time-varying models. These basic models were used to (a) demonstrate that the predictor variables were significantly associated

with recidivism, when tested alone, (b) provide a base standard against which to compare the later, more complex models, and (c) permit initial checks of assumptions (including those described earlier in this chapter, and additional model checks in the next section of this chapter). Attending to indicators of effect size, these univariate models express the degree to which each time-varying predictor alone is able to successfully predict recidivism. Recall from Chapter 2 that the *c*-index is interpreted as an AUC (Area Under the Curve) statistic, and that Xu and O’Quigley’s (1999) R^2 statistic, referred to in tables as “ R^2 (XO)”, is interpreted as the percentage of total variance accounted for by the model.

First, Table 4.11 displays the results from a model with RoC*RoI as a single predictor. As expected, static risk scores were significantly related to recidivism. Since RoC*RoI’s scale ranges from 0 to 1, the hazard ratio expressing the effect for one unit difference is not highly intuitive. This can be re-expressed, such that compared to the average static score, scores one standard deviation higher were associated with a 2.4 times higher likelihood of recidivism, during any time period of the follow-up year ($OR = 2.36$). Indicators of effect size suggest that static risk scores alone explain 44% of the total outcome variance (the order and timing of recidivism events), and represent a strong predictor of recidivism, even after controlling for individual differences in follow-up time (*c*-index = 0.71).

As with all DRAOR subscales presented in this section, note that *stable* in Table 4.12 is a dynamic predictor, such that as re-assessments occurred across time, updated scores were entered into the model. Results suggest that for each one-unit score in *stable*, recidivism was 20% more likely in the subsequent week of the

Table 4.11
*Results of a Discrete-time Hazard Model Using Static Risk (RoC*RoI) to Predict Survival Time to Criminal Recidivism*

Parameter	<i>B (SE)</i>	Wald	<i>df</i>	<i>p</i>	<i>Exp(B)</i>	95% CI
Constant	-6.08 (0.11)	2883.45	1	< .001	0.002	[0.002, 0.003]
Week	0.01 (0.007)	1.78	1	0.18	1.01	[0.99, 1.02]
Week ²	-0.0005 (0.00002)	9.65	1	0.002	0.99	[0.99, 1.00]
RoC*RoI	3.53 (0.15)	588.38	1	< .001	34.10	[25.64, 45.35]
Deviance	13164.46					
AIC	13172.46					
BIC	13210.18				<i>R</i>² (XO)	0.44
# of iterations	8				<i>c</i>-index	0.71
$\chi^2 \Delta$ from model without predictor	725.54 (1)					

Note. *N* = 91,945, *n* recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size (*N* = 3648). RoC*RoI = Risk of Reconviction* Risk of Re-imprisonment (Bakker et al., 1999). AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

Table 4.12
Results of a Discrete-time Hazard Model Using DRAOR Stable Subscale to Predict Survival Time to Criminal Recidivism

Parameter	<i>B (SE)</i>	Wald	<i>df</i>	<i>p</i>	<i>Exp(B)</i>	95% CI
Constant	-5.21 (0.09)	3130.96	1	< .001	0.005	[0.005, 0.007]
Week	0.014 (0.0007)	3.55	1	0.06	1.01	[0.99, 1.03]
Week ²	-0.0006 (0.00002)	13.09	1	< .001	0.99	[0.99, 1.00]
Stable	0.18 (0.01)	341.74	1	< .001	1.20	[1.18, 1.22]
Deviance	13549.71					
AIC	13557.71					
BIC	13595.43				<i>R</i>² (XO)	0.17
# of iterations	8				<i>c</i>-index	0.65
$\chi^2 \Delta$ from model without predictor	340.29 (1)					

Note. *N* = 91,945, *n* recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size (*N* = 3648). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

follow-up (i.e., a DRAOR *stable* score of 3 is associated with a 20% higher likelihood of recidivism the following week, compared to a score of 2). To re-express the effect in terms of standard deviations, scores that were +1 *SD* above the mean were associated with 1.7 higher odds of recidivism the following week, compared with average *stable* scores in the same, or any other, time period.

The model averages this effect in such a way that this interpretation remains consistent within these data across all time periods of the follow-up (i.e., both early and late periods of the one-year follow-up). Also, the model averages the magnitude of this effect in such a way that this interpretation remains consistent whether or not participants' scores changed across time. In other words, using the example from above, a score of 3 was always associated with a 20% higher likelihood of recidivism compared to a score of 2 in these data, even among (a) participants who were rated with a *stable* score of 3 in every time period, and (b) participants who were rated with a *stable* score of 3 in some, but not all time periods. Cox regression with time-varying predictors does not frame prediction effects as the likelihood that an *individual* will experience recidivism, but the likelihood of recidivism associated with predictor *scores*, which will at some times characterize individual participants, and some times will not.

Finally, Table 4.12 shows that the *stable* subscale was a good predictor of recidivism (*c*-index = 0.65) in the univariate model, and this model explained about 17% of the total variance in outcome. This suggests a good level of predictive accuracy can be achieved with a measure that is both relatively brief, and includes only theoretically dynamic risk items.

Similar time-varying prediction models are displayed for *acute* and *protect* in Tables 4.13 and 4.14. Note that when comparing the three subscales, prediction accuracy rates remained similar ($c = 0.65$ to 0.66). However, when the *acute* subscale was the predictor, the resulting model explained about 7-9% more outcome variance, compared to the other two subscales. For each one-unit score in *acute*, recidivism was about 25% more likely in the subsequent week, and for each one-unit score in *protect*, there was about one-fifth reduced likelihood of recidivism in the subsequent week.

Prior research with this dataset (Hanby, 2013) demonstrated that *baseline* DRAOR subscales were statistically significant when added incrementally to prediction models with static risk (RoC*RoI) scores. At the end of Chapter 5, this dissertation displays similar multivariate tests that include all predictors in the same model, but unlike the analyses presented by Hanby (2013), these models employ the time-varying versions of the DRAOR subscales.

Final Model Checks

After building a Cox regression model, some additional checks are required to ensure that (a) the model is correctly specified, (b) all cases are within range, and (c) the model's limitations are well understood. Similar to the assumption checking described earlier in this chapter, exploring the following three issues helps to justify more complex modeling, and provides context that qualifies how model results should be interpreted.

Table 4.13
Results of a Discrete-time Hazard Model Using DRAOR Acute Subscale to Predict Survival Time to Criminal Recidivism

Parameter	B (SE)	Wald	df	p	Exp(B)	95% CI
Constant	-5.43 (0.10)	3215.91	1	<.001	0.004	[0.004, 0.005]
Week	0.02 (0.01)	9.64	1	0.002	1.02	[1.01, 1.04]
Week ²	-0.0008 (0.0002)	20.04	1	<.001	0.99	[0.99, 1.00]
Acute	0.23 (0.01)	438.09	1	<.001	1.26	[1.23, 1.28]
Deviance	13479.99					
AIC	13487.99				R² (XO)	0.24
BIC	13525.71				c-index	0.66
# of iterations	7					
χ² Δ from model with no predictor	410.01 (1)					

Note. N = 91,945, n recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size (N = 3648). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

Table 4.14
Results of a Discrete-time Hazard Model Using DRAOR Protect Subscale to Predict Survival Time to Criminal Recidivism

Parameter	<i>B</i> (SE)	Wald	<i>df</i>	<i>p</i>	Exp(<i>B</i>)	95% CI
Constant	-2.86 (0.08)	1329.21	1	<.001	0.06	[0.05, 0.07]
Week	0.02 (0.007)	6.71	1	0.01	1.02	[1.01, 1.03]
Week ²	-0.0007 (0.0002)	16.81	1	<.001	0.99	[0.99, 1.00]
Protect	-0.20 (0.01)	351.89	1	<.001	0.82	[0.80, 0.84]
Deviance	13547.71					
AIC	13555.71					
BIC	13593.43				R² (XO)	0.16
# of iterations	7				c-index	0.65
χ² Δ from model with no predictor	342.29 (1)					

Note. *N* = 91,945, *n* recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size (*N* = 3648). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

Potential Non-linear Effects

As discussed earlier in this chapter, the Box-Tidwell approach was used prior to model specification to test for the possibility of non-linear effects. This section further explores whether predictors might demonstrate non-linear relationships with hazard, such that a one-unit difference between scores that are placed low on a predictor's scale may have less of an incremental effect on hazard for recidivism, compared to a one-unit difference between scores that are placed higher on a predictor's scale. Carrying forward the example from above, it may be that a *stable* score of 3 is associated with an 18% higher likelihood of recidivism, compared to a score of 2, but that a score of 7 may have a 22% higher likelihood of recidivism, compared to a score of 6.

In Table 4.15, results from four independent models are displayed, which each include quadratic terms to examine the potential of non-linear relationships between predictors and (logit) hazard. Note that in the bottom row of the table, chi-square statistics indicate whether including the quadratic term adds significantly to the model.

Overall, RoC*RoI was the only predictor that showed evidence of a non-linear relationship with logit hazard. Visual inspection of graphed data (not shown) suggests that RoC*RoI scores between about 0 and .35 show a relatively linear incline in their relationship with risk of recidivism before starting to show some slight upward acceleration between scores of .35 to 1.00. Note that the quadratic term is significant (as shown in Table 4.15), but adding a cubic term (not shown) did not result in statistically significant improvement in the model. By contrast, non-

Table 4.15

Results of Fitting Four Discrete-time Hazard Models Predicting Criminal Recidivism Using DRAOR Subscales, Including Quadratic Predictor Terms

Parameters	RoC*RoI	Stable	Acute	Protect
	<i>B (SE)</i>	<i>B (SE)</i>	<i>B (SE)</i>	<i>B (SE)</i>
Constant	-6.49** (0.19)	-5.39** (0.14)	-5.48** (0.14)	-2.89** (0.11)
Week	0.009 (0.007)	0.01 (0.007)	0.02* (0.007)	0.02* (0.007)
Week ²	-0.0005** (0.0002)	-0.0006** (0.0002)	-0.0008** (0.0002)	-0.0007** (0.0002)
Linear term	5.36** (0.69)	0.25** (0.04)	0.26** (0.04)	-0.19** (0.04)
Quadratic term	-1.70* (0.62)	-0.01 (0.003)	-0.002 (0.003)	-0.001 (0.003)
Deviance	13156.12	13546.72	13479.53	13547.57
AIC	13166.12	13556.72	13489.53	13557.57
BIC	13213.27	13603.87	13536.68	13604.72
$\chi^2 \Delta$ from model with no quadratic term	7.91* (1)	2.99 (1)	0.46 (1)	0.14 (1)

Note. $N = 91,945$, n recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size ($N = 3648$). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). RoC*RoI = Risk of Recidivism * Risk of Re-imprisonment (Bakker et al., 1999). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

* $p < .01$, ** $p < .001$; **Bolded values** indicate the most relevant statistically significant results.

significant quadratic terms for the DRAOR subscales suggest there were linear relationships between each subscale and hazard. Note that models presented in this dissertation that included RoC*RoI as a predictor were compared to models that additionally included a quadratic term for RoC*RoI; substantive conclusions did not differ between these two approaches.

Presence of Highly Influential Cases

After a model has been specified, it can be evaluated to determine if any individual observations within the dataset exerted disproportionate influence on the model's estimated parameters. Highly influential cases may be considered outliers in the solution, and when these exist, it is important to follow up on these unusual cases to determine if they should be considered separately from the rest of the sample. Assessing the presence of highly influential cases is accomplished by examining the *DFbeta* statistic, which will exhibit a large value if deleting the associated observation would substantially change the parameter's value. Thus, *DFbetas* flag outlying cases in the solution by giving high values to observations that have had substantial impact on the coefficient values.

Some (Tabachnick & Fidell, 2007) suggest that *DFbetas* with values larger than 1 are problematic, whereas others (Ryan et al., 2015) suggest the decision threshold should be values that are larger than 2. Examining the *DFbeta* values (from the models presented in Model Building above) indicated that none of the values exceeded 1. This suggests there were no disproportionately influential cases within these model solutions.

Unobserved Heterogeneity

Both logistic regression models and Cox regression models make a critical assumption that can neither be logically checked, nor definitively tested. These models assume that observed outcomes (i.e., recidivism vs. no recidivism) are associated *only* with the predictors included in the model, and thus, all potential predictors have been observed and no additional predictors have been omitted from the model. Put in other words, two participants with exactly similar scores on all predictors should logically exhibit exactly similar outcomes, since the model assumes no other unmeasured predictor exists to explain why these participants would differ.

In practice, it is highly evident that this is a nonviable assumption, except within a successful randomized controlled trial where participant assignment fully neutralized all group differences. In this case, it would be reasonable to conceive of group assignment as that model's sole explanatory predictor. By contrast, Cox regression models predicting criminal recidivism seem particularly vulnerable to misspecification due to unobserved heterogeneity, given that theoretical models of criminal behavior (as described in Chapter 1) posit there are an array of environmental influences on criminal recidivism, which researchers rarely attempt to include in prediction models, and even if they were to do so, it would be unlikely they would be able to include enough predictors to account for the full range of outcome variance.

Thus, unobserved heterogeneity is a problematic assumption that should encourage the field to exercise caution when interpreting the results of what

amounts to a large percentage of research within correctional psychology. This is particularly true when attempting to use Cox regression models to either (a) make causal statements about the relationship between predictors and outcomes, or (b) make statements about the likelihood that any particular individual will experience an outcome, either those within an observed sample, or those similar to the participants who were studied.

In the absence of explicit knowledge of unobserved variables, there are two ways to approach estimating how much bias may exist due to unexplained heterogeneity. First, it is helpful to interpret model results in the context of a predictor's (or model's) effect size, specifically the percentage of outcome variance explained. For example, as reported earlier in this chapter, including RoC*RoI as a single predictor resulted in a model that explained 44% of the variance in outcome. This estimate suggests there is an additional 56% of variance remaining, which encourages caution when interpreting the degree of impact associated with RoC*RoI as a predictor of recidivism, given that unobserved predictors likely exist that explain this additional variance.

On the other hand, as described next, when the primary goal of model building is to understand the relationship between a predictor and outcome, without having designs to accurately (and presciently) predict 100% of the outcomes, it can be justifiable to accept that unmeasured predictors may exist, as long as one can hypothesize that these unmeasured predictors are unlikely to be highly correlated with measured predictors.

Specifically, a second way to approach the problem of unmeasured heterogeneity is to conduct sensitivity analyses, which estimate the probable biasing impact of unknown variables on model parameters. Lin, Psaty, and Kronmal (1998) developed sensitivity analyses for Cox regression, and their approach suggests it is most problematic when unmeasured variables both (a) have strong independent effects on outcomes, and (b) exhibit high correlations with predictors that are included in the model.

For example, in Table 4.16, sensitivity analyses are presented for the univariate static risk (RoC*RoI) model. Note that under the conditions that (a) no unmeasured variable is associated with the outcome, or (b) the unmeasured variable has no correlation with RoC*RoI, the hazard ratio for the effect of RoC*RoI is unchanged from the value displayed in Table 4.11. (Note that comparing the tables does reveal a slight difference in hazard ratios; this is only due to using a slightly different estimation approach in order to utilize the formulae presented by Lin et al., 1998). However, the remainder of the values in Table 4.16 suggest the effects attributed to RoC*RoI scores would be substantially reduced in the event that unmeasured predictors exist that have increasingly higher correlations with RoC*RoI (displayed across *rows*) and/or have increasingly higher associations with recidivism outcomes (displayed across *columns*).

Examining the *columns*, Table 4.16 estimates how sensitive the parameter estimate would be to the existence of unmeasured variables with no (0.0), small (0.6), medium (1.8), or large (3.0) independent effects on recidivism. For comparison purposes, each of the DRAOR subscales was observed to have a

Table 4.16
*Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and Static Risk (RoC*RoI) Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, as Estimated Through Sensitivity Analyses (Lin et al., 1998)*

Hypothesized Correlation Between Predictor & Unobserved Variable (ρ)	Hypothesized Strength of Relationship Between Unobserved Variable and Outcome (I')			
	0.0	0.6	1.8	3.0
.00	32.9 [24.8, 43.6]	32.9 [24.8, 43.6]	32.9 [24.8, 43.6]	32.9 [24.8, 43.6]
.20	32.9 [24.8, 43.6]	20.1 [15.2, 26.7]	7.52 [5.67, 9.97]	2.81 [2.12, 3.73]
.50	32.9 [24.8, 43.6]	9.6 [7.3, 12.8]	0.82 [0.62, 1.09]	0.07 [0.05, 0.09]
.75	32.9 [24.8, 43.6]	5.2 [3.9, 6.9]	0.13 [0.09, 0.17]	0.003 [0.002, 0.004]
<i>Note.</i> RoC*RoI = Risk of Reconviction*Risk of Re-imprisonment (Bakker et al., 1999).				

relationship with recidivism in the *small* range, whereas RoC*RoI was observed to have a *large* relationship.

Examining the *rows*, Table 4.16 estimates sensitivity under the conditions that an unmeasured variable might have no correlation (0.00), a small correlation (0.20), a medium correlation (0.50), or a large correlation (0.75) with RoC*RoI. For comparison purposes, each DRAOR subscale and RoC*RoI were observed to have a correlation in the *small* range. Using these comparisons, sensitivity analyses in Table 4.16 estimate that the observed hazard ratio of 32.9 for RoC*RoI would in actuality be reduced to 20.1 if a DRAOR subscale (which has a small relationship with outcome, and a small correlation with RoC*RoI) were included in the model. Using actual data to verify this estimate, an analysis that includes both RoC*RoI and DRAOR *stable* scores in a prediction model reduces the hazard ratio associated with RoC*RoI to a value that is similar to this estimate (specifically, $OR = 23.2$).

As stated above, unless a researcher withholds some measured predictors from a model, it is logically impossible to know the true impact of truly unmeasured variables. Instead, the purpose of sensitivity analyses is to specify the conditions under which an observed effect is likely to be reduced to statistical non-significance. This knowledge encourages exercising caution during interpretation of results, and stimulates researchers to consider what unmeasured variables should be prioritized for future measurement in that field of study. As argued above, sensitivity analyses should also be considered alongside estimates of R^2 effect size, which help determine if a strong case can be made that many unmeasured variables likely exist.

As implied by this discussion, it is most logical and important to conduct sensitivity analyses for final models (that include all measured predictors). But, prior to presenting sensitivity analyses for the full model specified in Chapter 5, for the sake of completeness, Tables 4.17, 4.18, and 4.19 display sensitivity analyses for the univariate predictor models that tested prediction among the *stable*, *acute*, and *protect* DRAOR subscales (from the section titled Model Building above). Similar to Table 4.16, note that these tables suggest that effects would remain statistically significant (albeit somewhat reduced) as long as a correlated, unmeasured variable was characterized only by a small effect on recidivism outcome. Unmeasured variables with high correlations with DRAOR *and* large relationships with recidivism would substantially change the conclusions derived from the prediction models displayed in this chapter. Ultimately, only further research (i.e., examining whether DRAOR subscales correlate with other measurable life circumstances or environmental influences) can fully examine whether present conclusions are tenable.

Table 4.17

Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and DRAOR Stable Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, as Estimated Through Sensitivity Analyses (Lin et al., 1998)

Hypothesized Correlation Between Predictor & Unobserved Variable (ρ)	Hypothesized Strength of Relationship Between Unobserved Variable and Outcome (I')			
	0.0	0.6	1.8	3.0
.00	1.20 [1.17, 1.22]	1.20 [1.17, 1.22]	1.20 [1.17, 1.22]	1.20 [1.17, 1.22]
.20	1.20 [1.17, 1.22]	1.15 [1.12, 1.17]	1.05 [1.03, 1.07]	0.97 [0.95, 0.99]
.50	1.20 [1.17, 1.22]	1.08 [1.06, 1.10]	0.87 [0.85, 0.89]	0.71 [0.69, 0.72]
.75	1.20 [1.17, 1.22]	1.02 [1.00, 1.04]	0.74 [0.73, 0.76]	0.54 [0.53, 0.55]
<i>Note.</i> DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).				

Table 4.18

Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and DRAOR Acute Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, as Estimated Through Sensitivity Analyses (Lin et al., 1998)

Hypothesized Correlation Between Predictor & Unobserved Variable (ρ)	Hypothesized Strength of Relationship Between Unobserved Variable and Outcome (I')			
	0.0	0.6	1.8	3.0
.00	1.25 [1.23, 1.28]	1.25 [1.23, 1.28]	1.25 [1.23, 1.28]	1.25 [1.23, 1.28]
.20	1.25 [1.23, 1.28]	1.19 [1.17, 1.22]	1.08 [1.05, 1.10]	0.97 [0.95, 0.99]
.50	1.25 [1.23, 1.28]	1.10 [1.08, 1.13]	0.86 [0.84, 0.87]	0.66 [0.65, 0.68]
.75	1.25 [1.23, 1.28]	1.04 [1.01, 1.06]	0.71 [0.69, 0.72]	0.48 [0.47, 0.49]
<i>Note.</i> DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).				

Table 4.19

Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and DRAOR Protect Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, as Estimated Through Sensitivity Analyses (Lin et al., 1998)

Hypothesized Correlation Between Predictor & Unobserved Variable (ρ)	Hypothesized Strength of Relationship Between Unobserved Variable and Outcome (I')			
	0.0	0.6	1.8	3.0
.00	0.82 [0.80, 0.84]	0.82 [0.80, 0.84]	0.82 [0.80, 0.84]	0.82 [0.80, 0.84]
.20	0.82 [0.80, 0.84]	0.84 [0.83, 0.87]	0.93 [0.91, 0.94]	0.99 [0.98, 1.00]
.50	0.82 [0.80, 0.84]	0.91 [0.88, 0.93]	1.06 [1.05, 1.07]	1.18 [1.17, 1.19]
.75	0.82 [0.80, 0.84]	0.95 [0.94, 0.96]	1.15 [1.14, 1.16]	1.30 [1.30, 1.31]

Note. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Chapter 5: Testing Prediction with Proximal vs. Distal Assessments

The purpose of this chapter is to propose a framework for testing the hypothesis in Chapter 1 of this dissertation, and present results of these tests using the dataset described in Chapter 3. Specifically, the question of interest focuses on whether there is predictive value in measuring within-individual re-assessments on a risk measure that includes theoretically dynamic items. As discussed in Chapter 1, there may be two mechanisms that allow re-assessments to enhance outcome prediction, either because (a) more proximal assessments are better indicators of current risk state, or (b) larger combinations of re-assessments provide better point estimates of overall risk status.

This chapter describes a three-step process to test whether more proximal assessments result in better prediction of short-term outcomes. After examining the proximity hypothesis, the final, fourth section of this chapter presents prediction models that combine all predictors. These multivariate models establish whether it is useful in practice to continually re-assess all DRAOR subscales alongside a strongly predictive static risk (RoC*RoI) measure.

Baseline versus Proximal Assessments

In practice, it would not be cost effective to continually re-assess risk variables among parolees in the community if initial, one-time assessments were able to predict as accurately (or more accurately) than subsequent assessments. Thus, the first logical step for exploring whether more proximal assessments are beneficial for prediction requires establishing that re-assessments incrementally add to prediction models beyond baseline assessments. Figure 5.1 displays the logic

Figure 5.1. A single participant's observed values on the DRAOR *stable* subscale across time. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

of comparing these two approaches. Specifically, Figure 5.1 graphs the pattern of a typical sequence in the dataset (with data points designated by “x”s). This participant’s sequence was characterized by a six-month follow-up (at which point censoring occurred). DRAOR scores initially began high and then declined during the first three months of the follow-up period. In the case of this participant, scores declined by two subscale points. In the prediction models presented in this section, participants’ observed values were compared against an assessment strategy where baseline scores were continually recycled into the prediction model across all time periods (also displayed in Figure 5.1, designated by squares).

To compare these strategies, Table 5.1 displays the results of three prediction models: (a) baseline DRAOR *stable* scores only, (b) re-assessed, dynamic DRAOR *stable* scores only, and (c) both baseline and re-assessed variables in a combined model. Results displayed in Table 5.1 support the conclusion that baseline scores were slightly weaker predictors compared to more proximal scores. First, note that the dynamic, re-assessment model (Model B in Table 5.1) resulted in better model fit (smaller AIC and BIC values compared to Model A), and slightly enhanced effect sizes (explaining 5% more variance in outcome, and a minor increase on the *c*-index value, compared to Model A). Second, the dynamic variable remained statistically significant in Model C when it was added incrementally above baseline scores.

Past research (Brown et al, 2009; Howard & Dixon, 2013; Jones et al., 2010; see discussion in Chapter 1) suggests Cox regression models with statistically significant incremental time-dependent variables may not simultaneously

Table 5.1

Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Stable Subscale Score

Parameters	Model A		Model B		Model C	
	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]
Constant	-5.18** (0.10)	0.006	-5.21** (0.09)	0.005	-5.36** (0.10)	0.005
Week	0.002 (0.007)	1.00 [0.99, 1.02]	0.014 (0.007)	1.00 [0.99, 1.02]	0.01 (0.007)	1.01 [0.99, 1.03]
Week ²	-0.0005* (0.0002)	0.99 [0.99, 1.00]	-0.0006** (0.0002)	0.99 [0.99, 1.00]	-0.0006** (0.0002)	0.99 [0.99, 1.00]
Stable (Baseline)	0.18** (0.01)	1.19 [1.17, 1.22]	-	-	0.06** (0.02)	1.06 [1.03, 1.10]
Stable (Dynamic)	-	-	0.18** (0.01)	1.20 [1.18, 1.22]	0.14** (0.02)	1.15 [1.12, 1.18]
R² (X0)	0.12		0.17		0.16	
c-index	0.64		0.65		0.64	
Deviance	13628.72		13549.71		13535.50	
AIC	13636.72		13557.71		13545.50	
BIC	13685.87		13606.86		13592.65	

$\chi^2 \Delta$ Model C vs. Model A	93.21** (1)
$\chi^2 \Delta$ Model C vs. Model B	14.20** (1)

Note. $N = 91,945$, n recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size ($N = 3648$). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). CI = confidence interval. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

* $p < .01$, ** $p < .001$; **Bolded values** indicate the most relevant statistically significant results.

demonstrate greater prediction efficacy (as measured by higher AUC values), compared to static-only models. To test this, time-dependent AUC analyses (described in Chapter 2) were calculated. Figure 5.2 displays prediction accuracies that correspond to the models in Table 5.1 (Models A and B).

In Figure 5.2, note that the predictive accuracy associated with baseline scores (represented with a dashed line) began within the $AUC = 0.65$ to 0.70 range, declined over the first 15 weeks of follow-up, and then remained steady at approximately $AUC = 0.60$. By contrast, although the initial pattern was similar when using re-assessed scores (represented with a solid line), prediction accuracy “recovered” slightly more from the early time-period decline, and leveled at approximately $AUC = 0.63$ to 0.65 .

However, differences in predictive accuracy across time were small. Confirming this, 95% confidence intervals were generated using bootstrap methods; these confidence intervals were overlaid onto the graphical display, as shown in Figure 5.3. Note that the high degree of overlap in the two sets of confidence intervals demonstrates that the AUC values across time did not differ at a statistically significant level, when comparing the baseline approach to the re-assessment approach. As shown in Figure 5.3, the only exception to this occurred at approximately Weeks 21 and 22, where the two sets of confidence intervals briefly did not overlap. In terms of overall prediction, this short period does not represent a highly meaningful effect, but may be interpreted as promising, given that the effect is in the direction suggested by the proximity hypothesis.

Figure 5.2. Time dependent Area Under the Curve (AUC) statistics graphed across the one-year follow-up period for baseline DRAOR *stable* scores and dynamic DRAOR *stable* scores. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Figure 5.3. Bootstrapped 95% confidence intervals overlaid onto Figure 5.2.

These same strategies were used to examine the DRAOR *acute* and *protect* subscales. As with *stable*, for both subscales, re-assessment scores showed statistically significant incremental prediction over baseline scores (see Tables 5.2 and 5.3). Specifically for *acute*, compared to the baseline model, the dynamic model explained an additional 9% variance in outcome. For *protect*, there was an additional 3% variance explained.

For *acute* and *protect*, respectively, time-dependent AUC analyses are displayed in Figures 5.4 (with 95% confidence intervals overlaid in Figure 5.5) and Figure 5.6 (with 95% confidence intervals overlaid in Figure 5.7). Comparing the three subscales, note that prediction differences between models were most pronounced for the *acute* subscale. As displayed in Figure 5.4, by about Week 10, the dynamic model maintained a steady prediction rate at approximately AUC = 0.65, whereas the baseline model showed consistent decline in predictive accuracy. Notably, the 95% confidence intervals for the two models did not overlap across two periods of the follow-up (as displayed in Figure 5.5), specifically between approximately Weeks 15-25 and Weeks 35-45.

For *protect*, the overall pattern suggested a general decline in predictive accuracy across time (for both baseline and dynamic models), and differences between the two models were small across time. Despite this, one period characterized by non-overlapping 95% confidence intervals was observed during Weeks 25-35 of the follow-up (as displayed in Figure 5.7).

Table 5.2

Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Acute Subscale Score

Parameters	Model A		Model B		Model C	
	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]
Constant	-5.11** (0.10)	0.006	-5.40** (0.10)	0.004	-5.62** (0.11)	0.004
Week	0.0017 (0.007)	1.00 [0.99, 1.02]	0.02* (0.007)	1.02 [1.01, 1.04]	0.02* (0.007)	1.02 [1.01, 1.04]
Week ²	-0.0005* (0.0002)	0.99 [0.99, 1.00]	-0.0008** (0.0002)	0.99 [0.99, 1.00]	-0.0008** (0.0002)	0.99 [0.99, 1.00]
Acute (Baseline)	0.18** (0.01)	1.19 [1.17, 1.22]	-	-	0.07** (0.01)	1.07 [1.04, 1.10]
Acute (Dynamic)	-	-	0.23** (0.01)	1.26 [1.23, 1.28]	0.19** (0.01)	1.21 [1.18, 1.24]
R² (X0)	0.15		0.24		0.23	
c-index	0.63		0.66		0.65	
Deviance	13649.56		13479.99		13458.82	
AIC	13657.56		13487.99		13468.82	

BIC	13706.71	13537.14	13515.97
$\chi^2 \Delta$ Model C vs. Model A			190.73** (1)
$\chi^2 \Delta$ Model C vs. Model B			21.16** (1)

Note. $N = 91,945$, n recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size ($N = 3648$). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). CI = confidence interval. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

* $p < .01$, ** $p < .001$; **Bolded values** indicate the most relevant statistically significant results.

Table 5.3

Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Protect Subscale Score

Parameters	Model A		Model B		Model C	
	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]
Constant	-2.99** (0.08)	0.05	-2.86** (0.08)	0.06	-2.77** (0.08)	0.06
Week	0.003 (0.007)	1.00 [0.99, 1.02]	0.02* (0.007)	1.02 [1.01, 1.03]	0.02 (0.007)	1.02 [1.00, 1.03]
Week ²	-0.0005* (0.0002)	0.99 [0.99, 1.00]	-0.0007** (0.0002)	0.99 [0.99, 1.00]	-0.0007** (0.0002)	0.99 [0.99, 1.00]
Protect (Baseline)	-0.18** (0.01)	0.84 [0.82, 0.86]	-	-	-0.06** (0.02)	0.94 [0.92, 0.97]
Protect (Dynamic)	-	-	-0.20** (0.01)	0.82 [0.80, 0.84]	-0.16** (0.02)	0.85 [0.83, 0.88]
R² (X0)	0.13		0.16		0.17	
c-index	0.63		0.65		0.64	
Deviance (χ^2)	13652.63		13547.71		13534.84	
AIC	13660.63		13555.71		13544.84	

BIC	13709.78	13604.86	13591.99
$\chi^2 \Delta$ Model C vs. Model A			117.79** (1)
$\chi^2 \Delta$ Model C vs. Model B			12.87** (1)

Note. $N = 91,945$, n recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size ($N = 3648$). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). CI = confidence interval. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

* $p < .01$, ** $p < .001$; **Bolded values** indicate the most relevant statistically significant results.

Figure 5.4. Time dependent Area Under the Curve (AUC) statistics graphed across the one-year follow-up period for baseline DRAOR *acute* scores and dynamic DRAOR *acute* scores. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Figure 5.5. Bootstrapped 95% confidence intervals overlaid onto Figure 5.4.

Figure 5.6. Time dependent Area Under the Curve (AUC) statistics graphed across the one-year follow-up period for baseline DRAOR *protect* scores and dynamic DRAOR *protect* scores. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Figure 5.7. Bootstrapped 95% confidence intervals overlaid onto Figure 5.6.

Averaging versus Proximal Assessments

The tests in this section focus on whether re-assessments are better predictors of outcome *due to proximity*, or if re-assessments are best combined so that all existing information is averaged into a single point estimate, updated across time. Specifically, the analyses presented in this section compare a variable that represents (a) a rolling mean across time, to a variable that represents (b) the most proximal assessment. The logic of these comparisons is displayed in Figure 5.8. The *rolling mean* is an average of all assessments up to that point in time, starting the week following baseline. For the sequence displayed in Figure 5.8, which is generally typical in the dataset, note that as assessments were continually combined, the rolling mean declined smoothly across time until it terminated at the sequence's grand mean.

To compare these averaging and proximity strategies, Table 5.4 displays results from three prediction models: (a) rolling mean DRAOR *stable* only, (b) dynamic DRAOR *stable* scores only, and (c) both rolling mean and proximal variables in a combined model. Results displayed in Table 5.4 support the conclusion that a rolling mean was a weaker predictor compared to more proximal scores. The dynamic model (Model B) resulted in better model fit compared to the rolling mean model (Model A). In addition, the rolling mean variable did not reach statistical significance in the combined model (Model C), whereas the proximity variable remained statistically significant. Indicators of effect size suggested that 3% additional variance was explained when using proximal assessments rather than the rolling mean strategy.

Figure 5.8. A single participant's observed values on the DRAOR *stable* subscale across time, with a depiction of a rolling mean score overlaid onto the graph. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Table 5.4

Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Stable Subscale Score, Including a Rolling Mean

Parameters	Model A		Model B		Model C	
	<i>B</i> (<i>SE</i>)	<i>OR</i> [95% CI]	<i>B</i> (<i>SE</i>)	<i>OR</i> [95% CI]	<i>B</i> (<i>SE</i>)	<i>OR</i> [95% CI]
Constant	-5.25** (0.10)	0.005	-5.21** (0.09)	0.005	-5.23** (0.10)	0.005
Week	0.01 (0.007)	1.01 [0.99, 1.03]	0.014 (0.007)	1.00 [0.99, 1.02]	0.01 (0.007)	1.01 [0.99, 1.03]
Week ²	-0.0006* (0.0002)	0.99 [0.99, 1.00]	-0.0006** (0.0002)	0.99 [0.99, 1.00]	-0.0006** (0.0002)	0.99 [0.99, 1.00]
Stable (Rolling)	0.19** (0.01)	1.20 [1.18, 1.23]	-	-	0.02 (0.03)	1.02 [0.97, 1.08]
Stable (Dynamic)	-	-	0.18** (0.01)	1.20 [1.18, 1.22]	0.16** (0.03)	1.17 [1.11, 1.24]
<i>R</i>² (X0)	0.14		0.17		0.17	
<i>c</i>-index	0.65		0.65		0.64	
Deviance	13606.75		13549.71		13549.05	
AIC	13614.75		13557.71		13559.05	
BIC	13663.90		13606.86		13606.20	

$\chi^2 \Delta$ Model C vs. Model A	35.46** (1)
$\chi^2 \Delta$ Model C vs. Model B	0.653 (1)

Note. $N = 91,945$, n recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size ($N = 3648$). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). CI = confidence interval. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

* $p < .01$, ** $p < .001$; **Bolded values** indicate the most relevant statistically significant results.

Results were similar, but slightly less pronounced for the *acute* (see Table 5.5) and the *protect* (see Table 5.6) subscales. For these subscales, in the combined models (Model C), the proximity variables incrementally predicted beyond the rolling mean variables, but the rolling mean variables also remained statistically significant. Comparing Model B to Model A, effect size indicators suggested a 2% difference in amount of variance explained for *acute*, but no difference in effect sizes for *protect*. However, despite attenuated effects, the core findings from these analyses were the same as for *stable*, which suggested that proximal assessments added to the prediction models beyond the rolling mean.

Different Averaging Strategies versus Proximity Only

The final step in this three-step framework tested the hypothesis that *both* averaging and proximity may be important influences on a dynamic measure's predictive ability, but that these two influences offset each other to a degree. In other words, proximal assessments may be good predictors of short-term recidivism *as long as they combine enough information to be accurate point estimates*, and averaged assessments may be good predictors of short-term recidivism *as long as they are proximal enough to the outcome*. If these two features are simultaneously important, (a) the model with only the most proximal assessment should demonstrate poorer fit compared to models that combine *recent* assessments into an average, *and* (b) models that combine *too many* assessments into averages should demonstrate poorer fit compared to models that combine only a small number of recent assessments into an average. In other words, averages of increasingly larger numbers of assessments should show a pattern of increasing

Table 5.5

Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Acute Subscale Score, Including a Rolling Mean

Parameters	Model A		Model B		Model C	
	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]
Constant	-5.61** (0.11)	0.004	-5.40** (0.10)	0.004	-5.61** (0.11)	0.004
Week	0.02* (0.007)	1.02 [1.01, 1.04]	0.02* (0.007)	1.02 [1.01, 1.04]	0.03* (0.007)	1.03 [1.01, 1.04]
Week ²	-0.0008** (0.0002)	0.99 [0.99, 1.00]	-0.0008** (0.0002)	0.99 [0.99, 1.00]	-0.0008** (0.0002)	0.99 [0.99, 1.00]
Acute (Rolling)	0.25** (0.01)	1.29 [1.26, 1.32]	-	-	0.11** (0.03)	1.11 [1.06, 1.17]
Acute (Dynamic)	-	-	0.23** (0.01)	1.26 [1.23, 1.28]	0.15** (0.02)	1.16 [1.11, 1.21]
R² (X0)	0.22		0.24		0.24	
c-index	0.66		0.66		0.65	
Deviance	13509.00		13479.99		13461.65	
AIC	13517.00		13487.99		13471.65	
BIC	13566.15		13537.14		13518.80	

$\chi^2 \Delta$ Model C vs. Model A	47.35** (1)
$\chi^2 \Delta$ Model C vs. Model B	18.34** (1)

Note. $N = 91,945$, n recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size ($N = 3648$). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). CI = confidence interval. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

* $p < .01$, ** $p < .001$; **Bolded values** indicate the most relevant statistically significant results.

Table 5.6

Results of Fitting Three Discrete-time Hazard Models to Predict Time to Criminal Recidivism, Using Different Specifications of DRAOR Protect Subscale Score, Including a Rolling Mean

Parameters	Model A		Model B		Model C	
	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]
Constant	-2.80** (0.08)	0.06	-2.86** (0.08)	0.06	-2.79** (0.08)	0.06
Week	0.02 (0.007)	1.02 [1.00, 1.03]	0.02* (0.007)	1.02 [1.01, 1.03]	0.02 (0.007)	1.02 [1.00, 1.03]
Week ²	-0.0007** (0.0002)	0.99 [0.99, 1.00]	-0.0007** (0.0002)	0.99 [0.99, 1.00]	-0.0007* (0.0002)	0.99 [0.99, 1.00]
Protect (Rolling)	-0.22** (0.01)	0.81 [0.79, 0.83]	-	-	-0.09* (0.03)	0.91 [0.86, 0.97]
Protect (Dynamic)	-	-	-0.20** (0.01)	0.82 [0.80, 0.84]	-0.12** (0.03)	0.88 [0.84, 0.93]
R² (X0)	0.16		0.16		0.17	
c-index	0.65		0.65		0.63	
Deviance	13558.66		13547.71		13538.04	
AIC	13566.66		13555.71		13548.04	
BIC	13615.81		13604.86		13595.19	

$\chi^2 \Delta$ Model C vs. Model A	20.62** (1)
$\chi^2 \Delta$ Model C vs. Model B	9.67* (1)

Note. $N = 91,945$, n recidivism events = 1342. BIC was calculated using the person-period sample size, rather than the person sample size ($N = 3648$). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). CI = confidence interval. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

* $p < .01$, ** $p < .001$; **Bolded values** indicate the most relevant statistically significant results.

model fit, until reaching an apex before reversing to show degraded model fit, once more distal assessments start to be averaged into the point estimate.

Figure 5.9 displays the logic of these model comparisons, by graphing a variable that represents a *moving* rolling mean. In Figure 5.9, each average (displayed as diamonds) combines the most recent three assessments observed at that point in time. The rolling mean “moves” because the average was continuously updated across time. Figure 5.9 displays a three-assessment averaging strategy, but to test the competing influences of averaging and proximity, seven different averaging strategies (from two assessments to eight) were compared against the non-averaged, proximity-only variable. These strategies are listed in Table 5.7.

To assess model fit, AIC values were compared; this approach requires that all models include exactly the same data points, differing only in how the variables were calculated. For this reason, a reduced version of the full dataset was employed to compare the eight strategies displayed in Table 5.7. Specifically, sequences could only be included in these models if participants were observed for at least eight weeks or more. After this decision rule, the dataset was reduced to 2871 sequences, representing 88,407 person-period weeks, and 858 recidivism events. When different decision rules were tested (i.e., including sequences lasting at least four weeks or more), the conclusions were identical to those displayed in Table 5.7.

In Table 5.7, for each DRAOR subscale, bolded values signify the model that resulted in the best fit. As displayed in Table 5.7, the model with the most proximal assessment fit best when predicting recidivism outcome, for both *stable* and *acute*.

Figure 5.9. A single participant's observed values on the DRAOR *stable* subscale across time, with a depiction of a moving rolling mean score overlaid onto the graph. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Table 5.7

Model Fit (Akaike Information Criteria) and Effect Size (R-Squared, and Concordance) Indices for Cox Regression Models with Time-varying Predictors Using Different Specifications of DRAOR Subscales as Predictors, Including Differing Averaging Strategies

Number of Averaged Assessment Weeks	Stable			Acute			Protect		
	AIC	R ² (XO)	c- index	AIC	R ² (XO)	c- index	AIC	R ² (XO)	c- index
1 (no averaging)	9233.61	0.132	0.628	9192.53	0.195	0.633	9241.79	0.113	0.625
2	9236.23	0.132	0.624	9199.72	0.189	0.633	9240.48	0.117	0.626
3	9240.09	0.126	0.623	9206.14	0.185	0.632	9241.14	0.116	0.627
4	9243.56	0.122	0.623	9208.87	0.184	0.633	9244.79	0.114	0.626
5	9246.79	0.119	0.622	9210.43	0.184	0.633	9248.10	0.113	0.625
6	9248.98	0.118	0.621	9214.37	0.185	0.632	9248.72	0.113	0.625
7	9251.41	0.116	0.620	9216.40	0.184	0.631	9248.25	0.114	0.625
8	9251.51	0.116	0.619	9216.19	0.185	0.630	9245.33	0.117	0.626
Δ between models	17.9	1.6%	0.01	23.87	1.1%	-	6.9	0.4%	-

Note. *n* = 2871 sequences (*n* = 88,407 person periods, 858 recidivism events). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale. AIC = Akaike Information Criteria. **Bolded values** indicate the model with best fit, or largest effect size.

For *protect*, the model with best fit still generally favored proximity over averaging, but suggested that averaging the most recent two assessments resulted in the best model. Indicators of effect size are also displayed in Table 5.7; differences between models are negligible, but the pattern of these values supports the same conclusions that were drawn by comparing AIC values.

Multivariate Prediction Models

Thus far, the models tested in this dissertation have entered DRAOR subscales as univariate predictors. To evaluate the practical utility of including all predictors in a prediction model, DRAOR subscales were entered together alongside a measure of static risk (RoC*RoI). Table 5.8 displays the results of three prediction models: (a) static risk only, (b) static risk and DRAOR baseline scores, and (c) static risk and DRAOR re-assessment scores.

Results (Models B and C) demonstrated that DRAOR *stable* did not reach statistical significance when entered alongside all other predictors. Comparing model fit and effect sizes, the dynamic model was the best prediction model, with Model C explaining an additional 13% of variance in outcome, compared to the static only model, and explained 57% of the total outcome variance. Prediction accuracy (*c*-index = 0.71 versus 0.74) was also slightly enhanced.

The time-dependent AUC values associated with Models A and B of Table 5.8 are displayed in Figure 5.10. As seen in the figure, combining DRAOR baseline scores with static scores slightly improved predictive accuracy early in the follow-up period, but by Week 15, there were no absolute differences in AUC values across time. By contrast, the model that combined static and dynamic scores maintained

Table 5.8

*Results of Fitting Three Multivariate Discrete-time Hazard Models to Predict Time to Criminal Recidivism, by Entering DRAOR Baseline or Dynamic Subscale Scores Following Static Risk Scores (RoC*RoI)*

Parameters	Model A		Model B		Model C	
	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]	<i>B (SE)</i>	<i>OR</i> [95% CI]
Constant	-6.08** (0.11)	0.002	-6.09** (0.19)	0.002	-6.31** (0.20)	0.002
Week	0.01 (0.007)	1.01 [0.99, 1.02]	0.01 (0.007)	1.01 [0.99, 1.03]	0.03** (0.007)	1.03 [1.02, 1.05]
Week ²	-0.0005* (0.0002)	0.99 [0.99, 1.00]	-0.0005* (0.0002)	0.99 [0.99, 1.00]	-0.0007** (0.0002)	0.99 [0.99, 1.00]
RoC*RoI	3.53** (0.15)	34.1 [25.6, 45.4]	3.13** (0.15)	22.8 [17.0, 30.5]	3.03** (0.15)	20.7 [15.4, 27.8]
Stable (Baseline)	-	-	0.03 (0.02)	1.03 [0.99, 1.06]	-	-
Acute (Baseline)	-	-	0.08** (0.01)	1.08 [1.05, 1.11]	-	-
Protect (Baseline)	-	-	-0.08** (0.01)	0.93 [0.90, 0.95]	-	-
Stable (Dynamic)	-	-	-	-	0.02 (0.01)	1.02 [0.99, 1.05]
Acute (Dynamic)	-	-	-	-	0.13** (0.01)	1.14 [1.11, 1.17]

Protect (Dynamic)	-	-	-	-0.08** (0.01)	0.92 [0.90, 0.95]
R² (XO)	0.44	0.50		0.57	
c-index	0.71	0.72		0.74	
Deviance	13165.46	13005.62		12869.61	
AIC	13172.46	13019.62		12883.61	
BIC	13211.18	13085.62		12949.61	
χ² Δ Model B vs. Model A				158.85** (3)	
χ² Δ Model C vs. Model A				294.86** (3)	

Note. $n = 3648$ sequences ($n = 91,945$ person periods, 1342 recidivism events). BIC was calculated using the person-period sample size, rather than the person sample size ($N = 3648$). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007). RoC*RoI = Risk of Reconviction*Risk of Re-imprisonment (Bakker et al., 1999). Stable = DRAOR *stable* subscale. Acute = DRAOR *acute* subscale. Protect = DRAOR *protect* subscale. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria.

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

Figure 5.10. Time dependent Area Under the Curve (AUC) statistics graphed across the one-year follow-up period for static risk scores (RoC*RoI) and baseline DRAOR subscale scores, combined. RoC*RoI = Risk of Reconviction*Risk of Re-imprisonment (Bakker et al., 1999). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

higher prediction accuracies across time compared to a static-only model, as displayed in Figure 5.11 (which graphs the time-dependent AUC values associated with Models A and C of Table 5.8). Bootstrapped 95% confidence intervals were overlaid onto this graph, as shown in Figure 5.12, and despite findings showing a consistent trend supporting the conclusion that re-assessment enhanced prediction, there were no statistically significant differences in prediction across time, when comparing the static-only (RoC*RoI) to the static + DRAOR dynamic models.

Finally, sensitivity analyses were conducted for the statistically significant predictors in Model C. Results are displayed in Tables 5.9, 5.10, and 5.11. Table 5.9 suggests that the effect associated with RoC*RoI would remain statistically significant under the conditions that additional unmeasured variables are characterized by (a) a small relationship with recidivism, *or* (b) a small correlation with RoC*RoI. By contrast, the effects associated with DRAOR *acute* and *protect* would remain statistically significant only under the condition that additional unmeasured variables are characterized by (a) a small relationship with recidivism, *and* (b) a small correlation with these DRAOR subscales.

Figure 5.11. Time dependent Area Under the Curve (AUC) statistics graphed across the one-year follow-up period for static risk scores (RoC*RoI) and dynamic DRAOR subscale scores, combined. RoC*RoI = Risk of Reconviction*Risk of Re-imprisonment (Bakker et al., 1999). DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Figure 5.12. Bootstrapped 95% confidence intervals overlaid onto Figure 5.11.

Table 5.9
*Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and Static Risk (RoC*RoI) Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, Estimated from the Final Multivariate Model*

Hypothesized Correlation Between Predictor & Unobserved Variable (ρ)	Hypothesized Strength of Relationship Between Unobserved Variable and Outcome (I')			
	0.0	0.6	1.8	3.0
.00	20.0 [15.0, 26.7]	20.0 [15.0, 26.7]	20.0 [15.0, 26.7]	20.0 [15.0, 26.7]
.20	20.0 [15.0, 26.7]	12.2 [9.14, 16.3]	4.57 [3.42, 6.10]	1.71 [1.28, 2.28]
.50	20.0 [15.0, 26.7]	5.84 [4.37, 7.80]	0.50 [0.37, 0.67]	0.04 [0.03, 0.06]
.75	20.0 [15.0, 26.7]	3.16 [2.36, 4.22]	0.08 [0.06, 0.11]	0.002 [0.001, 0.003]

Note. RoC*RoI = Risk of Reconviction*Risk of Re-imprisonment (Bakker et al., 1999).

Table 5.10
Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and DRAOR Acute Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, Estimated from the Final Multivariate Model

Hypothesized Correlation Between Predictor & Unobserved Variable (ρ)	Hypothesized Strength of Relationship Between Unobserved Variable and Outcome (I')			
	0.0	0.6	1.8	3.0
.00	1.14 [1.11, 1.17]	1.14 [1.11, 1.17]	1.14 [1.11, 1.17]	1.14 [1.11, 1.17]
.20	1.14 [1.11, 1.17]	1.08 [1.05, 1.11]	0.98 [0.95, 1.00]	0.88 [0.86, 0.90]
.50	1.14 [1.11, 1.17]	1.00 [0.97, 1.03]	0.78 [0.76, 0.80]	0.60 [0.59, 0.62]
.75	1.14 [1.11, 1.17]	0.94 [0.91, 0.96]	0.64 [0.63, 0.66]	0.44 [0.43, 0.45]

Note. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Table 5.11

Hazard ratios [95% Confidence Intervals] Expressing the Relationship Between Criminal Recidivism and DRAOR Protect Scores Under the Assumption of Varying Degrees of Unobserved Heterogeneity, Estimated from the Final Multivariate Model

Hypothesized Correlation Between Predictor & Unobserved Variable (ρ)	Hypothesized Strength of Relationship Between Unobserved Variable and Outcome (I')			
	0.0	0.6	1.8	3.0
.00	0.93 [0.90, 0.95]	0.93 [0.90, 0.95]	0.93 [0.90, 0.95]	0.93 [0.90, 0.95]
.20	0.93 [0.90, 0.95]	0.97 [0.94, 0.99]	1.05 [1.02, 1.07]	1.12 [1.10, 1.14]
.50	0.93 [0.90, 0.95]	1.03 [1.00, 1.05]	1.20 [1.19, 1.22]	1.34 [1.32, 1.35]
.75	0.93 [0.90, 0.95]	1.08 [1.07, 1.10]	1.31 [1.29, 1.32]	1.47 [1.46, 1.48]

Note. DRAOR = Dynamic Risk Assessment for Offender Re-entry (Serin, 2007).

Chapter 6: Discussion

Researching the association between risk factors and criminal justice outcomes can benefit correctional psychology by both (a) enhancing theoretical understanding of the psychosocial variables related to criminal recidivism, and (b) offering practical guidance for correctional agencies tasked with offender management. Relevant to these goals, two broad findings emerged from the project reported in this dissertation, which employed a statistical approach that models *when* individuals were at risk for recidivism, using prospective, longitudinal re-assessment data, collected by officers supervising re-entering offenders in the community.

First, results showed that offenders more often recidivated shortly (typically defined as one week, but at most, up to six weeks) following a supervision meeting when they were rated relatively higher on theoretically dynamic risk factors, compared to other supervised offenders. This finding may suggest that the set of variables measured within DRAOR assessments are good candidates to be considered (a) markers of short-term risk to re-offend, and (b) descriptive indicators of the phenomenological experiences that lead up to the commission of a crime. In particular, the *acute* subscale appeared to demonstrate the strongest evidence of proximal association with subsequent outcomes, by demonstrating statistically significant improvement in prediction accuracy over baseline assessments (as described using time-dependent AUC statistics) during some weeks of the follow-up period. The *acute* subscale of DRAOR includes variables that assess current life situation (employment, housing, criminal opportunity), interpersonal

relationships, mood, and substance use. These findings are highly consistent with the potentially proximal risk variables identified in past research, as displayed in Table 1.3.

Second, results suggested that it is meaningful for correctional staff to re-assess offenders on dynamic risk factors, since prediction models that included more proximal re-assessments resulted in a better fit with recidivism outcomes, compared to models with only relatively distal, baseline assessments. Thus, it is recommended that agencies consider employing continued re-assessment of dynamic risk variables within their correctional practice, in order to better monitor and respond to offender clients attempting to remain crime-free in the community. However, because the present project represents one of the few studies to investigate the predictive value of re-assessments, and the only presently known study to demonstrate improved prediction at a statistically significant level when comparing AUCs (in contrast to prior research that demonstrated incremental prediction in Cox models only; e.g., Howard & Dixon, 2013; Jones et al., 2010), it is recommended that an agency regularly evaluate the predictive performance of any dynamic re-entry instrument it employs.

As such, a core purpose of this dissertation was to develop a framework for testing the proximity hypothesis, which represents an important defining feature of a risk instrument purported to measure truly dynamic risk factors. In general, more research has examined whether dynamic risk factors can be observed to change across time (across rehabilitation programs, community re-entry, or forensic psychiatric management), and if change scores are related to recidivism outcomes.

By comparison, little prior research has examined whether likelihood of recidivism varies with continually updated risk factor ratings, to the degree that researchers can confidently recommend that updated assessment scores enhance predictive validity.

This dissertation proposed a framework for testing this proximity hypothesis, by employing a *Cox regression with time varying predictors* statistical model in a three-stage process. First, logically, re-assessments have added value only if they show incremental prediction above initial assessments. It would be a waste of resources to re-assess dynamic risk items if a baseline assessments have equal predictive validity as later assessments. Second, a dynamic risk instrument that is truly measuring proximal risk status should show enhanced prediction for proximal scores over (a) a grand mean score that updates with each assessment, and (b) a moving rolling mean score that averages recent assessments.

Note that a dynamic risk instrument that is shown to favor proximity in the first stage of the framework, but not in the more rigorous latter two tests, would still have provided evidence that supports the practice of re-assessment. However, that pattern of results would suggest an alternate mechanism underlying enhanced predictive validity. Specifically, improved prediction associated with averaged re-assessments suggests re-assessments serve to provide a more accurate point estimate when combined, rather than more accurate information due to proximity to the outcome.

In the analyses with DRAOR data presented in Chapter 5, results supported the conclusion that proximal assessments were more closely associated with

recidivism outcomes compared to distal assessments. Support for the proximity hypothesis remained robust, even when the most proximal assessment was averaged with recent assessments for comparison. Thus, these findings were consistent (a) in every stage of the proposed three-stage framework, (b) with all three DRAOR subscales, with only minor exceptions, and (c) to the degree that statistically significant differences were sometimes observed between baseline and re-assessment models, when examining time-dependent AUC statistics (for the *acute* and *protect* subscales).

Possible Explanations for the Pattern of Results

There are at least four different underlying reasons that may explain why the analyses provided support for the proximity hypothesis. First, it may be that the most proximal assessment represents the most ideal assessment for predictive validity, because information taken during DRAOR re-assessments represents an up-to-date description of an individual's current risk for recidivism. It may be that a dynamic instrument that assesses theoretically acute dynamic risk factors (such as DRAOR) is able to behave as a "real time" flag for short-term risk. This is certainly the goal of building dynamic risk measures for re-entry, and there are certainly theoretical reasons to suggest this should be the case.

Specifically, as discussed in Chapter 1, crime is hypothesized to result from the immediate interplay between motivational factors, individualized perceptions of the desirability of criminal action, and opportunity for crime. Whereas risk instruments that assess only background static risk factors have long been observed to be highly robust predictors of recidivism, the items within static instruments are

arguably distal markers of the truly immediate influences on crime behaviors. It may be that DRAOR items added incremental predictive validity beyond a strongly predictive static risk instrument (the RoC*RoI) because they were more timely markers of the latent features that influence, or cause, criminal recidivism.

Separately entering unchanging (static) and changing (dynamic) risk variables into a statistical model that emphasizes *timing* suggests that some features differentiate individuals from each other (i.e., degree and severity of past criminal behavior), and other features differentiate at-risk situations across any one individual's period of re-entry (i.e., time periods when *acute* subscale risk factors were high).

From a theoretical point of view, risk assessment instruments that result in scores that *cannot* change are highly problematic, given that situational, temporal, and rehabilitative factors should impact risk of criminal behavior. Also, the robust findings that support the primacy of static risk measures within correctional psychology should be seen as paradoxical to evidence of life-course change from the broader criminology literature (as summarized by Piquero, 2008). This broader literature implies that assessments of risk are, and should be, inherently unstable across time, given offenders are expected to show general patterns of decline across the life course, as well as periods of relatively rapid transition from peak offending to intermittent offending to long-term disengagement from crime. But, theory and practice in correctional psychology have been slow to recognize the implications of the body of literature on life course change, largely because (a) the time-frame of active offender management is decidedly short-term, (b) the field is still in the early stages of building measurement strategies that aim to assess who changes how and

when, and (c) the research that has retrospectively identified life-course trajectories has not revealed psychosocial variables that are obvious prospective markers that can reliably forecast future trajectories. Thus, among instruments designed for assessing the general population of offenders, DRAOR is a unique instrument for both assuming offenders' features and circumstances will change, and aiming to predict outcomes in the short-term. Also, results presented in this dissertation are unique for demonstrating short-term associations between risk factors and relatively immediate risk among general offenders.

Second, although the above explanation assumes that each re-assessment of DRAOR is a blank-slate rating of an offender's immediate situation, this is not the case. Because supervision officers rate the same clients repeatedly, they make each rating with knowledge of prior behavior and prior evaluations of that behavior. As a result, it may be that supervision officers implicitly average information in their re-assessments. If this is the case, the most proximal assessment represents not only the offender client's current status, but also everything else the rater has previously learned about the client, which is arguably embedded in their most recent rating. Epstein (1983) reported that participants appear to engage in an intuitive averaging process when they observe the same individual multiple times. He suggested human decision-makers automatically aggregate information across occasions, in a non-conscious process that mimics what occurs when a researcher might aggregate multiple observations to attain a better point estimate of an individual's behavior.

Thus, potentially undermining the analyses presented in Chapter 5 that compared averaging strategies against proximal scores, there is the possibility that

the “proximal” comparison score may, in essence, have been an implicit average. This explanation assumes that a supervision officer’s final rating contains a better averaged point estimate than could be achieved by compiling recent ratings that were explicitly recorded by that same officer. In other words, supervision officers may have rated the most proximal assessment with some knowledge of the offender’s current functioning that was not captured in the items they were purportedly rating (and was not captured in prior ratings of the same items). This may appear a criticism of the training process, but is a possibility, given prior research has suggested well-trained officers may even consciously make rating decisions that will support their preferred course of offender management (Guy et al., 2014). Also recall that Jones and colleagues (2010) showed that officers were able to rate items in a way that attained predictive validity, even though their ratings were not related with self-report versions of the same construct, suggesting the officers were able to differentiate between recidivists and non-recidivists, even if they may not have been assessing the same constructs they were explicitly meant to be rating.

One way to test this possibility is to conduct qualitative interviews with DRAOR raters who can describe the phenomenological experience of rating the items during case management; this type of important research has been rarely conducted (Litwack, 2002), and may be an even more important avenue of research among officers who are tasked with re-assessing the same items on the same offender. Such qualitative research may hint at whether there may be some

unknown, conscious or non-conscious weighting process used by supervision officers that serves to maximize the predictive validity of their ratings.

Third, related to the point above, supervision officers may have had a particularly crucial piece of information at the time of rating the most proximal DRAOR that would not have been explicitly reflected in the assessment items: knowledge that they were on the verge of citing their client for a breach of conditions. In this project, it is a potential limitation that recidivism outcome was defined broadly to include technical violations. (However, DRAOR was specifically designed to predict general breaches and criminal behavior, rather than criminal behavior only.) If supervision officers were considering or were planning to initiate a technical violation at the time of the most proximal DRAOR assessment, it is likely this would have influenced the rating of the items.

In these data, the precise date of new criminal actions and technical violations were registered into the system; however, it is unknown how far in advance a supervision officer may have begun considering a client to have breached his or her release conditions. Although care was taken to stagger all assessment data so they were used to prospectively predict recidivism dates the following week, this may have been undermined by rater foreknowledge of upcoming technical violations. Still, supervision officers undergo a decision process where they are most likely to breach offenders with concerning behavior (Yesberg et al., 2015); thus, even if proximal assessment ratings most directly reflect supervision officer *decisions*, these decisions were likely based on some problematic behavior that was at least partially, and possibly fully, reflected in DRAOR content.

If this explanation has merit, within each sequence, one might expect there to be a dramatic difference between earlier ratings and final (proximal) ratings. In other words, if proximal ratings were highly influenced by the knowledge of upcoming breach proceedings, a substantial increase compared to earlier ratings should be observed during the final week of each recidivist's sequence. In general, this is not the case. Instead, among recidivists on the *stable* and *acute* subscales, the difference between the second-to-the-last assessment and the final assessment was 0.1 point higher at the final assessment. Although these differences were statistically significant,¹ comparison with Table 4.6 suggests that this amount of change is *less* than the typical week-to-week change observed across recidivists. Thus, these data provide no substantial evidence that the most proximal DRAOR assessments were unusually high among recidivists. Instead, the amount of change observed at the end of sequences was consistent with the general pattern of change across the entire follow-up period.

Thus, fourth, perhaps the most helpful way to interpret this project's findings is within the context of how eventual recidivists and non-recidivists differed from each other in their general patterns of change. Specifically, the underlying computation within Cox regression is designed, at each point in time, to compare the scores of recidivists to non-recidivists who still remain in the sample un-censored

¹ For recidivists, the mean change on *stable* comparing second-to-last and last assessments was 0.10, $SD = 1.15$ ($M = 6.49$ to 6.58 , $SD = 2.84$ for both, $t = 3.06$, $p = .002$), and on *acute* was 0.10, $SD = 1.42$ ($M = 5.75$ to 5.86 , $SD = 2.66$ and 2.64 , respectively, $t = 2.70$, $p = .007$). The mean difference between *protect* scores was not statistically significant ($M = -0.05$, $SD = 1.26$, $M = 6.10$ to 6.05 , $SD = 2.72$ and 2.70 , respectively, $t = -1.51$, $p = .13$). For non-recidivists, comparing mean change resulted in a statistically significant difference for *stable* scores only ($M = -0.03$, $SD = 0.65$, $M = 5.21$ to 5.18 , $SD = 2.80$ and 2.83 , respectively, $t = -1.98$, $p = .05$), and statistically non-significant comparisons for *acute* ($M = 0.01$, $SD = 1.07$, $M = 4.20$ to 4.21 , $SD = 2.31$ and 2.40 , respectively, $t = 0.50$, $p = .62$), and *protect* ($M = 0.03$, $SD = 0.78$, $M = 7.33$ to 7.37 , $SD = 2.60$ and 2.65 , respectively, $t = 1.81$, $p = .07$).

from observation. As a result, this analytic method is able to capitalize on two features to attain a pattern of results that prefers proximal over distal scores: (a) eventual recidivists began supervision with higher DRAOR risk scores than eventual non-recidivists, on average, and (b) non-recidivists showed more consistent and precipitous declines in risk scores over time. As a result, across all points in time, the most proximal scores were always the set of scores that showed the greatest gap between these two groups, and this gap only widened with time.

Put this way, re-assessment is important, not (at least not primarily) because the dynamic instrument is able to “flag” high-risk periods among imminent recidivists, but because it allows eventual non-recidivists to have their risk levels downgraded to appropriately lower levels. Rather than focusing on hypothesized, but largely unobserved peaks in risk, this explanation emphasizes that what is most meaningful is that (a) eventual recidivists did not show downward change, but maintained relatively higher DRAOR risk scores across time (although this group still showed decline, on average), and (b) eventual non-recidivists were regularly down-rated on risk scores across time, on average.

Conceptually, it is consistent that eventual recidivists would begin supervision with higher risk scores, and generally maintain their higher risk status over time. By contrast, more explanation is required to understand why non-recidivists tend to begin supervision with higher risk scores, and decline across time. That risk scores decline in severity across time is a known feature of re-entry risk data, and has been shown in other re-assessment datasets (Babchishin, 2013; Greiner et al., 2014; Jones et al., 2010; Michel et al., 2013; although, Jones et al., 2010,

showed a counter-intuitive finding where problems with finances and use of leisure time increased across time among non-recidivating offenders).

There are at least two possible explanations. First, at the start of supervision, it may be that eventual non-recidivists' "true" risk scores are low, but they are perceived as higher risk. This may be a result of supervision officers (a) taking a cautious stance toward all re-entering offenders, such that any positive features are viewed as temporary, unstable, or unsure, or (b) rating items for which they have little or no information, but choosing to rate the items with conservative caution in the meantime. Or, it may be that incarceration has temporarily stripped lower-risk offenders of any overt manifestations of their lower-risk status, making it impossible for a rater to perceive their true risk level upon re-entry. As such, patterns of decline in risk scores across time may not represent change, but declining scores may be increasingly better approximations of the individual's truly low risk status, as the client stabilizes in the community, and the rater begins to trust that positive features will remain stable. Similarly, Jones and colleagues (2010) suggested that supervision officers' expectations of clients' levels of functioning may change across time in a way that influences how they rate dynamic risk items. (Although, this is a hypothesis of measurement variance, which can be empirically tested, see discussion in Chapter 3.)

Second, it may be that eventual non-recidivists have a truly high-risk status at the start of supervision, but observed declines in risk scores represent true decreases in risk for recidivism. True latent risk may decline over time because (a) early re-entry periods are filled with stressors, situations, and temptations that

initially elevate risk, but dissipate over time, (b) offenders are initially inexperienced at enacting the personal changes they desire, but become more proficient with time, or (c) the experience of incarceration temporarily but initially impacts risk status after release. Certainly, at a group level of analysis, the early months following re-entry are the months with the highest rates of recidivism, but further research is required to build a case that intra-individual changes in risk factors are a legitimate phenomenon, rather than a manifestation of another process.

In summary, although the Cox regression approach does not explicitly model within-individual patterns of change (as explained in Chapter 2), it is important to interpret results in the context of the patterns of change observed within the sample. Results from the dataset used in this dissertation suggest that risk score ratings tended to decline across time, on average, and eventual recidivists and non-recidivists differed in their degree of decline in expected ways (see Hanby, 2013, and Chapter 4 of this dissertation). As a result, more proximal assessments generally had stronger predictive validity compared to more distal assessments because non-recidivists were rated to have increasingly lower risk over time.

Limitations & Future Directions

Several limitations of the dataset used in this project were described in Chapter 3. Specifically, recall that the dataset does not contain information that would allow analyses to separate out differential effects of supervision officers. In particular, officers may have differed from each other in how effectively they were trained to rate DRAOR items with fidelity, how effectively they rated DRAOR items in practice, and how effectively they implemented intervention strategies with

offender clients (e.g., building positive relationships, setting sanctions, providing resources, etc.). If it were possible to analyze these nested data with information about how offenders were nested within supervision officers, results would provide better estimates of the relationship between risk scores and outcomes, by accounting for some of this cluster-specific variance.

Although attaining all levels of information embedded within a dataset is important for analyses, the key conceptual implication is that implementation fidelity deserves focused research attention. Correctional psychology has convincingly demonstrated that risk instruments can be designed to have good predictive validity, but has generated little knowledge about how to implement risk instruments into the field with fidelity (Nonstad & Webster, 2011). Thus, future research with DRAOR should attend to fidelity in measurement, and fidelity in practice, including (a) inter-rater reliability, (b) construct validity, (c) training slippage, (d) instrument “buy-in” among supervision officers, (e) rater comfort assessing items with more psychological content, (f) rater fatigue when assessing the same client across time, and (g) how supervision officers use knowledge of item ratings to choose intervention strategies.

For assessing prediction accuracy, it can be considered a limitation that the supervision officers engaged in case management were also rating (and thus were not blind to) the content of their clients’ DRAOR items. In many ways, this is not a limitation; this reflects typical practice in community corrections settings (which allows these research results to have good external validity), and represents best practice for ensuring that supervision officers are familiar with and can readily

respond to their clients' needs. However, assuming that supervision officers attempted to intervene when they noted they were rating a client with a high DRAOR risk score (which should be a reasonable assumption), such intervention strategies would have effectively reduced the relationship between high-risk scores and outcomes that were averted due to effective management.² Thus, the contamination in the research design that occurs when raters enact the dual roles of risk assessment and risk management serves to place a lower limit on the degree of prediction accuracy that can be obtained, even though this feature maximizes public safety.

Little research has examined whether implementing a risk tool can impact recidivism rates (Desmarais & Singh, 2013), but there are a number of reasons to suggest that risk assessors should be more effective in their attempts to intervene. Specific to instruments similar to DRAOR, Day (2013) argued that continuous re-assessment should alter supervision practice by re-focusing the officer's attention from setting constraints on the offender to intervening in a way that will produce observable change. Also, arguably, training supervision officers to rate items conceptualized as protective factors should have a profound impact on the nature of the conversations officers would otherwise typically use to elicit information for rating items. And, as Yesberg and Polaschek (2014) argued, if using the DRAOR in practice alters officer-client interactions, it may be that these resulting case management strategies are more strongly associated with recidivism outcomes,

² However, presumably, after the intervention was successful, and the feared outcome was averted, the client would have remained in the community with newly lowered risk scores, effectively following the pattern of a non-recidivist in the dataset, and contributing in that way to the accurate predictive validity of DRAOR.

rather than the DRAOR scores themselves. Thus, future avenues of research should include directly testing the impact of DRAOR implementation on offender outcomes (i.e., by randomly assigning DRAOR training to different officers or sites), and examining whether the relationship between DRAOR scores and offender outcomes is potentially mediated by intervention strategies employed by supervision officers.

An additional feature of the dataset that limits how results should be interpreted deserves to be highlighted a second time. As discussed in earlier chapters, the New Zealand Department of Correction's policies about assigning length of parole results in there being an important distinction between the demographic characteristics of offenders who were observed up to a maximum of six months following release, and offenders observed longer than six months. Specifically, higher risk offenders were generally only supervised for no more than six months, whereas lower risk offenders may have been supervised for longer than six months. In some analyses, this demarcation had noticeable effects, and it is important to exercise caution when interpreting results in a way that summarizes findings across the entire year of follow-up. For example, some observed differences in predictive validity that occurred only after the six-month mark (for example, with the *protect* subscale) should be specifically interpreted as stronger predictive validity *among offenders who were observed for longer than six months*.

Finally, it may be a source of confusion that DRAOR's subscales have been named *stable* and *acute* for conceptual reasons (i.e., items chosen for these subscales are *theoretically* stable dynamic and acute dynamic risk factors), while the present project is among the first to provide empirical support for these labels. More

narrowly, this project (alongside Hanby, 2013) most accurately shows that these subscale *total scores* can be observed to be dynamic (i.e., scores on both subscales were observed to change across time) in relatively stable and acute ways, respectively (i.e., *stable* scores demonstrated more stability across time than *acute* scores). However, only longitudinal re-assessment research can determine whether an item is truly stable or acute (Yesberg & Polaschek, 2014), and the present research did not examine change or prediction on an item level. This is an important avenue for future research, to empirically determine how constructs captured within DRAOR items vary over time. A construct-level understanding of risk items is critical for building intelligible risk instruments (Hanson et al., 2013), even if it is expected that risk variables will be more strongly predictive in aggregate form, rather than when isolated into single variables (Kroner et al., 2007; Vose, 2008).

In addition, empirically separating DRAOR items into constructs may provide opportunities for advancing the correctional theory of criminal decision-making. Specifically, as argued in Chapter 1, it is important to investigate how immediate decisions to commit crime are influenced by immediate external circumstances, interpersonal influences, and personal features. When items with clear relevance to the theoretical causes of criminal actions are re-assessed in close proximity to criminal behavior, there is an opportunity for investigating these hypotheses, as long as the construct content of the items is understood. For theory building, it is equally important to supplement DRAOR data with self-report items assessed on a similar schedule, to both investigate the associations between DRAOR content and

self-report content, and to examine how constructs measured by actors versus observers mutually evolve across time, and may ultimately be related to recidivism outcomes. An individual's beliefs toward crime are a critical component of all the current comprehensive theories of criminal action, and advancing these theories will rest upon measuring these beliefs alongside other conceptualized influences on crime behavior. Whereas developmental psychologists hold an important piece by investigating how beliefs about crime develop in early life, correctional psychologists are responsible for understanding how these beliefs are maintained through adulthood, or are altered during re-entry.

Recommendations

- 1) Broadly, conducting longitudinal research on dynamic risk factors is highly recommended (Farrington, 2011; Hanson & Harris, 2000; Olver, Beggs Christofferson, et al., 2014), because only longitudinal research can effectively address the core assumptions embedded within contemporary correctional practice, including the hypotheses that (a) the risk factors that differentiate offenders and non-offenders are also the risk factors that influence within-individual patterns of criminal action, (b) rehabilitation programs are effective because they reduce known risk factors, and (c) offenders are at higher risk of recidivism when acute dynamic risk factors are present.
- 2) This dissertation proposes a three-step framework for investigating the hypothesis assumed by the definition of a *dynamic risk factor*, specifically, that more proximal assessments of risk factors have higher predictive

validity for short-term recidivism outcomes compared to prior assessments.

It is recommended that researchers consider applying this framework to their own data, with the suggestion that an instrument containing theoretically dynamic factors should only be considered to have potential for proximal prediction if these tests are successful, at minimum.

- 3) It is suggested that the proposed modeling approach (i.e., Cox regression with time-varying predictors) will be particularly useful to researchers whose re-entry re-assessment data have the following features: (a) data are taken on a variable schedule with a relatively large number of assessments per individual, (b) there is a clearly defined “beginning of *time*” for all participants, such as the date each individual was released from prison, (c) there is a clearly defined binary outcome event (i.e., recidivism) that removes the individual from further observation (i.e., revocation of release), and (d) the exact dates of assessments and outcome events are known, or can be reasonably ordered across time to ensure prospective prediction.
- 4) It is recommended that correctional researchers who employ Cox regression models (with or without time-varying predictors) report effect size indices within their results. Specifically, Xu and O’Quigley’s (1999) R^2 equivalent, the *c*-index (Harrell et al., 1982), and time-dependent AUC statistics (Heagerty & Zheng, 2005) are recommended as good candidates for assessing the comparative value of models. In particular, within this dissertation, results using these effect size indices served to temper over-zealous interpretation of statistically significant incremental prediction within the Cox model itself.

In essence, it is highly possible that a variable that adds incremental effects in a regression model may make minimal contributions to the amount of variance explained, or the accurate ranking of recidivists versus non-recidivists.

- 5) Finally, it is recommended that researchers who employ this statistical model and framework refrain from building tests on data that were ordered incorrectly across time. Although it may appear intuitive to test some of these models, experience with the present dataset revealed counter-intuitive results that required exerting some careful thought about the nature of the formulae's calculations in order to resolve. For example, one might use Cox regression to compare individuals' grand mean scores across all assessments against their most proximal scores; however, note that in the early time periods of the follow-up, an individual's grand mean has effectively "pulled forward" information that was not available at that point in time, in order to create the grand mean average. This violates a core assumption of Cox regression with time varying predictors, and will likely lead to counter-intuitive results. Similarly, other averaging strategies that served to insert assessment information earlier than its known sequence within time were attempted within this project, but not reported in this dissertation, and are not recommended, for reasons of conceptual clarity and model integrity.

References

- Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*, 341-382.
doi:10.1177/0011000005285875
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211. doi:10.1016/0749-5978(91)90020-T
- Ajzen, I., & Fishbein, M. (1969). The prediction of behavioral intentions in a choice situation. *Journal of Experimental Social Psychology, 5*, 400-416.
doi:10.1016/0022-1031(69)90033-X
- Andrews, D. A. (1995). The psychology of criminal conduct and effective treatment. In J. McGuire (Ed.), *What works: Reducing reoffending: Guidelines from research and practice* (pp. 35-62). New York: Wiley.
- Andrews, D. A., & Bonta, J. (1995). *The Level of Service Inventory-Revised (LSI-R)*. Toronto, ON: Multi-Health Systems, Inc.
- Andrews, D. A., & Bonta, J. L. (2010). *The psychology of criminal conduct*. Cincinnati, OH: Anderson.
- 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:10.1177/0011128705281756
- Andrews, D. A., 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

Andrews, D. A., & Robinson, D. (1984). *The Level of Supervision Inventory: Second report*. Toronto, ON: Ontario Ministry of Correctional Services.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, 16, 397-438. doi:10.1080/10705510903008204

Austin, J., Coleman, D., Peyton, J., & Johnson, K. D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument: Final report*. Washington, DC: Institute on Crime, Justice and Corrections at the George Washington University.

Babchishin, K. M. (2013). *Sex offenders do change on risk-relevant propensities: Evidence from a longitudinal study of the Acute-2007* (Unpublished doctoral dissertation). Carleton University, Ottawa, ON.

Babchishin, K. M., Hanson, R. K., & Helmus, L. (2012a). Communicating risk for sex offenders: Risk ratios for Static-2002R. *Sexual Offender Treatment*, 7(2), 1-12. Retrieved from <http://www.sexual-offender-treatment.org>.

Babchishin, K. M., Hanson, R. K., & Helmus, L. (2012b). Even highly correlated measures can add incrementally to predicting recidivism among sex offenders. *Assessment*, 19, 442-461. doi:10.1177/1073191112458312

Baglivio, M., & Jackowski, K. (2013). Evaluating the effectiveness of a victim impact intervention through the examination of changes in dynamic risk scores. *Criminal Justice Policy Review*, 26, 7-28. doi:10.1177/0887403413489706

- Baird, C. (2009). *A question of evidence: A critique of risk assessment models used in the justice system*. Oakland, CA: National Council on Crime and Delinquency. Retrieved from <http://www.nccdglobal.org>.
- Bakker, L., Riley, D., & O'Malley, J. (1999). *Risk of reconviction: Statistical models predicting four types of re-offending*. Wellington, NZ: Department of Corrections Psychological Service.
- Banse, R., Koppehele-Gossel, J., Kistemaker, L. M., Werner, V. A., & Schmidt, A. F. (2013). Pro-criminal attitudes, intervention, and recidivism. *Aggression and Violent Behavior, 18*, 673-685. doi:10.1016/j.avb.2013.07.024
- Barbaree, H. E., Langton, C. M., & Blanchard, R. (2007). Predicting recidivism in sex offenders using the VRAG and SORAG: The contribution of age-at-release. *International Journal of Forensic Mental Health, 6*, 29-46. doi:10.1080/14999013.2007.10471247
- Barbaree, H. E., Langton, C. M., Blanchard, R., & Boer, D. P. (2008). Predicting recidivism in sex offenders using the SVR-20: The contribution of age-at-release. *International Journal of Forensic Mental Health, 7*, 47-64. doi:10.1080/14999013.2008.9914403
- Barbaree, H. E., Langton, C. M., Blanchard, R., & Cantor, J. M. (2009). Aging versus stable enduring traits as explanatory constructs in sex offender recidivism: Partitioning actuarial prediction into conceptually meaningful components. *Criminal Justice and Behavior, 36*, 443-465. doi:10.1177/0093854809332283

- Barnao, M., & Ward, T. (2015). Sailing uncharted seas without a compass: A review of interventions in forensic mental health. *Aggression and Violent Behavior, 22*, 77-86. doi:10.1016/j.avb.2015.04.009
- Barnett, G. D., Wakeling, H., Mandeville-Norden, R., & Rakestrow, J. (2013). Does change in psychometric test scores tell us anything about risk of reconviction in sexual offenders? *Psychology, Crime & Law, 19*, 85-110. doi:10.1080/1068316X.2011.607820
- Beech, A., Friendship, C., Erikson, M., & Hanson, R. K. (2002). The relationship between static and dynamic risk factors and reconviction in a sample of UK child abusers. *Sexual Abuse: A Journal of Research and Treatment, 14*, 155-167. doi:10.1023/A:1014672231744
- Beggs, S. (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. (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
- Bergeron, C. L., & Miller, H. A. (2013). Tracking change through treatment with the inventory of offender risk, needs, and strengths. *Psychological Assessment, 25*, 979-990. doi:10.1037/a0033190
- Blanchard, A. J. E. (2013). *Dynamic risk factors in violence risk assessment: A multiple time-point evaluation of the HCR-20 and START* (Master's thesis). Available from ProQuest Dissertations and Theses database. (UMI No. MS24170)

- Blanton, H., & Jaccard, J. (2006a). Arbitrary metrics in psychology. *American Psychologist*, 61, 27-41. doi:10.1037/0003-066X.61.1.27
- Blanton, H., & Jaccard, J. (2006b). Arbitrary metrics redux. *American Psychologist*, 61, 62-71. doi:10.1037/0003-066x.61.1.62
- 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: Sage Publications.
- 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., & Heij, D. (2014). Improving agreement about intervention plans in probation by decision support. *International Journal of Offender Therapy and Comparative Criminology*. Advance online publication. doi:10.1177/0306624X14539127
- Brooks Holliday, S., Heilbrun, K., & Fretz, R. (2012). Examining improvements in criminogenic needs: The risk reduction potential of a structured re-entry program. *Behavioral Sciences & the Law*, 30, 431-447. doi:10.1002/bsl.2016
- Brown, S. L., St. Amand, M. D., & 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
- Buchanan, A., Nich, C., Douglas, K. S., Babuscio, T., & Easton, C. J. (2013). Risk factors of violence during a 4-week period in a psychiatric outpatient population.

The Journal of Nervous and Mental Disease, 201, 1021-1026.

doi:10.1097/NMD.0000000000000061

Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multilevel mixed modelling. *Statistics in Medicine*, 17, 1261-1291. doi:10.1002/(SICI)1097-0258(19980615)17:11<1261::AID-SIM846>3.0.CO;2-Z

Campbell, M. A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior*, 36, 567-590.
doi:10.1177/0093854809333610

Cardwell, S. M., Piquero, A. R., Jennings, W. G., Copes, H., Schubert, C. A., & Mulvey, E. P. (2015). Variability in moral disengagement and its relation to offending in a sample of serious youthful offenders. *Criminal Justice and Behavior*, 42, 819-839. doi:10.1177/0093854814567472

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 U.S. probationers and parolees* (Unpublished master's thesis). Carleton University, Ottawa, ON.

- Chadwick, N., Serin, R. C., & Lloyd, C. D. (2015, June). *Measuring offender change and the impact on predicting recidivism*. Paper presented to the North American Correctional and Criminal Justice Psychology Conference, Ottawa, ON.
- Chen, X. (2010). Desire for autonomy and adolescent delinquency: A latent growth curve analysis. *Criminal Justice and Behavior*, 37, 989-1004.
doi:10.1177/0093854810367481
- Childs, K. K., Sullivan, C. J., & Gullledge, L. M. (2011). Delinquent behavior across adolescence: Investigating the shifting salience of key criminological predictors. *Deviant Behavior*, 32, 64-100. doi:10.1080/01639621003748498
- Chu, C. M., Thomas, S. D., Ogloff, J. R., & Daffern, M. (2013). The short- to medium-term predictive accuracy of static and dynamic risk assessment measures in a secure forensic hospital. *Assessment*, 20, 230-241.
doi:10.1177/10731911111418298
- Cohen, J. (1988). *Statistical power analyses for the behavioral sciences* (2nd ed.). New York: Lawrence Erlbaum Associates.
- Coid, J. W., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Farrington, D., & Rogers, R. (2011). Most items in structured risk assessment instruments do not predict violence. *Journal of Forensic Psychiatry & Psychology*, 22, 3-21.
doi:10.1080/14789949.2010.495990
- Conner, M., & Armitage, C. J. (1998). Extending the theory of planned behavior: A review and avenues for further research. *Journal of Applied Social Psychology*, 28, 1429-1464. doi:10.1111/j.1559-1816.1998.tb01685.x

- Day, A. (2013). The supervision of known sex offenders in the community: Monitoring progress and providing feedback. *Sexual Offender Treatment, 8*(2), 1-9. Retrieved from <http://www.sexual-offender-treatment.org>.
- de Bruijne, M. H., le Cessie, S., Kluin-Nelemans, H. C., & van Houwelingen, H. C. (2001). On the use of Cox regression in the presence of an irregularly observed time-dependent covariate. *Statistics in Medicine, 20*, 3817-3829. doi:10.1002/sim.1083
- Dekker, F. W., de Mutsert, R., van Dijk, P. C., Zoccali, C., & Jager, K. J. (2008). Survival analysis: Time-dependent effects and time-varying risk factors. *Kidney International, 74*, 994-997. doi:10.1038/ki.2008.328
- Desmarais, S. L., & Singh, J. P. (2013). *Risk assessment instruments validated and implemented in correctional settings in the united states*. Washington, DC: Council of State Governments Justice Center. Retrieved from <http://csgjusticecenter.org>.
- Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict inpatient aggression: Reliability and validity of START assessments. *Psychological Assessment, 24*, 685-700. doi:10.1037/a0026668
- Desmarais, S. L., Van Dorn, R. A., Telford, R. P., Petrila, J., & Coffey, T. (2012). Characteristics of START assessments completed in mental health jail diversion programs. *Behavioral Sciences & the Law, 30*, 448-469. doi:10.1002/bsl.2022

- Devlin, R. S., & Gibbs, J. C. (2010). Responsible adult culture (RAC): Cognitive and behavioral changes at a community-based correctional facility. *Journal of Research in Character Education*, 8(1), 1-20. Retrieved from <http://character.org>.
- 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
- Dickson, S. R., Polaschek, D. L., & Casey, A. R. (2013). Can the quality of high-risk violent prisoners' release plans predict recidivism following intensive rehabilitation? A comparison with risk assessment instruments. *Psychology, Crime & Law*, 19, 371-389. doi:10.1080/1068316x.2011.640634
- 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
- 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. (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

- Farrington, D. P. (2011). The integrated cognitive antisocial potential (ICAP) theory. In D. P. Farrington (Ed.), *Integrated developmental and life-course theories of offending* (pp. 73-92). New Brunswick, NJ: Transaction.
- Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: Systematic review and meta-analysis. *British Medical Journal*, *345*, 1-12. doi:10.1136/bmj.e4692
- Fitzmaurice, G. M., & Ravichandran, C. (2008). A primer in longitudinal data analysis. *Circulation*, *118*, 2005-2010. doi:10.1161/CIRCULATIONAHA.107.714618
- 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
- Ghisletta, P., & Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics*, *29*, 421-437. doi:10.3102/10769986029004421
- Gottfredson, S. D., & Moriatry, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime & Delinquency*, *52*, 178-200. doi:10.1177/0011128705281748
- Gray, M. K., Fields, M., & Maxwell, S. R. (2001). Examining probation violations: Who, what, and when. *Crime & Delinquency*, *47*, 537-557. doi:10.1177/0011128701047004003

- Greiner, L. E., Law, M. A., & Brown, S. L. (2014). Using dynamic factors to predict recidivism among women: A four-wave prospective study. *Criminal Justice and Behavior*, 42, 457-480. doi:10.1177/0093854814553222
- Guy, L. S., Nelson, R. J., Fusco-Morin, S. L., & Vincent, G. M. (2014). What do juvenile probation officers think of using the SAVRY and YLS/CMI for case management, and do they use the instruments properly? *International Journal of Forensic Mental Health*, 13, 227-241. doi:10.1080/14999013.2014.939789
- Haas, S. M., & DeTardo-Bora, K. A. (2009). Inmate reentry and the utility of the LSI-R in case planning. *Corrections Compendium*, 34(1), 11-16. Retrieved from <http://www.jrsa.org>.
- Hagger, M. S., Chatzisarantis, N. L., & Biddle, S. J. (2002). A meta-analytic review of the theories of reasoned action and planned behavior in physical activity: Predictive validity and the contribution of additional variables. *Journal of Sport & Exercise Psychology*, 24(1), 3-32.
- Hanby, L. J. (2013). *The relationship between dynamic risk, protective factors, and criminal recidivism in a New Zealand offender release cohort* (Unpublished doctoral dissertation). Carleton University, Ottawa, ON.
- Hanley, J. A., Negassa, A., Edwardes, M. D. deB., & Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology*, 157, 364-375. doi:10.1093/aje/kwf215

Hanson, R. K. (2008). What statistics should we use to report predictive accuracy?
Crime Scene, 15(1), 15-17.

Hanson, R. K. (2009). The psychological assessment of risk for crime and violence.
Canadian Psychology, 50, 172-182. doi:10.1037/a0015726

Hanson, R. K., Babchishin, K. M., Helmus, L., & Thornton, D. (2013). Quantifying the
relative risk of sex offenders: Risk ratios for Static-99R. *Sexual Abuse: A
Journal of Research and Treatment, 25*, 482-515.
doi:10.1177/1079063212469060

Hanson, R. K., & Harris, A. J. (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., Harris, A. J. R., Scott, T., & Helmus, L. (2007). *Assessing the risk of sexual
offenders on community supervision: The Dynamic Supervision Project*. User
Report No. 2007-05. Ottawa: Public Safety Canada.

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

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders:
A comparison of three actuarial scales. *Law and Human Behavior, 24*, 119-
136. doi:10.1023/A:1005482921333

Hanson, R. K., Thornton, D., Helmus, L. M., & Babchishin, K. M. (2015). What sexual
recidivism rates are associated with Static-99R and Static-2002R scores?

Sexual Abuse: A Journal of Research and Treatment. Advance online publication. doi:10.1177/1079063215574710

Harrell, F. E. (2001). *Springer Series in Statistics: Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA: Journal of the American Medical Association*, 247, 2543-2546. doi:10.1001/jama.1982.03320430047030

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361-387.

doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Hart, S. D. (1999). Assessing violence risk: Thoughts and second thoughts.

Contemporary Psychology, 44(6), 486-487.

Hart, S. D., Michie, C., & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments. *The British Journal of Psychiatry*, 190, s60-s65.

doi:10.1192/bjp.190.5.s60

Heagerty, P. J., & Saha-Chaudhuri, P. (2015). *Riskset ROC curve estimation from censored survival data*. Retrieved from <https://cran.r-project.org/web/packages/risksetROC/risksetROC.pdf>.

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1), 92-105.

- Helmond, P., Overbeek, G., Brugman, D., & Gibbs, J. C. (2014). A meta-analysis on cognitive distortions and externalizing problem behavior: Associations, moderators, and treatment effectiveness. *Criminal Justice and Behavior*, 42, 245-262. doi:10.1177/0093854814552842
- Helmus, L., 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
- Helmus, L., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. R. (2012). Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples: A meta-analysis. *Criminal Justice and Behavior*, 39, 1148-1171. doi:10.1177/0093854812443648
- Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: A Journal of Research and Treatment*, 24, 64-101. doi:10.1177/1079063211409951
- 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
- Howard, P. D., & Moore, R. (2009). *Measuring change in risk and need over time using OASys*. (Research summary 10/09). Ministry of Justice. Retrieved from <http://webarchive.nationalarchives.gov.uk>.

- Hox, J. J. (2010). *Quantitative Methodology Series: Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.
- 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. doi:10.1037/0022-006X.59.1.12
- 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
- Kattan, M. W. (2003). Judging new markers by their ability to improve predictive accuracy. *Journal of the National Cancer Institute, 95*, 634-635.
doi:10.1093/jnci/95.9.634
- Kazemian, L., Farrington, D. P., & Le Blanc, M. (2009). Can we make accurate long-term predictions about patterns of de-escalation in offending behavior? *Journal of Youth and Adolescence, 38*, 384-400. doi:10.1007/s10964-008-9338-z
- Kent, J. T., & O'Quigley, J. (1988). Measures of dependence for censored survival data. *Biometrika, 75*, 525-534. doi:10.1093/biomet/75.3.525
- 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
- Krauss, D. A., Sales, B. D., Becker, J. V., & Figueredo, A. J. (2000). Beyond prediction to explanation in risk assessment research: A comparison of two explanatory theories of criminality and recidivism. *International Journal of Law and Psychiatry*, 23, 91-112. doi:10.1016/S0160-2527(99)00032-1
- 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
- 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
- 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
- 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

Liang, K. -Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22. doi:10.1093/biomet/73.1.13

Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 54(3), 948-963.

Lindsay, W. R., Murphy, L., Smith, G., Murphy, D., Edwards, Z., Chittock, C., . . . Young, S. J. (2004). The Dynamic Risk Assessment and Management System: An assessment of immediate risk of violence for individuals with offending and challenging behaviour. *Journal of Applied Research in Intellectual Disabilities*, 17, 267-274. doi:10.1111/j.1468-3148.2004.00215.x

Lipsitz, S., & Fitzmaurice, G. (2009). Generalized estimating equations for longitudinal data analysis. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Handbook of Modern Statistical Methods: Longitudinal data analysis* (pp. 43-78). Boca Raton, FL: Chapman & Hall/CRC.

Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development*, 31, 357-365. doi:10.1177/0165025407077757

Litwack, T. R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law*, 7, 409-443. doi:10.1037//1076-8971.7.2.409

- Litwack, T. R. (2002). Some questions for the field of violence risk assessment and forensic mental health: Or, "back to basics" revisited. *International Journal of Forensic Mental Health, 1*, 171-178. doi:10.1080/14999013.2002.10471171
- Lloyd, C. D., & Serin, R. C. (2014). Offender change in treatment. In G. Bruinsma & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 3301-3311). New York: Springer. doi:10.1007/978-1-4614-5690-2
- Lofthouse, R. E., Lindsay, W. R., Totsika, V., Hastings, R. P., & Roberts, D. (2014). Dynamic risk and violence in individuals with an intellectual disability: Tool development and initial validation. *The Journal of Forensic Psychiatry & Psychology, 25*, 288-306. doi:10.1080/14789949.2014.911946
- Maden, A. (2003). Standardised risk assessment: Why all the fuss? *Psychiatric Bulletin, 27*, 201-204. doi:10.1192/pb.27.6.201
- 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
- Maruna, S. (2001). *Making good: How ex-convicts reform and rebuild their lives*. Washington, DC: American Psychological Association. doi:10.1037/10430-000
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577-605. doi:10.1146/annurev.psych.60.110707.163612

- McArdle, J. J., & Hamagami, F. (2004). Methods for dynamic change hypotheses. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Mathematical modelling: Theory and applications: Recent developments on structural equation models* (pp. 295-335). Norwell, MA: Kluwer Academic Publishers.
- McDermott, B. E., Edens, J. F., Quanbeck, C. D., Busse, D., & Scott, C. L. (2008). Examining the role of static and dynamic risk factors in the prediction of inpatient violence: Variable- and person-focused analyses. *Law and Human Behavior*, 32, 325-338. doi:10.1007/s10979-007-9094-8
- McGrath, A., & Thompson, A. P. (2012). The relative predictive validity of the static and dynamic domain scores in risk-need assessment of juvenile offenders. *Criminal Justice and Behavior*, 39, 250-263. doi:10.1177/0093854811431917
- McNiel, D. E., Lam, J. N., & Binder, R. L. (2000). Relevance of interrater agreement to violence risk assessment. *Journal of Consulting and Clinical Psychology*, 68, 1111-1115. doi:10.1037//0022-006x.68.6.1111
- Michel, S. F., Riaz, M., Webster, C., Hart, S. D., Levander, S., Müller-Isberner, R., . . . Hodgins, S. (2013). Using the HCR-20 to predict aggressive behavior among men with schizophrenia living in the community: Accuracy of prediction, general and forensic settings, and dynamic risk factors. *International Journal of Forensic Mental Health*, 12, 1-13. doi:10.1080/14999013.2012.760182
- Miller, J., & Maloney, C. (2013). Practitioner compliance with risk/needs assessment tools: A theoretical and empirical assessment. *Criminal Justice and Behavior*, 40, 716-736. doi:10.1177/0093854812468883

- Mills, J. F., & Kroner, D. G. (2006). The effect of discordance among violence and general recidivism risk estimates on predictive accuracy. *Criminal Behaviour and Mental Health*, 16, 155-166. doi:10.1002/cbm.623
- Millson, W. A., Weekes, J. R., & Lightfoot, L. O. (1995). *The offender substance abuse pre-release program: Analysis of intermediate and post-release outcomes* (Research Report No. 40). Retrieved from Correctional Service of Canada website: <http://www.csc-scc.gc.ca/research/r40e-eng.shtml>.
- Monahan, K. C., & Piquero, A. R. (2009). Investigating the longitudinal relation between offending frequency and offending variety. *Criminal Justice and Behavior*, 36, 653-673. doi:10.1177/0093854809335527
- Morgan, R. D., Kroner, D. G., Mills, J. F., Serna, C., & McDonald, B. (2013). Dynamic risk assessment: A validation study. *Journal of Criminal Justice*, 41, 115-124. doi:10.1016/j.jcrimjus.2012.11.004
- Müller-Isberner, R., Webster, C. D., & Gretenkord, L. (2007). Measuring progress in hospital order treatment: Relationship between levels of security and C and R scores of the HCR-20. *International Journal of Forensic Mental Health*, 6, 113-121. doi:10.1080/14999013.2007.10471256
- Mulvey, E. P., Odgers, C., Skeem, J., Gardner, W., Schubert, C., & Lidz, C. (2006). Substance use and community violence: A test of the relation at the daily level. *Journal of Consulting and Clinical Psychology*, 74, 743-754. doi:10.1037/0022-006X.74.4.743
- Muthén, B., & Asparauhov, T. (2009). Growth mixture modeling: Analysis with non-gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G.

- Molenberghs (Eds.), *Handbook of Modern Statistical Methods: Longitudinal data analysis* (pp. 143-166). Boca Raton, FL: Chapman & Hall/CRC.
- Muthén, B., & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics, 30*, 27-58.
doi:10.3102/10769986030001027
- Nadesu, A. (2007). *Reconviction patterns of released prisoners: A 36-months follow-up analysis*. Wellington, NZ: Department of Corrections.
- Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed poisson model. *Criminology, 31*, 327-362. doi:10.1111/j.1745-9125.1993.tb01133.x
- Nonstad, K., & Webster, C. D. (2011). How to fail in the implementation of a risk assessment scheme or any other new procedure in your organization. *American Journal of Orthopsychiatry, 81*, 94-99. doi:10.1111/j.1939-0025.2010.01076.x
- Novaco, R. W., & Taylor, J. L. (2015). Reduction of assaultive behavior following anger treatment of forensic hospital patients with intellectual disabilities. *Behaviour Research and Therapy, 65*, 52-59. doi:10.1016/j.brat.2014.12.001
- Nunes, K. L., Pettersen, C., Hermann, C. A., Looman, J., & Spape, J. (2014). Does change on the MOLEST and RAPE scales predict sexual recidivism? *Sexual Abuse: A Journal of Research and Treatment*. Advance online publication.
doi:10.1177/1079063214540725

- Olver, M. E., & Wong, S. C. (2009). Therapeutic responses of psychopathic sexual offenders: Treatment attrition, therapeutic change, and long-term recidivism. *Journal of Consulting and Clinical Psychology, 77*, 328-336.
doi:10.1037/a0015001
- Olver, M. E., & Wong, S. C. P. (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
- Olver, M. E., Beggs Christofferson, S. M., Grace, R. C., & Wong, S. C. (2014). Incorporating change information into sexual offender risk assessments using the Violence Risk Scale-Sexual Offender version. *Sexual Abuse: A Journal of Research and Treatment, 26*, 472-499. doi:10.1177/1079063213502679
- Olver, M. E., Kingston, D. A., Nicholaichuk, T. P., & Wong, S. C. (2014). A psychometric examination of treatment change in a multisite sample of treated Canadian federal sexual offenders. *Law and Human Behavior, 38*, 544-559.
doi:10.1037/lhb0000086
- Olver, M. E., Wong, S. C. P., Nicholaichuk, T., & Gordon, A. (2007). The validity and reliability of the Violence Risk Scale-Sexual Offender version: Assessing sex offender risk and evaluating therapeutic change. *Psychological Assessment, 19*, 318-329. doi:10.1037/1040-3590.19.3.318
- O'Quigley, J., & Xu, R. (2001). Explained variation in proportional hazards regression. In J. Crowley (Ed.), *Handbook of statistics in clinical oncology* (pp. 397-410). New York: Marcel Dekker, Inc.

- O'Quigley, J., Xu, R., & Stare, J. (2005). Explained randomness in proportional hazards models. *Statistics in Medicine*, 24, 479-489. doi:10.1002/sim.1946
- Orrick, E. A. (2012). Criminal attitudes and reentry: An analysis of post-release from prison among serious violent offenders (Master's thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 3523601)
- Peterson-Badali, M., Skilling, T., & Haqanee, Z. (2014). Examining implementation of risk assessment in case management for youth in the justice system. *Criminal Justice and Behavior*, 42, 304-320. doi:10.1177/0093854814549595
- 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
- Philipse, M. W., Koeter, M. W., van der Staak, C. P., & van den Brink, W. (2006). Static and dynamic patient characteristics as predictors of criminal recidivism: A prospective study in a Dutch forensic psychiatric sample. *Law and Human Behavior*, 30, 309-327. doi:10.1007/s10979-006-9013-4
- Piquero, A. R. (2008). Taking stock of developmental trajectories of criminal activity over the life course. In A. Liberman (Ed.), *The long view of crime: A synthesis of longitudinal research* (pp. 23-78). New York: Springer.
- 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

- Potapov, S., Adler, W., & Schmid, M. (2015). *Estimators of prediction accuracy for time-to-event data*. Retrieved from <https://cran.r-project.org/web/packages/survAUC/survAUC.pdf>.
- Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Quantitative Applications in the Social Sciences: Latent growth curve modeling*. Thousand Oaks, CA: Sage.
- Quinsey, V. L., Book, A., & Skilling, T. A. (2004). A follow-up of deinstitutionalized men with intellectual disabilities and histories of antisocial behaviour. *Journal of Applied Research in Intellectual Disabilities*, 17, 243-253.
doi:10.1111/j.1468-3148.2004.00216.x
- Quinsey, V. L., Coleman, G., Jones, B., & Altrows, I. F. (1997). Proximal antecedents of eloping and reoffending among supervised mentally disordered offenders. *Journal of Interpersonal Violence*, 12, 794-813.
doi:10.1177/088626097012006002
- Quinsey, V. L., Jones, G. B., Book, A. S., & Barr, K. N. (2006). The dynamic prediction of antisocial behavior among forensic psychiatric patients: A prospective field study. *Journal of Interpersonal Violence*, 21, 1539-1565.
doi:10.1177/0886260506294238
- R Core Team (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>.

- Raynor, P. (2007). Risk and need assessment in British probation: The contribution of LSI-R. *Psychology, Crime & Law*, 13, 125-138.
doi:10.1080/10683160500337592
- Rogers, R. (2000). The uncritical acceptance of risk assessment in forensic practice. *Law and Human Behavior*, 24, 595-605. doi:10.1023/A:1005575113507
- 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.statcan.gc.ca>.
- Saha, P. (2009). *Time-dependent predictive accuracy: Extending binary classification accuracy methods for censored survival data* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3370556)
- Saha, P., Heagerty, P. J., & Zheng, Y. (n.d.). *Introduction to risksetROC: An R package for riskset-ROC (incident/dynamic) calculation*. Retrieved from <http://faculty.washington.edu/heagerty/Software/SurvROC/RisksetROC/risksetROCdiscuss.pdf>.
- Sampson, R. J., & Laub, J. H. (2005). A life-course view of the development of crime. *Annals of the American Academy of Political and Social Science*, 602, 12-45.
doi:10.1177/0002716205280075
- Savaya, R., Monnickendam, M., & Waysman, M. (2001). An assessment of the utilization of a computerized decision support system for youth probation officers. *Journal of Technology in Human Services*, 17, 1-14.
doi:10.1300/j017v17n04_01

- Scanlan, J., Fortune, C. -A., Polaschek, D., & Yesberg, J. (2015, June). *Women on community supervision sentences: How do initial and proximal DRAOR scores compare in predicting recidivism?* Paper presented to the North American Correctional and Criminal Justice Psychology Conference, Ottawa, ON.
- Schemper, M., & Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics*, 56, 249-255. doi:10.1111/j.0006-341X.2000.00249.x
- Schlager, M. D., & Pacheco, D. (2011). An examination of changes in LSI-R scores over time: Making the case for needs-based case management. *Criminal Justice and Behavior*, 38, 541-553. doi:10.1177/0093854811402300
- Schwalbe, C. S. (2007). Risk assessment for juvenile justice: A meta-analysis. *Law and Human Behavior*, 31, 449-462. doi:10.1007/s10979-006-9071-7
- Scoones, C. D., Willis, G. M., & Grace, R. C. (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
- Scott, K. L. (2004). Predictors of change among male batterers: Application of theories and review of empirical findings. *Trauma, Violence, & Abuse*, 5, 260-284. doi:10.1177/1524838003264339
- Serin, R. C. (2007) *The Dynamic Risk Assessment for Offender Re-entry (DRAOR)*. Unpublished user manual.
- 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. L., & Wilson, N. J. (2012) *The Dynamic Risk Assessment for Offender Re-entry (DRAOR)*. Unpublished user manual.

Serin, R. C. & Prell, L. (2012, March). *Pathways to crime desistance for probationers*.

Paper presented at American Psychology-Law Society Annual Conference, San Juan, Puerto Rico.

Shulman, E. P., Cauffman, E., Piquero, A. R., & Fagan, J. (2011). Moral disengagement among serious juvenile offenders: A longitudinal study of the relations between morally disengaged attitudes and offending. *Developmental Psychology*, 47, 1619-1632. doi:10.1037/a0025404

Simon, J. (2005). Reversal of fortune: The resurgence of individual risk assessment in criminal justice. *Annual Review of Law and Social Science*, 1, 397-421.

doi:10.1146/annurev.lawsocsci.1.041604.120006

Simons, R. L., & Barr, A. B. (2014). Shifting perspectives: Cognitive changes mediate the impact of romantic relationships on desistance from crime. *Justice Quarterly*, 31, 793-821. doi:10.1080/07418825.2012.704388

Simons, R. L., & Burt, C. H. (2011). Learning to be bad: Adverse social conditions, social schemas, and crime. *Criminology*, 49, 553-598. doi:10.1111/j.1745-9125.2011.00231.x

Simons, R. L., Burt, C. H., Barr, A. B., Lei, M. -K., & Stewart, E. (2014). Incorporating routine activities, activity spaces, and situational definitions into the social

schematic theory of crime. *Criminology*, 52, 655-687. doi:10.1111/1745-9125.12053

Simourd, D. J., Olver, M. E., & Brandenburg, B. (2015). Changing criminal attitudes among incarcerated offenders: Initial examination of a structured treatment program. *International Journal of Offender Therapy and Comparative Criminology*. Advance online publication. doi:10.1177/0306624X15579257

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Singh, J. P., Fazel, S., Gueorguieva, R., & Buchanan, A. (2014). Rates of violence in patients classified as high risk by structured risk assessment instruments. *The British Journal of Psychiatry*, 204, 180-187.

doi:10.1192/bjp.bp.113.131938

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science*, 20, 38-42.

doi:10.1177/0963721410397271

Skeem, J. L., Schubert, C., Odgers, C., Mulvey, E. P., Gardner, W., & Lidz, C. (2006). Psychiatric symptoms and community violence among high-risk patients: A test of the relationship at the weekly level. *Journal of Consulting and Clinical Psychology*, 74, 967-979. doi:10.1037/0022-006X.74.5.967

Smeth, A. (2013). *Evaluating risk assessments among sex offenders: A comparative analysis of static and dynamic factors* (Unpublished master's thesis). Carleton University, Ottawa, ON.

- Smeth, A., Serin, R. C., & Hanby, L. J. (2015, June). *Examining the predictive accuracy of the Dynamic Risk Assessment for Offender Re-entry (DRAOR) across sex offender subtypes*. Paper presented to the North American Correctional and Criminal Justice Psychology Conference, Ottawa, ON.
- 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 & Offenders*, 4, 148-169.
doi:10.1080/15564880802612581
- Studer, L. H., & Reddon, J. R. (1998). Treatment may change risk prediction for sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, 10, 175-181. doi:10.1023/A:1021361703167
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1, 1-26.
doi:10.1111/1529-1006.001
- Szasz, T. S. (1961). *The myth of mental illness: Foundations of a theory of personal conduct*. New York: Harper Perennial.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn and Bacon.
- Tamatea, A., & Wilson, N. (2009). *Dynamic Risk Assessment for Offender Re-entry (DRAOR): A pilot study*. New Zealand: Department of Corrections.
- Taxman, F. S., & Caudy, M. S. (2015). Risk tells us who, but not what or how: Empirical assessment of the complexity of criminogenic needs to inform

correctional programming. *Criminology & Public Policy*, 14, 71-103.

doi:10.1111/1745-9133.12116

Taxman, F. S., Caudy, M. S., & Pattavina, A. (2013). Risk-Need-Responsivity (RNR):

Leading towards another generation of the model. In F. S. Taxman & A.

Pattavina (Eds.), *Simulation strategies to reduce recidivism: Risk need
responsivity (RNR) modeling for the criminal justice system* (pp. 285-307).

New York: Springer. doi:10.1007/978-1-4614-6188-3_11

Therneau, T., & Crowson, C. (2015). *Using time dependent covariates and time*

dependent coefficients in the Cox model. Retrieved from [https://cran.r-
project.org/web/packages/survival/vignettes/timedep.pdf](https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf).

Thornton, D. (2002). Constructing and testing a framework for dynamic risk

assessment. *Sexual Abuse: A Journal of Research and Treatment*, 14, 139-153.

doi:10.1177/107906320201400205

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 der Knaap, L. M., Alberda, D. L., Oosterveld, P., & Born, M. P. (2012). The

predictive validity of criminogenic needs for male and female offenders:
Comparing the relative impact of needs in predicting recidivism. *Law and
Human Behavior*, 36, 413-422. doi:10.1037/h0093932

van der Laan, A. M., Blom, M., & Kleemans, E. R. (2009). Exploring long-term and

short-term risk factors for serious delinquency. *European Journal of
Criminology*, 6, 419-438. doi:10.1177/1477370809337882

- Vaswani, N., & Merone, L. (2014). Are there risks with risk assessment? A study of the predictive accuracy of the Youth Level of Service-Case Management Inventory with young offenders in Scotland. *British Journal of Social Work, 44*, 2163-2181. doi:10.1093/bjsw/bct059
- Viglione, J., Rudes, D. S., & Taxman, F. S. (2014). Misalignment in supervision: Implementing risk/needs assessment instruments in probation. *Criminal Justice and Behavior, 42*, 263-285. doi:10.1177/0093854814548447
- Vincent, G. M., Guy, L. S., Fusco, S. L., & Gershenson, B. G. (2012). Field reliability of the SAVRY with juvenile probation officers: Implications for training. *Law and Human Behavior, 36*, 225-236. doi:10.1037/h0093974
- Vincent, G. M., Paiva-Salisbury, M. L., Cook, N. E., Guy, L. S., & Perrault, R. T. (2012). Impact of risk/needs assessment on juvenile probation officers' decision making: Importance of implementation. *Psychology, Public Policy, and Law, 18*, 549-576. doi:10.1037/a0027186
- Venables, W. N., Smith, D. M., & the R Core Team. (2014). *An introduction to R*. Retrieved from <http://www.R-project.org/>.
- Vose, B. (2008). *Assessing the predictive validity of the Level of Service Inventory-Revised: Recidivism among Iowa parolees and probationers* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3323919)
- Wakeling, H. C., 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

Walters, G. D. (2002). Current and historical content scales for the Psychological Inventory of Criminal Thinking Styles (PICTS). *Legal and Criminological Psychology*, 7, 73-86. doi:10.1348/135532502168397

Walters, G. D. (2006). Appraising, researching, and conceptualizing criminal thinking: A personal view. *Criminal Behaviour and Mental Health*, 16, 87-99. doi:10.1002/cbm.50

Walters, G. D. (2015). Short-term goals and physically hedonistic values as mediators of the past-crime-future-crime relationship. *Legal and Criminological Psychology*, 20, 81-95. doi:10.1111/lcrp.12014

Ward, T. (2015). Detection of dynamic risk factors and correctional practice. *Criminology & Public Policy*, 14, 105-111. doi:10.1111/1745-9133.12115

Ward, T., & Beech, A. R. (2014). Dynamic risk factors: A theoretical dead-end? *Psychology, Crime & Law*, 21, 100-113. doi:10.1080/1068316x.2014.917854

Ward, T., & Gannon, T. A. (2006). Rehabilitation, etiology, and self-regulation: The comprehensive good lives model of treatment for sexual offenders. *Aggression and Violent Behavior*, 11, 77-94. doi:10.1016/j.avb.2005.06.001

Ward, T., & Marshall, W. L. (2004). Good lives, aetiology and the rehabilitation of sex offenders: A bridging theory. *Journal of Sexual Aggression*, 10, 153-169. doi:10.1080/13552600412331290102

- Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). *Reliability coefficients and generalizability theory*. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics* (Vol. 26, pp. 81-124). Amsterdam: Elsevier B. V.
- Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20. Assessing the risk of violence. Version 2*. Burnaby, BC: Simon Fraser University and Forensic Psychiatric Services Commission of British Columbia.
- Webster, C. D., Martin, M. -L., Brink, J., Nicholls, T. L., & Desmarais, S. L. (2009). *Manual for the Short-Term Assessment of Risk and Treatability (START) (Version 1.1)*. Coquitlam, BC: Forensic Psychiatric Services Commission.
- Wikström, P. -O. H., Ceccato, V., Hardie, B., & Treiber, K. (2010). Activity fields and the dynamics of crime: Advancing knowledge about the role of the environment in crime causation. *Journal of Quantitative Criminology*, 26, 55-87. doi:10.1007/s10940-009-9083-9
- Wikström, P. -O. H., & Treiber, K. (2009). What drives persistent offending? The neglected and unexplored role of the social environment. In J. Savage (Ed.), *The development of persistent criminality* (pp. 389-420). New York, NY, US: Oxford University Press. doi:10.1093/acprof:oso/9780195310313.003.0019
- Willett, J. B., & Bub, K. L. (2004). Latent growth curve analysis. In D. Rindskoph (Ed.), *Encyclopedia of statistics in the behavioral sciences*. Sussex, UK: John Wiley and Sons.
- Willett, J. B., & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research*, 61(4), 407-450.

- Willett, J. B., & Singer, J. D. (1993). Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology, 61*, 952-965. doi:10.1037/0022-006X.61.6.952
- Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10*(2), 395-426.
- Woessner, G., & Schwedler, A. (2014). Correctional treatment of sexual and violent offenders: Therapeutic change, prison climate, and recidivism. *Criminal Justice and Behavior, 41*, 862-879. doi:10.1177/0093854813520544
- Wright, K. N., Clear, T. R., & Dickson, P. (1984). Universal applicability of probation risk-assessment instruments. *Criminology, 22*, 113-134. doi:10.1111/j.1745-9125.1984.tb00291.x
- 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., Wong, S. C., & 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
- Yang, S., & Mulvey, E. P. (2012). Violence risk: Re-defining variables from the first-person perspective. *Aggression and Violent Behavior, 17*, 198-207. doi:10.1016/j.avb.2012.02.001

- Yang, S., Mulvey, E. P., Loughran, T. A., & Hanusa, B. H. (2012). Psychiatric symptoms and alcohol use in community violence by persons with a psychotic disorder or depression. *Psychiatric Services*, 63, 262-269.
doi:10.1176/appi.ps.201100354
- Yeaton, W. H., & Sechrest, L. (1981). Meaningful measures of effect. *Journal of Consulting and Clinical Psychology*, 49, 766-767. doi:10.1037/0022-006X.49.5.766
- Yesberg, J. A., & Polaschek, D. L. (2014). 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
- Yesberg, J. A., Scanlan, J. M., Hanby, L. J., Serin, R. C., & Polaschek, D. L. L. (2015). Predicting women's recidivism: Validating a dynamic community-based 'gender-neutral' tool. *Probation Journal*, 62, 33-48.
doi:10.1177/0264550514562851
- Zamble, E., & Quinsey, V. L. (1997). *The criminal recidivism process*. New York: Cambridge University Press.

Appendix

The purpose of this appendix is to outline how to calculate effect size estimates for Cox regression models within the statistics program, *R* (see www.r-project.org; *R* Core Team, 2015). This guide will likely be most useful for readers who (a) have some basic familiarity working with *R*, (b) understand the basic concepts and computations underpinning the effect size measures being calculated, (c) are familiar with the results from the Cox models run on their dataset, and (d) have specific interest in obtaining the effect size estimates described in this dissertation. Readers interested in gaining familiarity with the *R* program will likely find it useful to consult Venables, Smith, and the *R* Core Team (2014). Readers looking to gain familiarity with the effect sizes will find it useful to consult the references cited on pages 90-99 of this dissertation. There are additional effect size options available for Cox models, and interested readers will find guidance for running these other options in the *R* program by consulting Potapov, Adler, and Schnid (2015).

The *R* open-source program is highly appealing for its flexibility and the vast number of options it provides for different types of analyses, as well as numerous options within any particular analysis. This appendix is purposefully reductive with the aim to focus only on obtaining the two effect size indices most recommended for Cox regression with time-varying predictors. This guide draws heavily from (and largely re-packages information found within) the following sources: Heagerty & Saha-Chaudhuri (2015); Potapov, Adler, and Schnid (2015); Saha, Heagerty, and Zheng (n.d.); and Therneau and Crowson (2015).

Preparation for Analysis

Prior to analysis, data must be stored in a format that can be easily read by *R*. Many *R* guides recommend converting the data into a simple text file with a predetermined format, such as comma-separated values (.csv). Data can be easily stored in a .csv format through the *Save as...* function within Excel or SPSS.

Data with time-varying predictors must be stored in a person-period format. See Table X for a hypothetical example based on the dataset described in this dissertation.

Table X. Hypothetical data

ID	entry	exit	status	rocroi	stable	acute	protect
1	0	6	0	.59	6.98	7.22	1.99
1	6	13	0	.59	6.80	7.20	2.22
1	13	20	0	.59	6.72	7.18	2.40
1	20	27	1	.59	6.60	7.15	2.73

Table X displays data for a single participant (ID = 1) who survived without reoffending in the community for three weeks, but then re-offended after his fourth week (status = 1). Note that *R* requires that datasets with time-varying predictors use two columns to represent the variable indicating time. In this example, each time period represents a week, but the units are expressed in days. Thus, a week-long time period is represented by a 7-day span between day of entry and day of exit. This format follows the [start, stop] notation, which indicates that the time-varying score on Day 6 would be gathered from the row associated with the (0, 6] interval. Thus, in this example, the *stable* score associated with Day 6 is 6.98, rather than 6.80. Note that this format is equivalent to treating a single participant's data as a series of left- and right-truncated time periods.

The remaining columns in Table X display time-invariant and time-varying predictors for Participant 1. The *R* program can accommodate datasets with missing values, but rows that will not enter into the analysis due to missing data will create difficulties when running some of the syntax displayed later in this appendix. In order to follow the examples below exactly as shown, it would be best to remove rows with missing data, prior to reading the data into *R*.

Once the .csv file is completed, the following commands within the *R* program will make the data accessible for analysis. In this example, the file is named `draor.csv`.

```
draor<-read.csv("C:/.../draor.csv")

attach(draor)
```

Specialized analyses are not automatically available within the *R* base program, and supplemental *R packages* may need to be downloaded (for guidance, see Venables et al., 2014). To conduct the analyses in this guide, the following packages and commands must be activated.

```
library(survival)

library(survAUC)

library(risksetROC)
```

Calculating Xu & O'Quigley's (1999) R^2 -analogous Coefficient

```
train<-draor[1:45969,]
test<-draor[45970:91945,]
```

Following the example and recommendation of Potapov et al. (2015), this first step involves dividing the dataset into two approximately equally sized independent samples. In the *draor* dataset example, the first 45,969 rows represent approximately half of the participants' time periods. This first half of the dataset,

labeled `train`, will be used to develop the Cox regression model. The second half of the dataset, labeled `test`, will be used to test the fit of the model when calculating Xu and O'Quigley's R^2 statistic.

```
model<-coxph(Surv(entry,exit,status)~stable,data=train)
```

This command fits a Cox regression survival model with a time-varying predictor. Note that the command requires three elements from the dataset to create the outcome variable: time periods are designated by their `entry` times, `exit` times, and recidivism `status`. Using Cox regression, this outcome is regressed on the time-varying predictor variable, `stable`. The final element of the command (`data=train`) limits the model building only to the rows designated as the training portion of the dataset.

```
f0<-rep(0,nrow(test))
```

```
f1<-predict(model,newdata=test)
```

```
Surv.res<-Surv(test$entry,test$exit,test$status)
```

These three commands set up the remaining elements needed to calculate the R^2 statistic. The first command creates `f0`, a column (vector) of predictor values generated from a model that omits the time-varying predictor. In other words, this command generates expected values in a baseline (null) model. The `nrow(test)` command ensures that the vector holds one value for each time period (i.e., each row) listed in the `test` portion of the dataset.

The second command creates `f1`, a column (vector) of predictor values generated by using scores listed in the `test` portion of the dataset and fitting them to the Cox model developed from the `train` portion of the dataset. The resulting vector lists values that represent the linear combination of mean-corrected predictor values multiplied by the beta coefficient from the model (i.e., *stable score* – *mean* * *Beta*). Note that both the mean and beta values are drawn from the results of `train` portion of the dataset, but are then calculated into predictor values using data drawn from the `test` portion of the dataset. Thus, this vector represents the expected predictor values for each time period, given the Cox model.

The final command generates an object that represents *time period* + *status*. In other words, all [`entry`, `exit`] time periods, one for each row in the `test` portion of the dataset, are listed and weighted by their status (i.e., 0 = no event, 1 = event).

```
XO(Surv.res,f1,f0)
```

This final command generates Xu and O'Quigley's R^2 statistic (`XO`) by entering the three vectors described above: the time periods weighted by status (`Surv.res`), the expected values obtained from the Cox regression model (`f1`), and the expected

values obtained from the baseline, no-predictor model (f_0). As discussed in the main text, calculations rely on a ratio that compares observed vs. expected discrepancies when the predictor is included versus excluded, and the resulting value can be interpreted as the proportion of randomness explained by the model.

Calculating and Graphing Heagerty & Zheng's (2005) Time-Dependent AUC

```
newmodel<-coxph(Surv(entry,exit,status)~stable,data=draor)
```

This first command is highly similar to the second step shown in the section above. As before, this command fits a Cox regression survival model with a time-varying predictor. However, the final element of this command indicates that the model should be run using all rows (time periods) contained in the `draor` dataset.

```
eta<-newmodel$linear.predictor
```

This command creates a vector of linear predictor values, one value for each time period within the dataset. As previously shown above (see f_1 vector), this command obtains expected values by fitting observed values into the Cox model. Within the next command, these expected values are used to test how well the expected values under the Cox model are able to distinguish between cases and controls (recidivists and non-recidivists). In this way, the analysis tests the prediction accuracy of the model.

```
AUC.sch<-
risksetAUC(Stime=exit,entry=entry,status=status,marker=
eta,method="Schoenfeld",span=0.2,tmax=364,col="darkblue",lt
y=2,
main="DRAOR Stable Time-Varying Scores: AUC
Plot",data=draor)
```

```
AUC.sch
```

These commands generate the two desired outputs: (a) a graph of time-specific AUC statistics plotted across the follow-up time, and (b) the weighted average of time-specific AUC statistics, the global summary statistic (*c*-index). The first three elements within the parentheses of the first command are familiar from commands described previously. In particular, these indicate which variables should be designated as the entry time, exit time, and recidivism status. The next element designates the vector of linear predictors generated from the Cox model (`eta`) as the marker of interest for testing sensitivity and specificity.

The following two elements (`method` and `span`) indicate how the function will use regression-smoothing techniques, which allow the analysis to meet the assumption that hazard ratios vary smoothly across time. The method chosen in this example employs a technique where the Schoenfeld residuals are smoothed. Smoothing

allows the function to estimate the time-varying effect of the predictor values across the follow-up period. Related to this, the specification of `span` determines the size of the local neighborhood used within the smoothing technique (i.e., the amount of observations surrounding an event time that are considered). The Schoenfeld method is appropriate when time-varying predictor variables are employed in the analysis, and Heagerty and Zheng (2005) demonstrated that this method resulted in approximately unbiased estimates, except in the most extreme values of time, where a small amount of bias was observed. The syntax above can be adjusted in order to compare different choices of smoothing techniques and span sizes (see Saha et al., n.d.).

Next, the `tmax` element indicates the maximum time period that should be considered in the calculation of time-specific AUC statistics. In this example, follow-up time is truncated to one year, so `tmax` equals 364 days. If follow-up time is not truncated, the `tmax` value should match the final failure time observed within the dataset.

The remaining elements in the command designate graphing functions. Including these elements indicates that a plot should be created that graphs the time-specific AUC statistics across the follow-up time. The above example specifies that the plotted line will be a dashed, dark blue line. The above example also gives the graph a title.