Dynamic Risk Factors and Treatment Change: Exploring the Mechanisms of Sexual Offending Onset and Desistance

A thesis submitted in fulfillment of the
requirements for a Doctorate of
Philosophy in Psychology at the
University of Canterbury

by

Jacinta R. Cording

October 2018

Abstract

The concept of change and the ability for individuals to reduce their level of risk through targeted intervention is a core feature of current rehabilitation frameworks used with individuals who have engaged in sexually harmful behaviours. However, despite the large emphasis placed on dynamic risk factors and the acknowledgement of the ability for individuals to make prosocial change, relatively little attention is given to furthering our understanding of how dynamic risk meaningfully relates to the aetiology of sexual offending, and to developing theories of the mechanisms underlying the change process. The current thesis therefore addressed this gap in the literature by investigating the factors and characteristics that play a causal role in sexual offending behaviour, and by exploring the underlying nature of offender change to help inform ongoing theory generation in this area.

Study One began with a validation of an influential theory of the aetiology of sexual offending, Ward and Siegert's (2002) Pathways Model of Child Sexual Offending. The study used pre-treatment scores on a psychometric battery completed by 1,134 male sexual offenders against children to conduct a Latent Profile Analysis (LPA), which is a statistical technique used to identify meaningful latent classes of individuals within a given sample. Results suggested that the sample was best captured by five classes of individuals that mapped closely to the five hypothesised pathways in the Pathways model, with a few notable exceptions. Overall, the study provided tentative support for the Pathways Model and its proposed mechanisms and aetiological pathways, provided guidance for potential amendments to the model, and highlighted the heterogeneity in the offender population and causes of offending.

Studies Two, Three and Four then went on to explore the nature and characteristics of sexual offender change, with the aim of providing valuable insights to inform ongoing theory generation regarding the mechanisms and nature of change. Study Two provided the first known study to explore whether sexual offender treatment change is best conceptualised as categorical or dimensional, by using standardised residual change scores from 346 male sexual offenders against children to conduct a taxometric analysis of change. Results from the analysis suggested that offender change is best conceptualised as a categorical construct; that is, that differences in treatment change between individuals are best understood as differences in the types of change made, rather than simply the amount of change made.

Study Three explored the implications of Study Two's findings further by attempting to identify what these change categories might look like. The study used standardised residual change scores from 1,170 sexual offenders against children to conduct an LPA, which found that three classes provided a best fit for the data. These classes represented individuals who had made Poor Change, Moderate Change, or Good Change over the course of treatment, with individuals in the Good Change group reoffending at significantly lower rates than individuals in the other two groups. The study suggested that meaningful distinctions can be made between different kinds of change made over the course of treatment, but did not provide much information regarding the mechanisms underlying these change patterns.

Study Four therefore provided a further investigation of these groups, by assessing the pre-treatment needs, static risk, and historical or demographic characteristics associated with each change group. Results indicated that there were no significant differences in static risk or most historic or demographic factors between groups, but that individuals in the Good Change group showed significantly lower rates of pre-treatment needs than individuals in the Poor Change group (with individuals in the Moderate Change group falling in between). This suggested that perhaps individuals in the Good Change group were already on a pathway to desistance prior to entering treatment.

Together, the results from this thesis suggest that internal factors, such as motivation to change and cognitive transformation (i.e. the adoption of a pro-social identity), may be key mechanisms underlying change demonstrated by offenders. They also highlight the heterogeneity of pathways into sexual offending and related treatment needs, and add to a growing body of research supporting the need for individualised assessment and intervention that focusses on the promotion of prosocial identity and skill acquisition.

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Large portions of Chapter One and portions of Chapter Two have been extracted from the following published paper:

Cording, J. R., Beggs Christofferson, S. M., & Grace, R. C. (2016). Challenges for the theory and application of dynamic risk factors. Psychology, Crime and Law, 22(1-2), 84-103.

Please detail the nature and extent (%) of contribution by the candidate:

The candidate was the lead author of this publication, was heavily involved in developing the topics discussed in the paper, and contributed 80-90% of the writing in the paper.

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Sarah Christofferson* Signature: *SM Christofferson* Date: *30/10/18*

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Some small portions of Chapter One and Two have been extracted from the following book chapter that is currently in press:

Cording, J.R., Ward, T., & Beggs Christofferson, S.M. (in press). Risk prediction and sex offending. In W. T. O'Donohue & D. Bromberg (Eds.). Sexually Violent Predators: A clinical science handbook. New York: Springer.

Please detail the nature and extent (%) of contribution by the candidate:

The candidate was the lead author of this book chapter, was heavily involved in developing the topics discussed in the paper, and contributed 80-90% of the writing in the paper.

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Sarah Christofferson* Signature: *SM Christofferson* Date: *30/10/18*

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

All of Study One has been accepted for publication, subject to minor revisions, as the following:

Cording, J.R., Grace, R.C., & Beggs Christofferson, S.M. (accepted subject to minor revisions). Addressing the cause: An empirical test of Ward and Siegert's Pathways Model of child sexual abuse. Sexual Abuse.

Please detail the nature and extent (%) of contribution by the candidate:

The candidate was the lead author of this publication and conducted all data collection, cleaning and analysis outlined in the paper. She also contributed 95% of the writing in the paper.

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Sarah Christofferson* Signature: *SM Christofferson* Date: *30/10/18*

Acknowledgements

I am grateful to a large number of people who have supported me throughout my Doctoral journey and who have each contributed in their own way to my final thesis. A huge thank you to my supervisors Randy Grace and Sarah Christofferson, for their unwavering confidence and support in me and my research. The mentorship that they provided and the generosity in sharing their time and knowledge was invaluable for my research.

A special mention to the members of the University of Canterbury and Victoria University of Wellington forensic psychology labs, who generously provided their thoughts on my early research, and who had such a large impact on my growth as a researcher and scholar. A very special gratitude goes to Devon Polaschek for her willingness to introduce an academic orphan to her exceptional group of students and collaborators, and her acceptance of ongoing third-wheeling.

I am grateful to staff at the Department of Corrections (including the Kia Marama Special Treatment Unit) who provided access to the data for my research, space to work, and invaluable assistance throughout data collection. I would also like to thank Mark Olver for his assistance with VRS-SO scoring and responsiveness to requests for clarification or advice.

Lastly, I would like to acknowledge the unwavering support of my friends and family throughout this whole process. My husband Rowan, for the cooked meals, cups of tea, and for being long-suffering sounding board. My parents, Shelley and Richard, and my grandparents, Betty and Barry, for providing me with my love of learning and unconditional encouragement. My siblings, for always keeping me grounded and making me laugh.

Author Note

Portions of the data reported here have previously been presented at the annual meetings of the New Zealand Psychological Society, the Association for the Treatment of Sexual Abusers, the International Association for the Treatment of Sexual Offenders, and the Department of Corrections' National Training Event.

Some portions of this thesis have also been published, are in press, or have been accepted for publication pending revisions. These portions of the thesis have been declared in the forms inserted above.

Note that UK spelling is used in this thesis except for Study One, in which US spelling has been used due to the editorial requirements of the journal it was submitted to. Please also note that as the studies have been prepared for separate submission to journals, there are some small repetitions in content between some portions of the thesis.

Table of Contents

List of Figures

List of Tables

Chapter One: Exploring Dynamic Risk Factors

The accurate prediction of reoffending risk and identification of risk-relevant needs (i.e. criminogenic needs) are key components of modern efforts to reduce community harm and justify the continued restriction of personal liberty inherent in the custodial management of offenders. The assessment of risk and need plays an important role in decision-making at all stages of offender management and intervention, including for sentencing, custodial and living arrangements, and decisions regarding early release or indefinite detainment. Supporting these risk and need assessments is a large body of literature that provides evidence for the link between certain individual and environmental characteristics, and increased rates of recidivism (includng sexual offending; Hanson & Bussiere, 1998).

In general, these offence-related characteristics can be categorised as either historic factors that are changeable (e.g., number of previous victims), or psychological or environmental factors that do have the capacity to change (e.g., ability to regulate emotions). Unchangeable factors have historically been referred to as static risk factors, whereas those that are able to change (at least theoretically) are referred to as dynamic risk factors (Bonta & Andrews, 2016). Because of the utility of dynamic risk factors in informing treatment targets, and because of mounting evidence for the superior predictive ability of risk assessment tools that incorporate dynamic risk factors compared with purely static risk assessment tools (Hanson & Morton-Bourgon, 2009), dynamic risk is increasingly becoming a primary focus of sexual offending research and assessment.

At the core of the assessment of dynamic risk and its role in informing offender management and treatment is the concept that an individual's level of risk and need can change; that is, that past behaviour is not always the best predictor of future behaviour. Indeed, previous research has confirmed this concept by providing evidence that offenders are able to display changes in dynamic risk over time (Nunes, Babchishin, & Cortoni, 2011; Olver, Nicholaichuk, & Wong, 2014). That said, given the importance of this concept of change to both risk management and modern offender treatment programmes, it is somewhat surprising that very little research to date has focussed on understanding the nature of this change, and confirming the assumption that pro-social change in criminogenic needs is predictive of reduced recidivism (Helmus, 2018).

Although a large amount of research has focussed on the predictive validity of dynamic risk factors at pre- and post-treatment, this research is primarily driven by exploring statistical relationships between dynamic risk and recidivism, rather than generating theories about the causes of offender change and the constructs underlying the criminogenic needs being measured (Heffernan & Ward, 2015). This focus on the predictive ability of dynamic risk factors rather than the measurement of change across meaningful offence-related characteristics is perhaps why a number of previous studies have failed to find a significant link between changes in dynamic risk factors and comparable changes in recidivism rates (e.g., Barnett, Wakeling, Mandeville-Norden, & Rakestrow, 2013; Olver, Kingston, Nicholaichuk, & Wong, 2014). Current theory regarding the causes of sexual offending, desistance from this offending, and the mechanisms of sexual offender change is therefore an area that is still being explored in the literature.

The current thesis focussed on this gap in the literature by investigating the factors and characteristics that play a causal role in sexual offending behaviour, and by exploring the underlying nature of offender change to help inform ongoing theory generation in this area. Study One began by testing the validity of an existing causal theory of sexual offending, Ward and Siegert's (2002) Pathways Model of Child Sexual Offending. This theory represents a notable attempt in the literature to explain the underlying mechanisms driving offender behaviour, and validating this theory experimentally provides important information regarding the aetiology of sexual offending (and therefore identifies potential areas of interest for change or desistance research).

Studies Two, Three and Four then went on to explore the nature and characteristics of sexual offender change, with the aim of providing valuable insights to inform ongoing theory generation regarding the mechanisms and nature of change. Study Two began at the core of treatment change by exploring whether treatment change is better characterised as a dimensional or categorical construct using taxometric analysis; although treatment change is typically treated as dimensional in the existing literature, this assumption has not been tested up until this point. After finding that treatment change is best characterised as categorical in nature, Study Three explored this finding further by using latent class analysis to identify these categories of change and explore the patterns of change associated with these categories (including their link with recidivism). Study Four extended Studies Two and Three further by exploring the characteristics associated with change group membership and the implications of these characteristics for inferring potential explanatory theories regarding treatment change, and predicting how individuals will respond to treatment.

Dynamic Risk Factors and Sexual Offending

Although dynamic factors have received overall less attention than static factors in research on risk assessment, they remain highly promising not only with regards to the accurate prediction of recidivism, but also in identifying useful and effective treatment targets for offenders. The concept of dynamic risk factors and their ability to change over time is also central to the theory underlying modern sexual offending treatment and measurement of offender change.

Given the central importance of dynamic risk factors to the research presented in this thesis, the remainder of this chapter focusses on providing an overview of current knowledge of dynamic risk factors and their connection to sexual offending. A large amount of previous literature has discussed the utility of dynamic risk factors in the assessment of risk and identification of treatment targets, however focussed critique of the current conceptualisation and use of dynamic risk factors is comparatively rare (Mann, Hanson, & Thornton, 2010; Ward & Beech, 2014; Ward & Fortune, 2016). For this reason, this chapter focusses greater attention on critically exploring dynamic risk factors and how they are currently used in practice. This includes an exploration of some of the key challenges to the theory and utility of dynamic risk factors in explaining offending behaviour and informing treatment targets. The discussion begins with a brief overview of the context in which dynamic risk factors began to be applied to offender assessment, and is then followed by a discussion of some of the primary challenges to the concept and application of dynamic risk factors, including construct validity, methods used to identify dynamic risk factors, and how dynamic risk factors are being applied in practice.

A Brief History of Risk Assessment

Although consideration of the potential for further offending has always been one of the key considerations in offender management and treatment decisions, it is only relatively recently that specific tools have been developed for the quantifiable prediction of specific types of antisocial behaviour. The development of risk assessment approaches has previously been described as having progressed in four distinct generations (Andrews, Bonta, & Wormith, 2006). Although this was initially outlined in relation to the assessment of risk for general offenders, risk assessment approaches for the prediction of sexual offending has largely followed the same trajectory.

First-generation: Unstructured clinical judgement

First generation risk assessments refer to largely unstructured professional judgement, in which the clinician or other decision-maker forms a subjective opinion on risk level, based on their knowledge and experience. This approach to sex offending risk assessment is now discouraged as research has shown that unstructured judgement typically has poor predictive accuracy (Hanson & Morton-Bourgon, 2009) and is often biased towards over-estimating risk (Craig, Browne, Stringer, & Beech, 2004). This bias is suggested to be due to a failure to consider the relatively low base rate of sexual recidivism in judgements of risk. In a meta-analysis including 28,757 convicted sexual offenders from 100 samples, the observed rate of sexual recidivism was 11.5% over an average follow-up period of just under 6 years (Hanson & Morton-Bourgon, 2009). These base rates do not substantially increase even given a longer follow-up period; one study reported sexual recidivism rates of 26% for adult sexual offending and 32% for child sexual offending over a follow-up period of 25 years (Prentky, Lee, Knight, & Cerce, 1997), and a more recent study reported sexual recidivism rates of 11.2% for child sexual offending and 13.5% for adult sexual offending with an average follow-up of 15 years (Vess & Skelton, 2010). Low base rates of reoffending make over-estimation of risk more likely, in part leading to unnecessarily restrictive management and supervision decisions, and an over-allocation of treatment resources. However, although base rates of official sexual recidivism are relatively low, it is important to note that official recidivism rates are likely to under-represent the true rates of recidivism. This is because of the likelihood of undetected offences that are not captured in official records, and because of the impacts of judicial processes (such as plea-bargaining) that can obscure the true nature of the crime committed. This under-representation is something that should be considered when assessing the implications of official recidivism rates.

Second-generation: Structured actuarial tools

The next step in the progression of risk assessment approaches was the development of structured actuarial tools (Andrews et al., 2006). These tools typically include a pre-determined set of risk factors (i.e. characteristics shown to bear a statistical link with recidivism) that are rated according to a standardized scoring framework. Ratings of risk are then combined to derive a total risk score and/or risk band that corresponds to empirically-derived estimates of risk. This approach to risk assessment therefore provides a way to quantify expected recidivism rates that avoids issues with the subjectivity and bias inherent in unstructured clinical judgement. Structured actuarial tools were initially comprised of static risk factors only i.e. those not able to be changed through treatment, such as previous sexual offences or age at first sexual offence. Some key static risk factors for sexual offending include prior sexual offending, having male or stranger victims, and a history of treatment dropout (Hanson & Bussiere, 1998). Tools comprising static factors only include the Static-99 (Hanson & Thornton, 1999), the Sex Offender Risk Appraisal Guide (SORAG; Quinsey, Harris, Rice, & Cormier, 1998), and the Rapid Risk Assessment for Sexual Offense Recidivism (RRASOR; Hanson, 1997).

The increase in predictive validity provided by second generation tools was demonstrated by Hanson and Morton-Bourgon's (2009) meta-analysis, in which empirical actuarial tools were found to be more accurate in predicting sexual recidivism than unstructured professional judgement (d = 0.67 and 0.42, respectively). These effect sizes are indicative of the size of the mean difference in scores between the recidivists and the non-recidivists on a given measure, and can be interpreted as a medium effect for empirical actuarial tools and a small effect for unstructured professional judgement. Predictive accuracy can also be assessed using Area under the Curve (AUC) values. AUC values represent the probability that a randomly-selected recidivist will have a higher score on a given measure than a randomly-selected non-recidivist; a score of 0.5 means that the measure does no better than chance at predicting recidivism, whereas a score of 1 means that a measure perfectly predicts recidivism. The AUC value therefore gives an indication of the accuracy of a measure in terms of its rate of true positives versus false positives. Effect sizes of *d* = 0.67 and 0.42 convert to AUC values of 0.68 and 0.62, respectively. This means that in Hanson and Morton-Bourgon's study, empirical actuarial tools were found to have a 68% probability of assigning a recidivist a higher score than a non-recidivist, whereas unstructured clinical judgement had a 62% probability of doing the same.

In addition to improved predictive accuracy, risk assessments based on static risk factors are easily administered and scored, cost-effective, and enable the efficient screening of large numbers of individuals at a time. They can therefore be utilised relatively easy to inform sentencing and parole decisions (which must be undertaken for large numbers of offenders each year), and they can also contribute to the efficient allocation of individuals to appropriate treatment options or levels of intervention.

One of the major issues with risk assessments based on static measures, however, is their lack of ability to identify potential treatment targets, and their failure to take into account any of the environmental or situational factors that may influence offending. Static risk factors are often based on demographic or historical factors, and are therefore unchangeable, and are therefore unable to reflect changes in risk that occur due to variations in situations or external influences over time, including treatment. It is also important to note that while static factors may be useful in predicting long-term recidivism rates for an aggregate of offenders, because static factors are poor indicators of change, they are of little help to predicting *when* an offender will reoffend (Hanson & Harris, 2001). For these reasons, and because of a growing body of literature demonstrating the empirical validity of dynamic factors in predicting risk (e.g., Hanson & Harris, 2000; Hudson, Wales, Bakker, & Ward, 2002), there was a subsequent shift to the inclusion of dynamic factors in risk assessment.

Third-generation: Risk/Need assessments

Andrews and colleagues (2006) define third generation risk assessments as being distinct from second generation tools in their systematic and objective consideration of individual needs, typically through the incorporation of dynamic risk factors. Dynamic risk factors are factors linked with offending that are amenable to change; key dynamic risk factors for sexual offending include substance abuse problems, pro-offending attitudes, and deviant sexual interest (Hanson & Harris, 2000). This provides an advantage over the second generation tools in that assessments are theoretically able to capture changes in risk over time or in response to direct interventions, and they are also able to be used to inform the selection of appropriate treatment targets. Examples of commonly-used third-generation risk assessments for sexual offending include the STABLE 2007 and the ACUTE 2007 (Hanson, Harris, Scott, & Helmus, 2007), and the Violence Risk Scale - Sexual Offense Version (VRS-SO; Wong, Olver, Nicholaichuk, & Gordon, 2003).

The sensitivity of third-generation measures to changes in risk-related factors over time suggests that they may be able to improve upon the predictive accuracy of static risk assessments, particularly for individuals who have completed treatment or for whom circumstances have substantially changed over time. Indeed, a number of studies have been able to show that assessments incorporating dynamic risk factors are able to provide incremental predictive validity beyond assessments incorporating static factors alone (Allan, Grace, Rutherford, & Hudson, 2007; Beggs & Grace, 2010; Craissati & Beech, 2003; Hanson et al., 2007). The size of this difference varies across studies and dynamic risk measures, however one study found an increase of 0.08 in AUC value when using post-treatment dynamic risk assessment as opposed to static risk assessments (Beggs & Grace, 2010). They found that the assessment incorporating dynamic risk measures gave a higher risk score to recidivists than non-recidivists 80% of the time, as opposed to 72% of the time for the static risk measure. This increase in predictive accuracy lends support to the idea that dynamic risk factors are tapping into a distinct facet of risk that is not being captured by historical or other static factors alone.

Structured professional judgement

Before moving on to fourth-generation risk assessment tools, it is important to briefly discuss risk assessment approaches based on structured professional judgement (SPJ). Approaches based on SPJ use empirically-derived frameworks to guide the assessment of risk in a structured, but flexible, manner. Specific SPJ measures for sexual offending include the Sexual Violence Risk–20 (SVR-20; Boer, Hart, Kropp, & Webster, 1997) and its evolved version, the Risk for Sexual Violence Protocol (RSVP; Hart et al., 2003). Although Andrews and colleagues (2006) considered SPJ tools to fall within the first generation of risk assessment approaches, it could be argued that it is more appropriate to consider these tools as an alternative third generation approach. This is because although clinical judgement plays a primary role in this approach to risk assessment, in this case clinical judgement is applied in a guided manner to a pre-determined set of risk domains; examples from the RSVP include history of sexual violence, psychological adjustment, social adjustment, and mental disorder. This means that there are formal systems in place that reduce the level of subjectivity inherent in the unstructured clinical judgement approach.

Instead, SPJ tools arguably allow for a greater utilisation of the unique knowledge of the offender and their circumstances that is held by the clinician to inform the assessment of treatment progress and risk, particularly in areas that are not fully captured by other dynamic tools. Such an approach acknowledges the complexity of risk assessment in the real world, in which a variety of factors, both psychological and external, can influence an individual’s behaviour at any given point in time. This complexity may be difficult to incorporate into a fully actuarial tool, which must balance breadth of variables covered with the practicalities of the time and resources required to complete the measure. Approaches that allow for structured professional judgement therefore potentially provide a useful means of assessing real-world risk because of their ability to reflect change in the attitudes and behaviours of the individual across a wide range of areas, whilst still being based on a credible empirical foundation. Support has been found for the interrater reliability of SPJ approaches (e.g., intraclass correlation coefficients (ICC) in the “fair” to “excellent” ranges for the RSVP across the four studies overviewed by Judge, Quayle, O’Rourke, Russell, & Darjee [2014]). Structured professional judgement tools also potentially allow for the development of risk predictions for populations for which there are no specific actuarial tools, due to a lack of available empirical information (e.g., female sex offenders).

Despite the advantages of the SPJ approach, it is important that its limitations in terms of predictive accuracy are noted. Although a large meta-analysis found that SPJ tools were predictive of sexual recidivism, the predictive accuracy of this approach was lower than that obtained by empirical actuarial tools (d = 0.46 and 0.67, respectively; Hanson & Morton-Bourgon, 2009). These effect sizes relate to small differences between mean scores on SPJ tools for recidivists compared with non-recidivists, and moderate differences for empirical actuarial tools. The effect sizes can also be interpreted as AUC values, with an AUC of 0.63 for SPJ tools and 0.68 for empirical actuarial tools. Thus, although SPJ tools may be a useful inclusion to the overall approach to evaluating risk and treatment outcomes, it is important that this is augmented by information obtained from actuarial tools as part of a wider assessment of risk. Additionally, some research suggests that professional judgement may be more accurately applied to risk assessment only when adjusting risk levels downwards on the basis of additional information, rather than to increase risk level (Wormith, Hogg, & Guzzo, 2012).

Fourth-generation: Case management

Due to some concerns that risk assessments were being administered for individuals but that the results of these assessments were not then being used to inform case management and decision-making (Bonta & Andrews, 2016), fourth-generation tools were developed to more explicitly highlight the necessary links between assessment and case management. The tools include measurement of common risk-related factors to assess risk level, but they also include measurement of specific responsivity needs, planning of treatment targets and intervention, and recording of treatment progress.

There are currently no fourth-generation tools that have been developed specifically for sexual offenders, although there has been some support for the ability for existing tools to predict sexual recidivism. Wormith and colleagues (2012) examined the predictive accuracy of the Level of Service/Case Management Inventory (LS/CMI; Andrews, Bonta, & Wormith, 2004) with a sample of 1,905 sex offenders and 24,545 non-sexual offenders (i.e. individuals who had a history of offending that did not include a sexual offence). They found that the LS/CMI was significantly predictive of sexual reoffending for both sexual offenders (AUC = .77) and non-sexual offenders (AUC = 0.75), indicating that fourth-generation tools may provide a promising direction for the development of future sexual offender risk assessments. Indeed, given previous findings that predictive accuracy is higher for measures when predicting outcomes that they were specifically designed for (Hanson & Morton-Bourgon, 2009), it is important to assess whether fourth-generation tools developed specifically for the prediction of sexual offending may provide even greater levels of accuracy than tools developed for general or violent offending. Because of the lack of fourth-generation tools specific to sexual offending and the resultant lack of information about their efficacy with this specific population, it is difficult to draw any strong conclusions or make strong comparisons between these and other existing measures in the prediction of sexual offending. They remain an area in need of further research, development and validation.

The Construct Validity of Dynamic Risk Factors

The most common interpretation of construct validity, as applied to dynamic risk factors, relates to whether the particular measure represents or measures what it is supposed to; this has been referred to as "fundamental" construct validity (Colliver, Conlee, & Verhulst, 2012). As such, much research has focused on determining the concurrent validity – the extent to which a particular measure correlates with existing measures of the same constructs – and the predictive validity – the extent to which a measure accurately predicts a specific relevant outcome measure – of dynamic risk assessments. Overall, research has supported the concurrent validity of common dynamic risk assessment tools, in that offenders who are categorised as high risk on one particular measure are likely to also be categorised as high risk on other measures (Beggs & Grace, 2010; Loza, Dhaliwal, Kroner, & Loza-Fanous, 2000; Nunes & Babchishin, 2012). Thus on the surface it appears as if there is a strong empirical basis for the construct validity of dynamic risk assessment. However, scholars have recently questioned the idea that concurrent and predictive validity are the two most important measures of construct validity (Borsboom, Mellenbergh, & van Heerden, 2004; Colliver et al., 2012; Haig, 2012). Instead, they argue that a given measure should be considered to have good construct validity if it is able to demonstrate a causal or explanatory link between the attributes it measures and the outcome of interest (Borsboom et al., 2004).

The implication is that dynamic risk factors could be considered to have good construct validity only if researchers are able to demonstrate a causative or explanatory link between dynamic risk measures and recidivism. In order to assess whether this is possible given the current evidence base, we first must take a step back and ask an important question:

what *are* dynamic risk factors? As defined by Bonta and Andrews (2016), dynamic risk factors are theoretically changeable factors that predict criminal behaviour, and that with changes we see corresponding changes in the likelihood of criminal behaviour. While this conceptualisation of dynamic risk factors, and therefore the assessment of dynamic risk, appears to at least theoretically meet the explanatory requirement of construct validity, there has been some recent doubt cast on whether the current reliance on correlational analyses and significance testing is a valid method of identifying truly causal factors relating to recidivism risk (Haig, 2012; Heffernan & Ward, 2015; Ward & Beech, 2014). Some scholars instead argue that the current conceptualisation of what constitutes as a "risk factor" may not be demonstrably valid in a more meaningful sense of the term. In particular, in many cases a statistical relationship between a given risk factor and future reoffending does not necessarily indicate that the factor is meaningful in terms of having a causal relationship to offending behaviour.

A further problem with validity in the area of dynamic risk relates to the multi-dimensional and indistinct nature of many dynamic risk factors (Heffernan & Ward, 2015). "Cognitive distortions" provides a good example. Typically, cognitive distortions are conceived of as non-normative belief structures that include justifications and rationalisations for sexual offending, and are regarded as a dynamic risk factor (T. A. Gannon, Ward, & Collie, 2007). However, Ó Ciardha and Gannon (2011) noted that "cognitive distortions" has been applied to a multitude of different constructs including "maladaptive beliefs" (Ward, Hudson, Johnston, & Marshall, 1997), "defensiveness" (Rogers & Dickey, 1991), "rationalisations" (Neidigh & Krop, 1992), "incorrect or deviant cognitive practices" (Ward & Casey, 2010), and "etiological cognitions" (Ó Ciardha & Gannon, 2011). Although definitions may play a larger role in the conceptualisation of scientific phenomena than perhaps they should (Haig, 2012), the degree of variation in how 'cognitive distortions' are defined poses a significant problem for developing valid measures of this dynamic risk factor (or dynamic risk assessments that incorporate cognitive distortions). Without clear definitions of risk factors, it is uncertain which features of each factor are linked to recidivism, and how to create clear scoring guidelines. Such uncertainty will not only potentially degrade the accuracy and discriminative validity of a measure, but also inter-rater and test-retest reliability. It also has implications for construct validity – how can we be sure that we are measuring what we want to measure, when we are unable to define clearly what that is?

The Data-Driven Approach to Dynamic Risk Factors

One of the possible reasons for the lack of research exploring the construct validity of dynamic risk factors and their causal role in sexual offending is the data-driven nature of research used to identify dynamic risk factors. This data-driven approach often leads to an over-reliance on empirical evidence that is available to researchers – which may be affected by particular choices of questionnaires or measures used as dynamic risk indicators - and under-reliance on theory or aetiology to guide our understanding of dynamic risk factors as scientific phenomena, and how these factors might be combined into an overall meaningful measure of risk; as explained by Haig (2013, p. 137), '[d]ata themselves are of scientific interest and importance only because they serve as evidence for the phenomena under investigation.'

Overreliance on the hypothetico-deductive methodology in psychological research, where empirical data are used to identify, describe and/or discover correlates of constructs to inform theory, has long been criticised (e.g., Cohen, 1994; Falk & Greenbaum, 1995; Rozeboom, 1960, 1997). By contrast, the abductive method may represent a more meaningful and valid approach to research (Borsboom et al., 2004; Haig, 2005, 2009, 2014). Haig (2014) describes the abductive approach as 'reasoning from factual premises to *explanatory* conclusions' (p.60), noting that 'phenomena, not data, serve as evidence for the abducted theories' (p.61). This differs from a purely inductive approach, whereby conclusions or theories are the 'same in kind' as the data used to generate them, meaning that they are more descriptive than explanatory in nature (Haig, 2014). As such, in using an abductive approach to theory generation we are more likely to develop meaningful knowledge about the phenomena of interest, including an understanding of aetiology, causal networks, and the potential for change or adaptation.

The abductive critique of the hypothetico-deductive method is clearly applicable to prior research on dynamic risk factors and assessment, which has been largely data-driven (attempting to identify the best predictors of recidivism among a set of candidate measures using regression-based statistical methods), rather than theory-driven (Heffernan & Ward, 2015; Ward, 2016). One could argue that variables which were studied as potential dynamic risk factors – such as lack of empathy for victims – were selected based on prior theoretical grounds (e.g., Marshall, Hamilton, & Fernandez, 2001), however acceptance of these variables as dynamic risk factors was reliant largely on evidence of their ability to predict recidivism, ideally beyond the contribution made by static factors. For example, an influential series of articles by Beech and colleagues showed that cluster analysis (a data-driven exploratory technique) could be used to classify sexual offenders as 'high deviance' or 'low deviance', and the deviance classification was subsequently shown to predict recidivism beyond the Static-99 (Beech, 1998; Beech, Friendship, Erikson, & Hanson, 2002; Fisher, Beech, & Browne, 1999).

Recent conceptualisations of validity suggest that a construct (or attribute) is valid only insofar as it is shown to relate causally to a criterion (Borsboom, 2005; Borsboom et al., 2004). However, data-driven approaches which merely demonstrate that dynamic risk factors are correlated with recidivism fall short of providing evidence of a causal linkage. This is likely to result in an incomplete picture of the risk posed by an individual offender that lacks an explanation of the aetiology or maintenance of behaviour. Consequently, implications for treatment formulation in terms of the most important needs to target to reduce risk are compromised. A greater understanding of the aetiology of serious offending would allow us to develop more effective strategies for early intervention, ideally to reduce first-time sexual and violent offending rather than reoffending.

The data-driven identification of dynamic risk factors can also have a negative impact other research in the area, including the measurement and exploration of offender change. As discussed further in the next chapter, although change is an area still relatively unexplored, results have been mixed for studies attempting to link change in dynamic risk to corresponding changes in recidivism rates (e.g., Barnett et al., 2013; Olver, Kingston, et al., 2014; Olver, Nicholaichuk, et al., 2014). Authors of a recent meta-analysis of dynamic risk assessments noted that a potential reason for the relatively weak predictive validity of change scores is that many current dynamic risk factors used in assessment are best conceptualised as correlates of phenomena linked to offending (i.e., symptoms of offending), rather than factors related to the causes of offending (van den Berg et al., 2018). This weakens the ability of change in measured dynamic risk to predict or explain changes in recidivism rates, and calls into question the validity of using dynamic risk assessments to inform treatment targets and progress. If dynamic risk factors were more closely associated with the direct causes of offending, one would expect change in dynamic risk to be more strongly predictive of changes in recidivism rates.

Leaving aside the issue of reliance on statistical significance testing (see Cumming, 2012), there are two major potential problems with the data-driven approach for dynamic risk: 1) there is no guarantee that the risk factors are clinically meaningful in the sense that they can be used to explain the aetiology or maintenance of offending (Heffernan & Ward, 2015; Mann et al., 2010); and 2) the identification of important risk factors is reliant upon the data or measures available to a given researcher.

The variability in the identification of risk factors caused by differences between the information contained in different datasets is displayed clearly by the emergence of competing “second-order” risk domains – composite risk factors that are predictive of reoffending, usually obtained by exploratory factor analysis on data from a psychometric battery. One example of this approach from the sex offender literature is from Allan, Grace, Rutherford, and Hudson (2007). These authors used exploratory factor analysis to identify dynamic risk factors from a large psychometric battery (a total of 20 different measures, including multiple sub-scales for some measures) that had been completed by a sample of sexual offenders against children prior to undergoing prison-based treatment. Four risk domains were identified that were each significantly predictive of sexual recidivism: Social Inadequacy (containing measures relating to low social competence and negative mood); Sexual Interests (containing measures relating to levels of sexual fantasies); Anger/Hostility (containing measures relating to anger expression and regulation); and Pro-offending Attitudes (containing measures relating largely to distorted cognitions and attitudes). Allan et al. combined these risk factors into a measure of ‘Overall Deviance’ which was shown to increase the predictive accuracy for recidivism beyond the Static-99. Previous researchers have also taken this approach for the development of general risk domains, however with slightly different results. For example, Olver, Nicholaichuk and colleagues (2014) identified three domains of dynamic risk for sexual offenders – Socioemotional Functioning, Anger/Hostility, and Misogynist Attitudes; Beech (1998) also identified three risk domains in his analysis of a psychometric battery, although these domains assessed conceptually different types of functioning: Social Competency, Pro-offending Attitudes, and Sexual

Interests. Although there is evidence of some overlap between these factors, it is clear that the identification of relevant measures of risk for a given offender population varies depending on the measures available to a given researcher.

Moving beyond a data-driven approach will require different ways to identify risk factors, and possibly different conceptualisations of risk factors. One approach that has been suggested elsewhere (Haig, 2005, 2012) is to modify our research methodologies to be more in line with an abductive approach to science, whereby theories are formed to explain patterns identified within the data (also called “phenomena”), rather than data analysis being used to directly generate theories. In terms of research into dynamic risk assessment, this would require increased utilisation of techniques such as confirmatory factor analysis, which can be used to test proposed models for dynamic risk factors. These could then be used to generate theories and hypotheses relating to relationships between certain risk factors or domains and recidivism (Haig, 2005). It would also require a greater emphasis on validating and replicating the findings of other research in order to ensure that we are identifying true phenomena rather than idiosyncrasies of particular datasets (Beech, 1998; Cumming, 2012). It is hoped that through changing how we identify dynamic risk factors, we might be able to develop a deeper and more meaningful understanding of how these factors contribute to the generation and maintenance of offending, as well as how these factors combine to determine the overall level of risk of an offender.

The Application of Dynamic Risk Factors

The Dual Uses of Dynamic Risk Assessments

Reasons for the use of the hypothetico-deductive method are understandable when one considers the primary goal of risk assessments: to predict the likelihood of future offending for a given individual. Thus it is logical to identify dynamic risk factors by their ability to predict reoffending beyond the static, actuarial factors that had already been shown to have predictive validity. However, this becomes an issue due to the use of risk assessments to assess individual needs (and therefore inform treatment decisions) as well as predict risk. Currently, both of these tasks are typically performed using the same tools, despite the differences in what is required of the tools for these two tasks: the assessment of needs is in essence a diagnostic task, in that the measure is being used to determine the presence or absence of a certain condition or characteristic (and in treatment contexts result in warranted causal inferences; Ward & Fortune, 2016), whereas the assessment of risk is a prognostic task, in that the measure is being used to assess the probability of a future outcome (Helmus & Babchishin, 2017).

The difference between these two tasks has several practical implications for scale development, including the selection of items to include in the measure. Diagnostic scales are inherently norm-referenced (i.e., they are trying to capture the degree to which an individual displays a particular characteristic), whereas prognostic scales are inherently criterion-referenced (i.e., they are designed to specifically predict a particular outcome (Helmus & Babchishin, 2017). Whereas norm-referenced scales should ideally include multiple items that assess the same construct in different ways to ensure the reliability of the 'diagnosis', criterion-referenced scales are solely concerned with predictive accuracy, and are therefore largely atheoretical. Because practical reasons often require measures to be as short as possible whilst still serving their purpose, criterion-referenced scales should ideally include a small number of items that each represent a distinct factor that has been linked with offending; overlap between what is predicted by different items is to be avoided. The primary concern here is efficiency and predictive accuracy of the scale, rather than the theoretical implications of the construct being measured. The implication is that in order to meet these competing requirements whilst ensuring that measures have construct validity where this is important, we may need to develop different measures for assessing risk as opposed to identifying treatment targets (or incorporate two different scales into the one measure, similar to the format of fourth-generation tools). This implication suggests that it is unwise to translate dynamic risk factors from risk prediction measures into causal constructs to be used in the explanation of offending and to direct treatment, without considerable theoretical reworking (Ward & Fortune, 2016).

This is particularly important where the risk assessments are currently being used in treatment. Merely demonstrating correlations between dynamic risk factors and recidivism falls short of providing evidence of a causal linkage, or of providing a strong explanatory theory behind the correlation. This is likely to result in an incomplete picture of the risk posed by an individual offender that is lacking in an explanation of the historical causes or current maintenance of behaviour. Consequently, implications for treatment formulation in terms of the most important needs to target to reduce risk are compromised. A greater understanding of the aetiology of serious offending would allow us to develop more effective strategies for early intervention, ideally to reduce first-time sexual and violent offending rather than reoffending. In moving away from the purely data-driven approach to risk/needs tool development, it is hoped that we can develop a deeper and more meaningful understanding of how these various factors may contribute to the generation and maintenance of offending, as well as how they combine to determine the overall level of risk of an offender. It also avoids issues with developing risk assessments that are generalisable and that can be used for offenders for whom it is difficult to obtain data.

The Use of Dynamic Risk Factors in Practice Contexts

Despite the concerns with the validity and explanatory depth of dynamic risk factors explored above, the assessment of dynamic risk factors is a common component of current treatment delivery and decision-making. As with research conducted on dynamic risk factors, the ability for dynamic risk factors to be measured validly and reliably is of central import when these factors are applied to offender treatment and management. However, most studies of the reliability and validity of dynamic risk factors and assessments have been conducted in research rather than applied settings, in that the measures are generally used by researchers or developers rather than by professionals in a correctional context (e.g. parole officers, forensic psychologists and custodial officers). This raises questions about whether the reliability and validity of these measures will be maintained when they are no longer being scored by trained researchers or research assistants, but instead by staff who may have many other responsibilities.

Reasons why we might expect differences in the scoring of these measures by correctional compared to research staff are the lack of specialised, standardised training; high levels of work-related stress (National Institute of Justice, 2007); and large workloads and time pressures leading to a greater reliance on clinical or professional judgment rather than a strict adherence to scoring guidelines (Jones, Brown, & Zamble, 2010; Public Safety Canada, 2008). This would be particularly salient for measures that use a largely unstructured scoring format, allowing for a greater influence of personal heuristics and cognitive biases (Payne, Bettman, & Johnson, 1993). It has also been suggested by some researchers that this reliance on clinical judgement rather than the structured scoring criteria might result from a reluctance of professionals to accept the idea that their judgements might be less accurate than purely quantitative methods of assessing risk (Schlager, 2009; Schneider, Ervin, & Snyder-Joy, 1996). There is also some evidence to suggest that the fear of political and professional implications of having rated someone as low risk who later goes on to reoffend (even if the rating was correct), leads correctional staff to manipulate dynamic scores in a way that over-estimates risk, with large resource and financial implications for the corrections service as a whole (Lanterman, Boyle, & Ragusa-Salerno, 2014; Schlager, 2009; Schneider et al., 1996).

A further threat posed to the validity of risk assessment in applied settings is the quality of training provision for these measures. Previous research has indicated that there is a positive association between the quality of training provided to scorers and the predictive validity of risk assessment (Lowenkamp, Latessa, & Holsinger, 2004). For example, formal instruction on the use of the measures led by trained practitioners is superior to "bootstrap" training led by inexperienced or untrained colleagues, and the provision for hands-on practice improves quality (Lowenkamp et al., 2004; U.S. Department of Justice, 2007). The importance of training with risk assessment tools was highlighted by a series of meta-analyses using a total of 101 validation studies on the Level of Service Inventory (LSI) risk assessment tool (Andrews et al., 2011). The studies were assessed on a number of factors that might moderate the predictive validity of the LSI, including length of follow-up, sample characteristics (such as gender and nationality), and LSI "allegiance" (defined as the level of involvement of the LSI developers in data collection for the study). Andrews et al. found that LSI allegiance was a significant moderator of the predictive validity of the LSI, with the stronger the allegiance, the higher the predictive validity. They suggested that LSI allegiance might be best understood as a proxy for the quality of the implementation and integrity in research methods (such as the selection of appropriate outcome measures). Andrews et al.'s (2011) results highlight the importance of a close adherence to scoring and implementation guidelines that are developed during quality training, in order to ensure maximal utility of the risk assessment tool in question.

Because performance in applied settings is essential for the utility of a risk assessment measure, it is important that we understand fully how reliability and validity of a given measure are affected by extending use from a research to an applied setting. Jones et al.

(2010) investigated the extent to which the predictive ability of risk as assessed by parole officers differed from the predictive ability of risk scored by researchers. They used a prospective design where risk was measured at multiple time points in order to best emulate real-world use of the measures. In order to further emulate real-world circumstances, parole officers provided crude proxy ratings of each area of dynamic risk based on their perceptions of offender circumstances and were not subject to quality assurance processes. Researchers, on the other hand, provided detailed and structured assessments of risk based on multi-dimensional case review (including semi-structured interviews and file reviews) and were subjected to routine quality assurance (such as inter-rater reliability checks). Contrary to expectation, researchers found that the predictive ability of the ratings of the parole officers and researchers were not significantly different, with AUCs of .76 and .79, respectively, indicating medium-high levels of predictive validity. In addition, ratings of external acute risk factors (such as employment) were highly correlated between the two groups, although those for internal acute risk factors (such as stress) were not significantly correlated. Thus, the ability of parole officers to predict recidivism based on crude proxies of risk was equal to that of highly structured and multi-dimensional assessments of risk. Jones et al. suggested that perhaps the extensive level of interaction between parole officers and offenders enabled them to gain a better picture of important collateral information about the offender, such as family, education and interaction with other health providers.

Although Jones et al.'s (2010) results are encouraging in that they suggest that risk assessment can be valid in an applied setting, it is important to note that the parole officers in this study were not expected to strictly adhere to scoring guidelines for each measure, but instead rated their perceptions of how a risk factor related to a given individual, in a similar way to the procedure used for structured clinical judgement tools. In other words, while it appears as though risk assessment can be accurate and valid in applied contexts, it is not so clear that the validity of actuarial tools can be transferred as successfully between research and practical contexts. It is important to note that as actuarial measures of dynamic risk continue to improve in their level of predictive validity and in the provision of estimated base-rates of offending for different risk bands (e.g. for the VRS:SO; Olver, Beggs Christofferson, Grace, & Wong, 2014), it will become increasingly important that professionals are able to utilise structured risk assessment tools accurately. The ability of professionals to provide estimates of recidivism rates by risk level becomes even more significant as the possible sentencing options for high-risk offenders become increasingly restrictive and intrusive in the lives of offenders (e.g., preventive detention and extended supervision orders in New Zealand; Ryan, Wilson, Kilgour, & Reynolds, 2013). Thus, given that the extant literature largely supports the notion that the validity of risk assessment can change substantially depending on implementation and adherence to guidelines, more effort is warranted to ensure protocols are in place for effective training and ongoing quality assessment for those responsible for risk assessment in a professional context.

Threats to the Validity of Dynamic Risk Measures: Socially Desirable Responding

One important threat to the validity of dynamic risk measures, particularly those in which offender self-reports play a role, is impression management or socially desirable responding (SDR; see Tan & Grace, 2008, for review). SDR refers to the tendency of some individuals to respond in ways that are likely to elicit approval from others, and to refrain from responding in ways that would be met with disapproval (Crowne & Marlowe, 1964). In terms of self-report measures, this tendency means that individuals may be influenced to respond to individual items not only based on their beliefs relating to the item content, but also what they believe to be a socially appropriate response. Such patterns of responding pose a unique challenge within an offender population, where an idiosyncratic desire to appear “overly positive” (Paulhus, 2002) is further augmented by a penal system that creates clear incentives for individuals to present in a positive way for parole boards, judges, probation officers, and other individuals making decisions affecting the length and type of custodial sentences (Davis, Thake, & Weekes, 2012).

To the extent that offender self-reports are considered in classification and parole decisions, assessing the credibility of those reports is obviously important. The inaccuracies in measurement that could potentially result from SDR threaten not only the classification and parole decisions, but also affect the ability to assess accurately the level of dynamic risk or need of an individual offender. Some psychometric measures used to assess dynamic risk factors have highly transparent items, so that it is fairly obvious to the responder as to what the test is measuring and therefore what the socially acceptable responses might be. For example, the Abel-Becker Cognitions Scale (ABCS; Abel et al., 1989), which is commonly used to assess offence-supportive beliefs and attitudes with sexual offenders against children, includes items such as “I show my love and affection to a child by having sex with her (him)” and “A child who doesn’t physically resist an adult’s sexual advances really wants to have sex with the adult”. Using transparent measures makes it relatively easy for offenders to minimise or deny problematic attitudes, and to exaggerate any positive or pro-social traits.

Because of these potential problems with offender self-reports, researchers have often used measures of SDR as part of psychometric batteries to assess dynamic risk (e.g., Beech, 1998). Variance associated with SDR is then partialled out prior to making a dynamic risk classification (cf. Saunders, 1991). Although this is a common practice, there is little evidence that correcting for SDR in this manner improves the accuracy of decision making in applied settings in general (R. E. McGrath, Mitchell, Kim, & Hough, 2010). Results with forensic samples are similar. Mills and Kroner (2006) found that using the impression management subscale of the Balanced Inventory of Desirable Responding (BIDR; Paulhus & John, 1998) to correct the self-reports of incarcerated offenders on a measure of dynamic risk decreased, not increased, the predictive validity for recidivism (although note that the decrease was not statistically significant). Recently, Stevens, Tan and Grace (2016) showed that correcting sexual offenders' self-reported dynamic risk scores for SDR (measured by the Marlowe-Crowne Social Desirability Scale; Crowne & Marlowe, 1964) had virtually no effect on predictive validity for sexual recidivism.

Because removal of SDR variance does not improve the correlation of self-report dynamic risk measures with recidivism, researchers have suggested that SDR scales like the BIDR or MCSDS may actually be measuring a personality trait or enduring disposition related to need for social approval. According to this view, SDR may be correlated with dynamic risk factors (indeed, SDR was negatively correlated with dynamic risk in Mills & Kroner (2006) and Stevens et al. (2016); see also Mathie & Wakeling (2010), but is not related to recidivism risk directly). This view is consistent with a recent reinterpretation of SDR by Uziel (2010), who suggested that instead of response bias, measures of SDR should be regarded as 'interpersonally oriented self-control', that is, SDR reflects the individual's ability to adjust to social situations and seek approval from others.

Although further research is needed in order to more fully understand what SDR actually is, how it is structured, and how it affects risk assessment in different offender populations, overall research suggests that SDR may not be the large challenge that many assume it to be, and that its impact on the ability to measure dynamic risk with self-reports may be less severe than originally thought.

Conclusions on Dynamic Risk Factors

In terms of the overall accuracy and utility of risk assessment tools, the move towards a greater consideration of dynamic risk factors when assessing risk has been a promising step forward. Not only has the inclusion of dynamic factors shown to improve the predictive validity of actuarial tools (Craissati & Beech, 2003; Gendreau, Little, & Goggin, 1996; Hanson & Harris, 2000), but these measures are also theoretically able to inform treatment targets, allow for the assessment of change in risk over time, and incorporate more meaningful risk factors that can be connected to the aetiology and maintenance of antisocial behaviour, all of which are important considerations for a rehabilitative approach to criminal justice (and which will be discussed further in the following chapter). However, it is important to note that the way in which these tools are developed and utilised will moderate the relationship between these theoretical benefits of incorporating dynamic risk factors and how these tools function in reality.

As discussed above, the current research methods used to identify dynamic risk factors, and the ways in which we combine and apply these factors to the measurement of risk, is possibly creating a disconnect between the theoretical conceptualisation of dynamic risk and what these tools are measuring in practice. While dynamic risk assessment tools may be reasonably accurate in their predictions of reoffending, it is important that we recognise that this does not necessarily mean that the risk factors used in these measures are psychologically meaningful, or that they contribute to our understanding of the aetiology of, or indeed to the desistance from, antisocial behaviour (Heffernan & Ward, 2015). As such, it is vital that we consider the methodology utilised when identifying important predictors of risk and remember that the development of meaningful knowledge about risk necessarily includes knowledge of aetiology and causal networks (Borsboom et al., 2004; Haig, 2005, 2012).

Addressing these applied and theoretical challenges related to the assessment of dynamic risk factors requires a change in the methodologies and analytical techniques used in dynamic risk research (e.g. increased use of causal modelling and confirmatory factor analysis), as well as a move towards a more theoretically-driven identification of relevant dynamic risk factors. The current thesis provides a step in this direction by first validating an existing influential theory of child sexual offending aetiology, and then further exploring the nature of offender change, including whether change is best conceptualised as categorical or dimensional in nature. The aim of this thesis was to provide a more solid understanding of the meaningful causal factors involved in offending, and then apply this information to an abductive exploration of offender change (an area that is relatively unexplored currently), with the hopes that findings could be used to inform ongoing causal theory generation in this area. To provide the necessary background and context for this research, the following chapter focusses on the current literature regarding offender change.

Chapter Two: Measuring and Conceptualising Offender Change

As discussed at the beginning of the previous chapter, one of the major benefits of dynamic risk assessments is that they are theoretically able to assess changes in risk over time. However, due to the potential issues with the accuracy of clinical judgement outlined previously, it is important that any approach to offender change is structured and minimises the possibility of bias. Given previous findings, it is fair to hypothesise that unstructured clinical judgement will *over-estimate* the influence of time or treatment on offender change; although dynamic risk factors are indeed theoretically able to change, it is important to remember that most factors measured are relatively enduring, with pre-treatment assessments of risk commonly found to remain significantly predictive of recidivism with follow-ups of several years *post*-treatment (Beech et al., 2002). This suggests that dynamic factors may be more stable than one might initially consider them to be, leading evaluators to incorrectly quantify the extent to which these factors change over the course of an often relatively short period of treatment. Indeed, previous research has underscored issues with clinical judgement of change, showing that factors unrelated to treatment progress can affect ratings; for example, individuals with a more positive view of treatment and sex offenders are more likely to report identification of positive treatment outcomes (Kivlighan & Tarrant, 2001).

Any adjustments to risk assessment on the basis of change observed over the course of a treatment programme will therefore need to acknowledge the gradual and complex nature of desistance from sexual offending. One issue here is the relatively small amount of research that has assessed whether a) dynamic factors really do change over time, and b) whether these changes can reliably be linked with changes in risk.

Do Dynamic Risk Factors Change Over Time?

The link between offender change and the concept of dynamic risk is clear – after all, a crucial component of the concept of dynamic risk is that it is just that – dynamic, or in other words, changeable. Although it is important to acknowledge the criticism made by some authors that the measurement of change in dynamic risk is not meaningful in terms of addressing the causes of offending (e.g., Ward, 2016), nonetheless measured change may have practical importance in terms of predicting future risk. However, the ability for dynamic risk factors to demonstrate change has seemed to be more readily accepted in theory than actually tested in research. For some oft-cited dynamic risk factors, although an empirical link with recidivism has been established and there may be face validity in terms of being changeable, studies have often only examined them at one point in time and this failed to demonstrate their ability to change in practice (e.g., Beech et al., 2002; Dempster & Hart, 2002).

Where more recent studies have begun to explore change, findings have been mixed. For example, a negligible amount of change was found in one study on psychometric test scores intending to tap into dynamic risk factors for violence, across a 20-month treatment programme for inpatient forensic mental health patients (Hildebrand & de Ruiter, 2012). Perhaps of more relevance to sexual offending-specific factors, one study of treated sexual offenders found no significant change on measures of cognitive distortions and deviant attraction post treatment (Jung & Gulayets, 2011). This raises questions about the nature of dynamic risk factors and whether they truly are changeable (and therefore whether assessments of risk *should* be amended), however this becomes less clear when one considers other possible reasons for the lack of measureable change on these factors, including ineffective treatment, or inappropriate or insensitive measurement techniques. This raises a problem evident in the psychometric assessment of dynamic risk factors: the difficulty in determining whether null results regarding change across treatment are the result of true lack of change (i.e. a poor treatment effect due to programme ineffectiveness or participant factors such as poor motivation), insensitive measurement relating to the tests chosen, or, that the 'dynamic' factors under investigation are not really dynamic.

In contrast to studies reporting insignificant amounts of change over time, other studies have reported significant improvements between pre-treatment and post-treatment psychometric test scores relating to dynamic risk areas such as pro-criminal attitudes, family/marital relationships, and education/employment (Brooks Holliday, Heilbrun, & Fretz, 2012). Likewise, substantial apparent changes from pre- to post-treatment have tended to be found when dynamic factors have been measured psychometrically among sex offenders (Hudson et al., 2002; Marques, Wiederanders, Day, Nelson, & Van Ommeren, 2005). One recent study examined treatment change based on a psychometric battery administered pre- and post-treatment to a sample of 267 men who had been convicted of a contact sexual offence (Olver, Nicholaichuk, & Wong, 2014). They found that on average the men made significant pre-post changes across almost all measures included in the battery, including measures of socioemotional functioning, anger/hostility, and misogynist attitudes. Furthermore, the authors found that the effect size of the changes were typically in the moderate range ($d > .50$). This replicated the findings of a previous study that examined pre-post treatment change across a number of measures for 313 males who had been convicted of a sexual offence (Nunes et al., 2011). This study also found that, on average, the men demonstrated significant changes that were approximately moderate in size, including on measures of loneliness, cognitive distortions, and dynamic risk (measured using the Stable-2000). When change was analysed at the individual level, they found that approximately one-third of participants displayed reliable change and reached normative levels of functioning post-treatment.

Results are therefore mixed regarding whether dynamic risk factors are truly dynamic, although the preponderance of more recent research suggests that individuals are indeed able to make change in these areas across time or after specialised treatment. However, even if dynamic risk factors are proven to be changeable, there is still a further question regarding whether this change is meaningful in terms of reoffending. This question is discussed in the following section.

Does Change in Dynamic Risk Predict Change in Offending?

In many studies in which dynamic risk changes have been explored, recidivism outcomes were not included in the investigations. Arguably, as well as needing to be empirically linked to recidivism, and changeable, it is also inherent in the definition of dynamic risk factors that any changes should be *meaningful* (i.e. linked to changes in actual reoffending risk). In fact, this is a central tenet of the needs principle of offender rehabilitation as described by Bonta and Andrews (2016). However, as noted by Serin, Lloyd, Helmus, Derkzen and Luong (2013), the question of whether changes (i.e. in dynamic risk factors) are reliably associated with recidivism likelihood is relatively unexplored. Research has only more recently begun to test the assumption empirically.

In one test of the link between treatment change and outcome, Beggs and Grace (2011) demonstrated that specific within-treatment changes in dynamic risk factors in the desired direction, measured psychometrically, can be linked with reduced recidivism at follow-up. In doing so, they reported on the problematic nature of analysing raw change scores: on any given test, individuals with pre-treatment scores towards the more problematic end of the scale (indicating higher levels of dynamic risk) have the opportunity to show greater levels of change across treatment, as they have more ‘room to move’. However a further problem is also clear: given that both riskier scores and lower change should theoretically be linked with higher recidivism, yet those with riskier scores have the opportunity to attain higher change scores, use of raw change scores on measures that were not specifically developed to measure change is inherently flawed. To manage this issue, Beggs and Grace employed a method of regression in which pre-treatment scores were partialled out of the prediction equation. This allowed a more meaningful pattern of results to emerge, in which positive treatment change was associated with reduced sexual recidivism overall, and for three out of four dynamic risk domains (employing the Allan et al. (2007) framework: social inadequacy, sexual interests, and anger/hostility; the fourth domain, pro-offending attitudes, approached significance). This technique has subsequently been applied by Olver, Nicholaichuk, Kingston and Wong (2014) in their exploration of therapeutic change and recidivism using a psychometric risk prediction instrument (the Violence Risk Scale-Sexual Offense Version; VRS-SO).

Additionally, and as mentioned previously, a recent meta-analysis assessed the predictive validity of change scores from six previously published studies, based on six unique samples comprising 1,980 participants (van den Berg et al., 2018). They found a fixed-effect weighted hazard ratio of 0.91 (95%CI [0.87-0.95]) after controlling for static and pre-treatment dynamic risk, indicating that prosocial change on dynamic risk assessments was significantly predictive of reductions in sexual recidivism. There was not enough information to run moderator analyses. Although change was found to be significantly linked with recidivism in this meta-analysis, the authors concluded that the small effect size suggested that only a small part of change in recidivism was attributable to the measured change in dynamic risk. They suggested that perhaps current approached to measuring treatment change were too focussed on correlates of offending (i.e. symptoms of offending), rather than measuring meaningful causal factors in offending. This suggests that in order to more effectively and meaningfully measure offender change and desistance, we may need to change the typical approach to measuring change.

In response to issues with the measurement of change in dynamic risk factors, other studies have employed a different method known as clinically significant change methodology to explore the link between within-treatment changes in dynamic risk and recidivism (e.g., Barnett et al., 2013; Olver, Beggs Christofferson, & Wong, 2015). This method avoids the problems associated with raw change scores, as in addition to considering change magnitude, post-treatment scores are evaluated against non-deviant norms to determine whether the individual has qualitatively "improved", "recovered", is "already ok" (i.e., never scored outside the normative range), or remained "unchanged". While this method offers a user-friendly and readily interpretable classification system for individuals based on what their dynamic risk test scores say about their treatment outcome, Olver and colleagues (2015) and others (e.g., Barnett et al., 2013) have overviewed the limitations of the method and noted mixed findings, in particular that the usefulness of the output is dependent on the quality of the measures used. In general, Olver et al. (2015) suggested that the use of a single, purpose-designed risk tool containing multiple dynamic factors, such as the VRS-SO or the STABLE 2007, may offer advantages over the psychometric battery approach for the consistent and meaningful applied measurement of dynamic risk factors, and change in these across treatment.

One common factor in the change studies discussed above is that the assessment of dynamic risk occurred at only two points in time – prior to, and then following, treatment. Whilst studies employing this design have been very useful in terms of establishing empirical relationships between specific changes in dynamic risk and decreased recidivism, and exploring the clinical measurement of within-treatment changes, they have focused exclusively on the period of treatment engagement as the change mechanism for risk.

However, theoretically speaking, other factors could influence the presence or expression of dynamic risk (resulting in change), such as maturation (Hirschi & Gottfredson, 1983), social context and influences (Sampson & Laub, 1995), or in the case of sexual offending, age-related decline in sexual response (Blanchard & Barbaree, 2005). It has also been suggested that including at least three waves of assessment increases the probability of detecting change (Brown, Amand, & Zamble, 2009). A recent multi-wave study by Greiner, Law, and Brown (2014) illustrated the tracking of seven major theorised dynamic risk factors (employment, personal/emotional factors, substance use, criminal attitudes, criminal associates, family functioning, and community functioning) among female offenders following their release from prison, across four assessment waves at six-monthly intervals. They found that all seven factors were significantly related to survival time without reoffending in the community, and that prediction was improved by their use of multiple assessments of dynamic risk across time. On the other hand, change across multiple waves using a well-validated dynamic risk tool for sex offenders, the STABLE 2007, has been found to not be associated with recidivism (Hanson et al., 2007). As such, although assessing dynamic risk factors at multiple time points both during and after treatment appears to be a promising technique in terms of improving the assessment of change, it is apparent that there are other factors that contribute to the mixed results of studies on change. It is possible that part of the problem lies with our current conceptualisation of what constitutes a dynamic risk factor, as discussed in the previous chapter.

Clearly, more research is needed on the assessment of changes in dynamic risk, with numerous challenges having been identified for applied settings. As discussed above, dynamic risk evaluations in applied settings typically have a great impact on individuals' progress through the criminal justice system, and assessments of change (across treatment or with continued repeat assessments) are certainly no different. For clinicians this carries a great responsibility, and the need to ensure that the methods we select to assess the changes made by our clients are both capable of detecting change that has occurred, and meaningful in terms of being predictive of actual reductions in the likelihood of recidivism. As noted by van den Berg and colleagues (2018), the modest ability for offender change to predict recidivism is perhaps indicative if the need to more closely scrutinise the current way in which change is measured and applied to judgements of offender progress and risk. In light of this, the next section discusses current approaches to the measurement of treatment change amongst offenders, and is followed by an exploration of how this measured change is used to conceptualise the mechanisms underlying change.

Measuring Change in Dynamic Risk Factors

Clearly, more research is needed on the assessment of changes in dynamic risk factors, with numerous challenges having been identified for applied settings. As discussed above, dynamic risk assessments in applied settings typically impact greatly on individuals' progress through the criminal justice system, and assessments of change (across treatment or with continued repeat assessments) are certainly no different. For clinicians this carries a great responsibility, and the need to ensure that the methods we select to assess the changes made by our clients are both capable of detecting change that has occurred, and meaningful in terms of being predictive of actual reductions in the likelihood of recidivism. The issue of how to incorporate information related to individual change into assessments of risk becomes especially relevant when one considers the implications that a reduction in estimated risk may have for an individual and their progress through the system (and, by extension, implications for the community they will be returned to), including their chances for parole and what their parole conditions and level of oversight will be, decisions around whether to extend their custodial sentence, and the availability of particular rehabilitation programmes.

A majority of the existing literature assessing change across treatment for sexual offending has measured treatment change as if it were a dimensional construct (see Beggs, 2010, for review). For example, one common method of measuring treatment change is to calculate the quantitative difference between pre- and post-treatment scores across a range of psychometric measures. Differences in the amount of change made are then used to test associations between change and outcomes of interest (such as recidivism; e.g., Allan, Grace, Rutherford, & Hudson, 2007; Beech & Ford, 2006) or factors thought to influence the amount of change made (such as therapist features; e.g., Marshall et al., 2002, 2003). Studies utilising this method to measure treatment change have typically found that individuals make pro-social change over the course of treatment on average, however there is a lack of reliable and consistent findings linking this pro-social treatment change to reduced recidivism (discussed further below).

A similar method of measuring treatment change that is used in the literature is through assessing changes in dynamic risk based on risk assessment tools that incorporate some measure of treatment progress (Beggs, 2010). Some guided clinical judgement tools (such as the Structured Anchored Clinical Judgment [Thornton, 1997] and the Multifactorial Assessment of Sex Offender Risk for Recidivism [Barbaree, Seto, Langton, & Peacock, 2001]) incorporate items relating to treatment progress into their post-treatment assessments of risk; however, these tools are used more for an overall assessment of post-treatment risk rather than a measure of overall treatment progress. Treatment change is sometimes measured using these risk assessment tools by calculating change as the difference between pre- and post-treatment risk scores. A recent meta-analysis used this approach to assess the predictive validity of change scores derived from dynamic risk assessment tools across nine studies and six unique samples (van den Berg et al., 2018). Overall, they found that change scores based on dynamic risk assessments were able to successfully predict sexual recidivism (d = .26, 95% CI [.10 - .42]), and added incremental predictive validity beyond that provided by static risk and pre-treatment dynamic risk alone. That said, they noted that the effect sizes for the predictive validity of treatment change are relatively small, which potentially indicates that current measurement of change is not focussed on factors that directly cause offending behaviour, or that we are not measuring change appropriately or effectively.

Although most of these dynamic risk assessment tools do not explicitly incorporate measurement of treatment change, there is one exception: the Violence Risk Scale - Sexual Offense version (VRS-SO; Olver, Wong, Nicholaichuk, & Gordon, 2007), which includes a structured method for scoring overall treatment change. The VRS-SO calculates post-treatment risk across a series of dynamic risk factors by adjusting pre-treatment scores based on progression through a series of "stages of change". These stages of change are drawn from the Transtheoretical Model (Prochaska, Diclemente, & Norcross, 1992), and represent the internal change process occurring in individuals during treatment, providing a way to categorise individuals based on their intentions or demonstrated efforts to make change. The VRS-SO reduces dynamic risk factor scores by 0.5 points for each stage progressed through past the point of "Contemplation"; therefore, the method by which the VRS-SO calculates change is based on change categories, but is translated into a dimensional overall change score. Notably, there have been a number of studies that have shown total VRS-SO change scores to be a reliable and valid predictor of recidivism after controlling for static and pre-treatment dynamic risk (Beggs & Grace, 2011; Olver, Sowden, et al., 2018; Olver et al., 2007). The interrater reliability of VRS-SO stages of change scoring was also supported in a study by Olver and colleagues (2007), who found a "good" level of agreement in scoring (ICC = 0.68). This particular measure therefore allows for the identification of individual treatment targets, and provides a measure of static and dynamic risk, some assessment of responsivity issues, and information on treatment gains all incorporated into a single measure (for these reasons, the VRS-SO could arguably be considered a fourth-generation, rather than a third-generation, dynamic risk assessment tool).

A third way in which treatment change can be measured is through the use of tools specifically developed to measure change (Beggs, 2010). Examples of these tools include the Standard Goal Attainment Scaling (SGAS) for sexual offenders (Hogue, 1994) and the Sex Offender Treatment Intervention and Progress Scale (SOTIPS; McGrath, Lasher, & Cumming, 2012). Change is typically measured on these tools using structured scoring frameworks across items that encompass both common dynamic risk factors and treatment-specific motivation or engagement items. Additionally, these measures can be scored multiple times across treatment to track progress and inform treatment targets. For example, the SOTIPS has previously been used to successfully guide collaborative treatment planning with offenders (Lasher, McGrath, Wilson, & Cumming, 2015). Importantly, change measured using both the SGAS and the SOTIPS has been found to significantly predict reoffending (Beggs & Grace, 2011; R. J. McGrath et al., 2012).

Treatment Change: Categorical or Dimensional?

Although there are multiple ways in which treatment change is measured in the literature, there is one shared feature of these approaches already mentioned above: treatment change is typically measured as a dimensional construct rather than as categorical. In each of the approaches outlined above, change is represented on a continuous scale, calculated through some method of collating differences between pre-and post-treatment scores across a series of dynamic risk factors or engagement/motivation-related factors. Even when treatment change is categorised for use in risk communication or prediction (e.g., as with the VRS-SO, for developing normed recidivism estimates [Olver, Mundt, et al., 2018]), categories are typically arbitrarily assigned based on non-theoretical cut-points (e.g., cut-points based on one standard above or below the mean change score).

However, whether treatment change is truly dimensional in nature has not explicitly been explored in the literature, for both sexual offending and offending more broadly. Despite the seemingly dimensional nature of treatment change, it is entirely possible that treatment change is best conceptualised as categorical rather than dimensional; its categorical nature could be obscured by the current use of continuous measures to capture treatment change. The fact that previous studies have been able to identify significant associations between treatment change (measured on a continuum) and recidivism may suggest that this distinction is not important in practical terms. However, there are important differences between categorical and dimensional constructs that have crucial implications for the way in which constructs are most effectively measured (in addition to theoretical implications; this is discussed further in Study Two). For example, measures of dimensional constructs are typically more complex and contain more items than measures of categorical constructs, as more information is needed to comprehensively capture the spectrum of possible scores (J. Ruscio, Haslam, & Ruscio, 2006). Conversely, measures of categorical constructs are generally shorter, with a focus on accurately discriminating between groups at points where there might be greater group overlap.

In addition to the suggestion that we might be measuring the wrong factors when assessing treatment change (van den Berg et al., 2018), utilising measures that treat change as a dimensional construct when it is really categorical could be another explanation for the relatively small effect sizes found when previous studies have explored the link between treatment change and recidivism. It is notable that one of the dynamic risk tools that has been more successful at demonstrating a link between treatment change and recidivism – the VRS-SO – utilises an underlying categorical framework to assess treatment progress (the

Transtheoretical Model's stages of change), despite using continuous scales for overall treatment change and risk.

The development of offender categories (or typologies) is not a new concept in forensic psychology (see Byrne & Roberts, 2007, for review). Indeed, typologies have also been developed based on individual characteristics or offending behaviour that are specific to sexual offending (see Robertiello & Terry, 2007, for review). Perhaps the most basic level of typology is the distinction often made between individuals who have offended against adults and those who have offended against children (Looman, Gauthier, & Boer, 2001). However, other more specific typologies have also been developed, including for sexual offending against the elderly (Burgess, Commons, Safarik, Looper, & Ross, 2007); juvenile sex offenders (Fox & DeLisi, 2018); high risk sex offenders (Kaseweter, Woodworth, Logan, & Freimuth, 2016); and groupings based on static risk (Ennis, Buro, & Jung, 2016), dynamic risk (Martínez-Catena, Redondo, Frerich, & Beech, 2017), and level of overall deviance (Beech, 1998). Perhaps the most researched sexual offender typologies are the Massachusetts Treatment Centre (MTC) typologies, developed separately for offenders against adults and offenders against children (Knight & Prentky, 1990). Subsequent validations and refinement of the MTC typologies have demonstrated that they are replicable and are meaningfully linked with offending aetiology and recidivism (Ennis et al., 2016); however, limitations have been noted regarding the relatively small representation of some of the groups identified in the development sample (Ward, Polaschek, & Beech, 2006).

The basic concept behind the development of typologies is the idea that offenders represent a heterogeneous population of individuals, who differ in terms of their behavioural patterns and psychological characteristics. By identifying common groupings of individuals within this broader offender population, it is hoped that risk management and intervention can be better targeted to the unique characteristics of each group (Byrne & Roberts, 2007;

Marshall, 1997). One more recent study focussed on this goal explicitly, by developing a typology of sexual offenders based on characteristics and behaviours relevant to treatment responsivity (Woessner, 2010). This concept of typologies can also be applied to treatment change in particular: it is possible that unique groupings of individuals exist that are distinct from one another in terms of the kind of change made over the course of treatment (i.e. as discussed above, it is possible that treatment change is categorical). For example, some individuals might have a greater propensity to make change in the area of sexual deviance but less so in the area of emotional dysregulation, and vice versa. This propensity could in turn be linked to the individual causal mechanisms driving the offending behaviour for a given individual (a concept explored further in Study One). Although many possibilities of categorical structure exist, current literature that has attempted to categorise individuals based on the change they have made over treatment is generally limited to arbitrary dichotomous categorisations based on low or high levels of change.

For example, Marques and colleagues (2005) compared the recidivism rates of 259 male sex offenders (child and adult offenders) who had been randomly assigned to treatment with 225 controls who volunteered for treatment but were randomly assigned to the control group, and 220 controls who were eligible for treatment but chose not to participate. As part of a wider set of analyses, Marques and colleagues used a 9-point scale derived from a range of post-treatment measures to identify whether treated participants derived benefit from the programme, or “got” the treatment. A median split was then used to divide the treated participants into groups of those who “Got It” and those who “Did Not Get It”. Although there was no significant difference between recidivism rates for the two groups overall, the authors found that high-risk offenders who “Got It” reoffended at a significantly lower rate than high-risk offenders who “Did Not Get It”. They also found significant differences in recidivism for offenders against children who “Got It” compared with offenders against children who “Did Not Get It”. Overall, the study showed that there may be some benefit derived from a simple categorisation of treatment change or progress (particularly for high-risk offenders or offenders against children). However, as mentioned above, the categories in this study were determined by deciding arbitrary cut-off points on a continuous scale, and therefore may not be considered a “pure” categorical approach to measuring treatment change.

A similar approach was taken by Scalora and Garbin (2003), who compared recidivism rates between a treated sample of 76 men who had been convicted of child sexual offending, and 118 untreated controls. Based on discharge documentation regarding treatment involvement and attainment of intervention goals, men were categorised as either “successfully treated” (those who had achieved treatment goals and were recommended for less restrictive placements) or “unsuccessfully treated” (those who prematurely disengaged from treatment or who made limited treatment progress). Overall, the authors found that successfully treated offenders were significantly less likely to reoffend than unsuccessfully treated offenders and offenders who did not participate in treatment at all. This represents an approach to treatment change classification that is less reliant on continuous scores of progress, however the authors note that overall judgements of progress were heavily based on risk assessment information contained in participant files. It is also notable that the definitions or characteristics of each treatment group were arbitrarily determined (in this case, based on whether perceived progress was high or low), rather than being theoretically or empirically derived.

Findings in the literature do not always support the predictive accuracy of treatment change categories. Seager, Jellicoe and Dhaliwal (2004) categorised 109 treated male sex offenders as either “successfully” or “unsuccessfully” completing treatment, based on clinical evaluations of progress in key treatment components. The authors found that there was no significant difference in recidivism rates between the successful and unsuccessful participants. In another study of treatment progress, Quinsey, Khanna and Malcolm (1998) mixed a categorical with a dimensional approach by rating treatment change for 193 male sex offenders on a 4-point scale, from “poor” (few or very few gains, with areas still requiring intervention) to “very good” (significant gains in all targeted areas). They also found that there was no significant difference in recidivism rates by rated treatment change.

One notable feature of the two studies failing to find a significant association between treatment change group and recidivism was that ratings of progress in these studies were based purely on clinical judgement; as discussed in the previous chapter, clinical judgement is known to have lower predictive accuracy than more structured assessments of risk or change. Indeed, where studies have used more structured assessments to underlie the measurement or categorisation of change, stronger links between treatment change and offending are often discovered. This indicates that if treatment change were indeed found to be categorical in nature, it would be important that any methods developed to discriminate between categories are based on structured and empirical judgements of the factors of treatment change that are relevant to each of the treatment change groups identified. Identifying what these meaningful factors are for each group would be the first step in developing these kinds of treatment change measures.

Treatment Change: A Summary

The accurate measurement of treatment change is vital to modern approaches to offender management and intervention. Despite its importance in case management and decision-making (including parole decisions), relatively little attention has been paid to the predictive accuracy of treatment change in the literature, and to how treatment change might best be evaluated. From the relatively small number of studies that have been conducted, mixed findings are emerging regarding the meaningfulness of treatment change in terms of its association with reductions in offending. This provides a challenge to our current approach to offender treatment and intervention that cannot be ignored.

It is clear from the literature that the underlying nature of treatment change and how it should best be conceptualised and measured is still unclear. Although most of the literature treats change as if it were a dimensional construct, there are currently no studies that have explicitly tested this assumption. Furthermore, studies that have bucked this trend somewhat and treated change as a categorical construct typically employ an arbitrary and atheoretical approach to discriminating between treatment groups. This makes it difficult to interpret findings from these studies, particularly when there is no significant link found between treatment change groups and recidivism.

A lack of understanding of the underlying nature of treatment change may be resulting in misinformed or inappropriate methods of measuring change, potentially partially explaining the mixed results regarding a link between treatment change and recidivism. It is therefore important that we take a step back and focus on exploring treatment change in more detail if we are to progress our knowledge and improve our methods for assessing change.

The Current Research

The preceding review of the literature provides an overview of the current state of the literature regarding dynamic risk factors and their application to assessing offender change. The review highlights that there are a number of notable gaps in the literature and important concerns regarding both dynamic risk factors and the measurement of treatment change. The current thesis is an attempt at addressing some of these gaps, with the aim of supporting ongoing research and theory generation that can be used to improve our current approach to offender management and treatment. A brief overview of the studies that follow is provided below; note that full introductions (including rationale for each of the studies) are provided at the beginning of each study chapter.

Study One

Study One was a validation study of an influential theory of the aetiology of child sexual offending: Ward and Siegert's (2002) Pathways Model of sexual offending against children. The Pathways Model is arguably one of the more comprehensive explanatory theories of child sexual offending, incorporating proposed causal mechanisms behind each of the hypothesised pathways to sexually harmful behaviour. The validation of a causal theory of child sexual offending provides an important contribution to our understanding of how this harmful behaviour can be prevented and managed, and is a direct response to the issues raised in Chapter One regarding the current lack of theory-based or abductive research related to sexual offending. Exploring the causal mechanisms of sexually harmful behaviour provides vital knowledge regarding sexual offending as a phenomenon, and informs ongoing efforts to identify dynamic risk factors or criminogenic needs that are meaningful and causally linked to offending.

Study One utilised data obtained from a pre-treatment psychometric battery completed by 1,134 men convicted of sexual offences against children who engaged in an in-prison treatment programme. This data included a range of measures capturing psychological deficits hypothesised in the Pathways Model (including anti-social cognitions, sexual deviance, emotional dysregulation, and intimacy deficits), as well as information collected on the history and demographics of participants. The validity of the pathways predicted in the Pathways Model were tested by conducting latent profile analysis (LPA) using the psychometric battery scores. LPA is a statistical technique that aims to detect mutually independent groups of individuals that are meaningfully different from one another based on the variables used in the analysis. The aim of the study was to identify whether groups could be identified in the sample that represented the psychological profiles hypothesised for each of the pathways proposed by Ward and Siegert. Demographic variables and historical information were also assessed to explore fit with the hypotheses of the model.

Study Two

Study Two provided an exploration of the nature of change to identify whether treatment change is best categorised as a categorical or dimensional construct; this is the first study known to the author to assess this distinction. As the Pathways Model suggests, sexual offenders are generally accepted to be a heterogeneous group of individuals that differ in terms of their motivations for offending, and the mechanisms that maintain offending (Ennis et al., 2016). It seems to make intuitive sense that if different mechanisms underlie the sexually harmful behaviour that some individuals engage in, then needs targeted during treatment should also differ between individuals. This is indeed an underlying theory of the influential Risk-Needs-Responsivity (RNR) treatment model (Bonta & Andrews, 2016) and Good Lives Model (GLM; Ward, 2002). If distinct needs are causally related to offending for different individuals, and treatment is individually targeted based on these needs, then it is plausible that offenders may change in meaningfully different ways over the course of treatment too.

Despite this possibility, current approaches to the measurement of treatment change either treat change as a dimensional construct, or use arbitrary and atheoretical methods to categorise offenders based on perceived amounts of treatment progress. Study Two therefore provided a vital first step toward developing a stronger understanding of the nature of change, and therefore how change is most effectively and accurately measured. This was achieved by obtaining pre- and post-treatment scores on the same psychometric battery used in Study

One, for 346 men who had been convicted of sexually offending against children. Raw change scores were calculated by subtracting post-treatment psychometric scores from pre-treatment scores. To control for the impact of pre-treatment risk on overall change (Beggs & Grace, 2011), raw change scores were then converted into standardised residual change scores. The standardised residual changes scores were then used in a series of taxometric analyses to assess whether change is best conceptualised as categorical or dimensional in nature.

Study Three

Study Three extended the findings from Study Two (that treatment change is categorical) by using an exploratory approach to identifying categories of individuals who demonstrated similar types of treatment change to one another, but that are meaningfully distinct from other groups. Understanding the types of change demonstrated by the different groupings of individuals is important for developing measures that most appropriately assess change, and for developing theories of the mechanisms underlying change that are congruent with the categorical nature of change identified in Study Two. Identifying the most appropriate method for measuring and conceptualising change are in turn important for reliably and validly assessing any link between treatment change and recidivism, and has important implications for how change should be communicated and incorporated into risk assessment.

Study Three therefore explored whether the change made by individuals over the course of treatment could be used to categorise individuals into meaningful groups, and if so, what the patterns of change looked like in each of these groups. This study used the same approach to measuring change as Study Two (i.e. standardised residual change scores, based on a pre- and post-psychometric battery), however the study employed a larger sample of

1,170 men who had been convicted of sexually offending against children. An LPA was conducted on the standardised residual change scores to identify any meaningful treatment change groups in the sample. The change and recidivism patterns of the resultant groups were also assessed.

Study Four

Study Four extended the exploration of treatment change presented in Studies Two and Three by identifying whether there were key characteristics of individuals assigned to each of the treatment change groups that could help inform the development or validation of theories regarding the mechanisms underlying change. Where Studies Two and Three represented an exploration of the treatment change data to better understand treatment change as a phenomenon, Study Four was explicitly focussed on taking the next step in an abductive approach to science by examining what the research suggested in terms of causal theories to explain that phenomena (Haig, 2005). A further aim of Study Four was to replicate the findings of Study Three using a separate measure of treatment change, to ensure that the findings were robust, reliable and generalisable.

To replicate the findings of Study Three, Study Four involved scoring the VRS-SO for a sub-sample of 292 men who had been included in the previous study's sample. The change profiles on the VRS-SO for this sub-sample were then compared to the change profiles expected based on their Study Three grouping, to assess whether change profiles could be replicated across different change measurement methods. This attempted replication was then followed by an assessment of the risk, criminogenic needs, and historical information of individuals across treatment groups, using the same full sample of 1,170 men included in Study Three. It was hoped that exploring these psychological and environmental factors would provide important information that could be used to link the findings from

Studies Two and Three to possible explanatory theories of the mechanisms underlying the new conceptualisation of treatment change identified in the previous studies.

Chapter Three/Study One: An Empirical Test of Ward and Siegert's Pathways Model of Sexual Offending against Children

Understanding the etiology of sexual offending against children is of great value to researchers and clinicians, and can inform prevention efforts targeting at-risk individuals before their first offense. As such, there have been a number of notable attempts to develop etiological theories of sexually harmful behavior against children. However, a disconnect is noticeable between the development of these theories and resulting empirical work to test their accuracy and apply findings to our conceptualisation of etiology and risk (Ward, 2014), which could partially be explained by the data-driven approach to much of the research on dynamic risk factors or criminogenic needs, as discussed in Chapter One. The current study aims to address this gap in the literature by testing the key hypotheses made by one promising theory: the Pathways Model of sexual offending against children.

The Pathways Model

The Pathways Model (Ward & Siegert, 2002) is arguably one of the more comprehensive models developed for explaining the etiology and maintenance of sexually harmful behavior against children, due to its multifactorial nature, explanatory depth, and the inclusion of developmental causal mechanisms as part of the model (T. Gannon, Terriere, & Leader, 2012). The Pathways Model proposes that the traits and behaviors displayed by individuals who sexually offend against children are caused by the interaction between four core sets of dysfunctional mechanisms (or vulnerabilities): intimacy/social skill deficits, emotional dysregulation, antisocial cognitions, and deviant sexual scripts. These vulnerabilities are proposed to arise from complex interactions between a range of biological, cultural, and environmental/situational factors.

The Pathways Model proposes that interaction between these primary vulnerabilities results in a number of distinct etiological pathways that lead to sexually harmful behavior. Each of these pathways is characterized by a core vulnerability that drives the sexually harmful behavior and is responsible for generating the cluster of psychological and behavioral characteristics unique to that pathway. The five hypothesized pathways are briefly outlined below.

Pathway one: intimacy deficits. The first etiological pathway describes individuals who experience a primary or core dysfunction in their social skills and ability to form intimate relationships with other adults. The intimacy deficits experienced by these individuals are hypothesized to be caused by insecure attachment patterns, which result in unsatisfactory or dysfunctional adult relationships and related feelings of intense loneliness. Adults are the preferred sexual and intimate partners for individuals in this pathway, however difficulties with establishing intimacy with age-appropriate partners means that they might be prepared to substitute children into this partner role, viewing them as a “pseudo-adult”. These individuals are likely to focus their sexual needs and desires on children as well, thus leading to the sexually harmful behavior.

Pathway two: deviant sexual scripts. The second etiological pathway includes individuals whose primary dysfunctions are associated with their sexual scripts; namely, these individuals possess distortions in their representation of appropriate contexts in which sex is sought. The cause of these distortions is theorized to be in premature sexualisation resulting from sexual abuse experienced as children. For these individuals sexual activity is seen as equivalent to intimacy, leading to a desire for impersonal sex, and a high number of short-term relationships or overt promiscuity. Children are chosen as sexual partners by these individuals largely as a result of opportunity or sexual need, rather than being a preferred choice of partner.

Pathway three: emotional dysregulation. The core vulnerability for individuals in the third etiological pathway is a lack of emotional competence. Although this emotional dysfunction may present in a range of ways, it is hypothesized that there are two main issues experienced by these individuals: under-regulation of emotions, and the use of maladaptive strategies to cope with negative emotions. Under-regulation of emotions describes an inability to control emotional states, which then leads to a loss of control over behavior. For example, an individual who is unable to control their anger proactively may sexually abuse children in order to "punish" the parent of the child. The second main dysfunction relates to the use of maladaptive coping strategies, including sex, to manage negative mood states. For these individuals, sex becomes intrinsically tied with well-being, possibly because of excessive masturbation during early adolescence and the absence of other strategies to improve self-esteem and mood. Thus, when negative moods such as anger, loneliness or anxiety are experienced, these individuals seek sex in order to escape these mood states.

Pathway four: antisocial cognitions. The fourth etiological pathway includes individuals with a primary vulnerability relating to the possession of beliefs or attitudes that support antisocial behavior in general. These individuals have normal sexual scripts, and their sexually harmful behavior is typically seen within the context of a wider pattern of criminal behavior, including property and violent offenses. This pattern of antisocial and/or criminal behavior is also likely to have been demonstrated in childhood and adolescence. For these individuals, children are selected as the target of their sexual desires as a result of their general exploitation of any opportunity for self-gratification.

Pathway five: multiple dysfunctional mechanisms. The final etiological pathway outlined in the Pathways Model contains individuals who display difficulties with all four of the primary vulnerabilities outlined above. It is hypothesized that at the core of this pattern of multiple dysfunction are distorted sexual scripts driving both the general dysfunction and the sexually harmful behavior. It is hypothesized that early sexualisation resulting from experiences of sexual abuse or exposure to sexual materials from a young age has led to a sexual preference for children or young people. It is also hypothesized that these individuals would display additional deviant sexual behavior, such as sadism or bestiality.

Validations of the Pathways Model

The comprehensiveness of the Pathways Model both in terms of the behaviors explained and the extent to which it incorporates previous models and literature is a major strength of the model. Because of its potential to increase our knowledge of psychologically meaningful risk factors, independent validation of the model is important (Gannon et al., 2012). To date, there have been three published empirical assessments of the Pathways Model; each of these is outlined below.

Connolly (2004) used a qualitative approach to identify whether the early experiences and offending patterns of men convicted of sexual offences against children were congruent with the profiles hypothesized by the Pathways Model. Unstructured interviews were conducted with 13 New Zealand men incarcerated for sexual offenses against children. An inductive thematic analysis of these interviews revealed that a majority of the men (10 out of 13) presented with profiles that were reasonably congruent with the hypothesized profile from one of the pathways, however Connolly was unable to identify any profiles that matched the hypothesized Antisocial Cognitions pathway.

Middleton, Elliott, Mandeville-Norden, & Beech (2006) conducted an empirical study to investigate the applicability of the Pathways Model beyond contact offending, focussing instead on Internet offenders. A sample of 72 males who had been convicted of an index offense involving child pornography completed a range of psychometrics selected to measure the hypothesized core vulnerabilities for each pathway. Individuals were then manually assigned into pathways according to relative elevations in their scores. Using this method, Middleton et al. were able to assign 33 out of 72 individuals (46%) to one of the hypothesized pathways; all of the five hypothesised pathways were identified as matching the profile of at least individual in this study.

Although both these studies provided tentative support for the Pathways Model, there were several issues that required further investigation, including the possibility of additional etiological pathways beyond those originally proposed. Thus Gannon et al. (2012) conducted an exploratory study attempting to validate the Pathways Model with a sample of 97 men convicted of contact sexual offenses against children. Participants completed a number of psychometrics that measured characteristics related to the key vulnerabilities hypothesized in the model. A non-hierarchical k-means cluster analysis was then used to identify meaningful groupings of individuals on the basis of their psychometric scores; this allowed for the possibility of discovering unique clusters of individuals that differed from the original five pathways proposed in the model. As hypothesized, a five-cluster solution provided the best fit for the data, with the psychological profiles of three of these clusters matching those hypothesized for the Intimacy Deficits (*n* = 12), Antisocial Cognitions (*n* = 12), and Multiple Dysfunction (*n* = 4) pathways. However, they failed to identify clusters that matched the Emotional Dysregulation and Deviant Sexual Scripts pathways. Instead, they identified one cluster they named Boy Predators (*n* = 18) because of their selective sexual interest in male children, and another they named Impulsive (*n* = 49) due to a small elevation in impulsivity scores; of note, this latter group showed relatively unremarkable scores on most other measures.

Overall, research that has attempted to test the Pathways Model has obtained mixed results. However, prior studies have been limited by small sample size, with only Middleton et al. (2006) identifying clusters with more than 15 individual members. This severely limits the ability to further validate the identified clusters using corroborating historical or offense-related characteristics. In addition, only Gannon et al. (2012) tested for the possibility of pathways additional to those presented in the original model. Given that two additional pathways were identified, it is important that this more exploratory approach to model testing is replicated independently in order to identify whether these additional pathways are evident in other samples.

The aim of the present study was to test Ward and Siegert's (2002) Pathways Model using a similar approach to Gannon et al. (2012), except with a larger sample. In this way, we hoped to obtain clusters of sufficient size to allow for post-hoc analyses to further define the characteristics of the group members. We used Latent Profile Analysis (LPA) to extract our groupings from the pre-treatment psychometrics, as opposed to the k-means cluster analysis used by Gannon et al. (2012); the benefits of this technique are outlined below. Based on prior results and predictions of the Pathways Model, we predicted that five distinct groups of individuals would be obtained from the data, with each of these corresponding to one of the five pathways described by the Pathways Model.

Method

Participants

All men (N = 1,474) who had participated in a high-intensity, prison-based treatment programme for sexual offenders against children in New Zealand between 1990 and 2007 were identified. For 280 men the available file information on pre-treatment psychometrics was insufficient, and thus the final sample size was N = 1,134; this final sample includes all participants included in the Allan et al. (2007) paper. The majority of the sample (69.4%) identified as being of European descent, with 23.6% as New Zealand Māori, 5.2% as Pasifika, and 1.8% as other ethnicities. The majority of the final sample (*n* = 750; 66.1%) had participated in treatment at the Kia Marama Special Treatment Unit in Rolleston, New Zealand, with the remaining 384 men (33.9%) participating at Te Piriti Special Treatment Unit, Auckland, New Zealand (based on the same treatment model as Kia Marama). Because the data was extracted from psychometrics administered pre-treatment, treatment non-completers were eligible for inclusion in the sample. All men had provided written consent for their information to be used for research and evaluation purposes prior to the commencement of assessment and treatment. Ethics approval was obtained from the authors' University ethics board prior to the research commencing.

Psychometric battery

The measures used in this study were completed by participants as part of a psychometric battery administered by clinical staff at the beginning of treatment. Because of the continuing evolution of the treatment programmes and the development or updating of psychometrics over time, the set of psychometrics completed by each participant was not identical across the sample. For this reason multiple psychometrics were used to measure each key psychological deficit outlined in the Pathways model. The use of different scales also ensured that multiple facets of each of the hypothesized deficits were able to be adequately captured. Brief descriptions of the self-report measures used to capture each deficit are provided below; for further information on the psychometric properties of these measures, see Allan et al. (2007).

Anti-social cognitions

The Abel-Becker Cognitions Scale (ABCS; Abel et al., 1989) was used to measure distorted attitudes and beliefs about sexual offending against children. Respondents indicate the extent to which they agree with 29 pro-paedophilic statements. Agreement is rated on a 5-point scale (1 = strongly agree, 5 = strongly disagree), such that lower scores are indicative of greater levels of support of sexually assaultive behavior against children. Example items include "A child who doesn't physically resist an adult's sexual advances really wants to have sex with the adult" and "I show my love and affection to a child by having sex with her (him)".

The Hostility Towards Women scale (HTW; Check, 1985) was used to measure negative beliefs about women, including the acceptance of aggressive motivations and behaviors directed at women. Responses for the 30 scale items are scored true/false, with higher scores being indicative of greater levels of aggression towards women. Example items include "It is safer not to trust women" and "I feel upset even by slight criticism by a woman".

The Rape Myth Acceptance Scale (RMAS; Burt, 1980) was used to measure attitudes supportive of sexual violence and aggression. The scale consists of nineteen statements about rape; the respondent's agreement with each statement is rated on a 7-point scale such that higher scores are indicative of greater endorsement of rape myths. Sample items include "Any healthy woman can resist rape if she really wants to" and "When women go around braless or wearing short skirts and tight tops, they are just asking for trouble".

Deviant sexual scripts

Wilson's Sex Fantasy Questionnaire (WSFQ; Wilson, 1978) was used to measure the frequency or strength of different types of sexual fantasies. Four sub-scales comprised of 10 items each are used to rate frequency of different types of fantasies, including intimate themes (e.g. sex with a partner), exploratory themes (e.g. group sex), impersonal themes, (e.g. sex with a stranger), and sado-masochistic themes (e.g. sex involving pain or use of force). Items are rated on a six-point frequency scale from 0 to 5, allowing for subscale totals to range between 0 and 50.

Emotional Dysregulation

The State-Trait Anxiety Inventory (STAI; Spielberger, 1983) was used to measure general anxiety (T-scale) and current anxiety (S-scale). Both subscales consist of 20 items rated on a 4-point scale, with subscale totals ranging between 20 and 80.

The State-Trait Anger Expression Inventory (STAXI; Spielberger, 1988) was used to measure several aspects of anger and anger expression. The STAXI consists of 44 items grouped into five major subscales rated on a 4-point frequency scale: (1) State Anger (10 items) – the current intensity of anger; (2) Trait Anger (10 items) – level of general tendency towards experiencing anger; (3) Anger-in (8 items) – the degree to which anger is internally supressed; (4) Anger-out (8 items) – the degree of outward expression of anger towards others, either verbally or physically; and (5) Anger Control (8 items) – the degree to which expressions of anger are controlled. Because of continual revising of the assessment battery over time, some participants in our sample completed the revised version of the STAXI rather than the original measure. The STAXI-2 (Spielberger, 1999) is largely similar to the STAXI, however there are differences in the overall number of items (57 in the STAXI-2), and there are a total of 6 major subscales: (1) State Anger (15 items); (2) Trait Anger (10 items); (3) Anger Expression-In (same as Anger-in in STAXI; 8 items); (4) Anger Expression-Out (same as Anger-out in STAXI; 8 items); (5) Anger Control-Out (8 items) – the amount to which outwards expression of anger is controlled; and (6) Anger Control-In (8 items) – the amount to which feelings of anger are controlled through calming down or cooling off.

Intimacy Deficits

The Revised UCLA Loneliness Scale (UCLS; Russell, Peplau, & Cutrona, 1980) was used to measure experiences of loneliness. The scale consists of twenty statements rated on a 4-point scale relating to perceived satisfaction or dissatisfaction with interpersonal relationships, with higher scores indicating greater dissatisfaction.

The Fear of Intimacy Scale (FIS; Descutner & Thelen, 1991) was used to measure anxiety about intimate dating relationships. The scale comprises 35 items rated on a 5-point scale, with individuals rating how characteristic the items are of them. Higher scores indicate a greater fear of intimacy.

The Assertion Inventory (AI; Gambrill & Richey, 1975) was used to measure degree of discomfort in situations requiring assertiveness (e.g. turning down a request for a meeting or date), and an individual's likelihood of making an assertive response in these situations. Forty items are rated twice on a 5-point scale, first for level of discomfort (AI-D) and secondly for response probability (AI-RP). Only AI-RP scores were available for the current sample. Higher scores are indicative of lower response probability.

The Social Self-Esteem Inventory (SSEI; Lawson, Marshall, & McGrath, 1979) was used as a measure of self-esteem in social situations. The SSEI consists of 30 items rated on a 6-point scale, with higher scores indicating greater self-esteem.

Socially Desirable Responding

The Marlowe-Crowne Social Desirability Scale (M-CSD; Crowne & Marlowe, 1960) was used to assess the tendency to respond in ways that are anticipated to gain approval from others, or to refrain from responding in ways that are likely to gain disapproval. The M-CSD consists of 33 true/false items, with higher scores indicating a greater tendency to response in socially desirable ways.

Procedure

Data for the current sample was extracted from a database held by the New Zealand Department of Corrections, containing scores from a psychometric battery completed pre-and post-treatment by all individuals engaging in the treatment programme, as well as demographic information collected prior to treatment entry.

In total, there were 1,474 cases initially extracted from the database. The data was then cleaned and checked for input errors; any scores that fell outside of the possible minimum and maximum scores for a given measure were deleted and were thereafter considered missing for that individual. Cases with unacceptably high levels of missing data – defined as those that did not have at least one completed measure from each of the four domains outlined above - were excluded from the study. This left a final sample of 1,134 men. Information on how missing data was handled is outlined in the following section.

Because two versions of the STAXI (the STAXI and the STAXI-2) were used in the database, raw scores were transformed prior to analysis in the following manner. For individuals who completed the STAXI-2 (n = 168) rather than the STAXI (n = 966), totals for Anger Control-Out and Anger-Control-In were added together to provide one overall Anger Control score. This meant that the number of subscales for each measure was equivalent for the two versions. All scores on individual subscales were then standardized for both versions of the measure, to control for the differences in number of items per subscale. In order to retain consistency across all measures, raw scores on all remaining psychometrics were also standardized before being used in subsequent analyses.

Planned Data Analysis

In order to test the validity of the pathways predicted in the Pathways Model, a latent profile analysis (LPA) was conducted with MPlus 7.4 software, using the pre-treatment psychometric scores; full information maximum likelihood estimation was used for cases with missing data. Latent profile analysis is a statistical technique that aims to detect mutually independent groups of individuals that are qualitatively distinct from one another, using two or more indicator (observed) variables (Collins & Lanza, 2010; Francis, Bowater, & Soothill, 2004). In the current study, LPA was used with the goal of identifying whether there were distinct categories of offenders within the sample that displayed pre-treatment psychometric profiles that were consistent with those predicted by the Pathways Model. Compared to similar statistical techniques such as cluster analysis or factor analysis, LPA is considered to be better suited to research in the social sciences because it is less restrictive in terms of assumptions (e.g., normality) and provides estimates of standard errors for fitted models (Fox & Farrington, 2016). Furthermore, rather than identifying latent groups on the basis of arbitrary distance measures (as with cluster analysis), LPA first builds a model that describes the distribution of the observed data, and then assigns individuals probabilities of belonging to each of the latent classes identified (Hagenaars & McCutcheon, 2009). Latent Profile Analysis could therefore be considered more of a “top-down” abductive approach to class detection, and is therefore more likely to result in the identification of theoretically meaningful latent classes than is cluster analysis (Haig, 2014).

Following the LPA, men were assigned to groups according to their highest level of membership probability indicated by the model. Analyses (ANOVAs and cross tabulations) were then run with SPSS 23 software to identify statistical differences between the groups in terms of pre-treatment psychometric needs, as well as with regards to demographic variables and static risk factors available on file.

Results

Latent Profile Analysis

A series of one- to seven-class models was tested with LPA. Identifying the optimum number of classes for a given sample involves several factors, including goodness-of-fit criteria, parsimony, interpretability of results, and theoretical expectations. The goal is to determine the number of classes whereby all groups are distinct, but the addition of an extra class does not provide additional explanatory power (Fox & Farrington, 2016). There is currently no consensus on the best fit criteria for statistically identifying the optimum number of classes, and as such, most researchers use a variety of fit indicators for this task (Nylund, Asparouhov, & Muthén, 2007; Tein, Coxe, & Cham, 2013). For the current study, the Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and sample-size adjusted BIC (SBIC; Sclove, 1987) were used as initial indicators of model fit (lower values indicate better fit). As Table 1 shows, AIC, BIC and SBIC values continued to decrease with each additional class up to the seven-class solution, indicating that additional goodness-of-fit criteria needed to be assessed in order to determine optimum model fit.

Table 1. Fit indices and intropy for all class solutions

Class Solution	Loglikelihood	AIC	BIC	SBIC	Entropy	LMRT
1 class	-27357.14	54786.27	54967.48	54853.13	-	-
2 classes	-25910.85	51931.70	52208.54	52033.84	0.84	<.001
3 classes	-25290.04	50728.08	51100.56	50865.51	0.85	.006
4 classes	-24868.23	49922.45	50390.57	50095.18	0.87	.002
5 classes	-24624.05	49472.10	50035.85	49680.11	0.87	.177
6 classes	-24422.84	49107.67	49767.06	49350.97	0.86	.283
7 classes	-24204.12	48708.24	49463.27	48986.82	0.87	.226

Note: AIC = Akaike information criterion; BIC = Bayesian information criterion; SBIC = sample-size-adjusted BIC; LMRT = Lo, Mendell, Rubin test.

Thus we used two additional indicators, the Lo-Mendell-Rubin test (LMRT; Lo, Mendell, & Rubin, 2001) and the bootstrap likelihood ratio test (BLRT; McLachlan & Peel, 2000). These are both likelihood ratio tests that compare the relative fit of two models, namely a model with k classes versus a model with k-1 classes. A statistically significant result indicates that the model with the greater number of classes provides a significantly better fit with the observed data. As shown in Table 1, the results of the LMRT indicated that model fit was improved by the addition of each class up to the 4-class solution, with additional classes after that point no longer providing a significant improvement to model fit. Because a 5-class solution was expected on a theoretical basis, we sought to confirm this result by applying the BLRT to both the 4-class and 5-class solutions. In both cases, the BLRT returned a statistically significant result, indicating that the 5-class solution provided a significantly better fit to the observed data than the 4-class solution. Previous simulation studies have indicated that the BLRT outperforms the LMR in identifying the optimum number of classes in an LPA (Nylund et al., 2007), suggesting that five classes may be a better fit to our data. Entropy levels were nearly equal for both the 4-class and 5-class solutions, indicating that the models did not differ substantially in how distinct the classes were.

Because the fit indicators did not indicate a clear consensus regarding the optimum number of classes, both the 4-class and 5-class solutions were selected for closer examination on the basis of interpretability of results and parsimony with existing theory; agreement with extant theory is an important consideration used widely in the literature to identify an optimum LPA solution (Nylund et al., 2007). Although both solutions resulted in groupings of individuals that showed distinctive psychological profiles (and thus were both readily interpretable), the 5-class solution resulted in groups of individuals that closely aligned with those expected from the previous literature, including the pathways hypothesized by Ward & Siegert (2002). Each class in the 5-class solution contained at least 100 men, also indicating that the addition of the extra class was producing a meaningful distinction between large groups of individuals. In addition, the previous cluster analysis conducted by Gannon et al. (2012) identified a 5-class solution as being optimum for their psychometric data. For these reasons, it was decided that the 5-class solution was the best fit for our data, however it is important to note that a 4-class solution could also have been viable. All groups extracted in the 4-class solution were also extracted in the 5-class solution; a class containing individuals with elevated scores on the emotional dysregulation measures was the additional class extracted in the 5-class solution. This indicates that selecting the 5-class over the 4-class solution does not have an appreciable impact on the conclusions of the study.

Table 2 shows the number of cases scored on each measure, the average raw score for each measure by cluster, and an indication of the significant differences between clusters in terms of these scores. In order to investigate the possibility that estimations of the model for cases with missing data affected our results, the LPA was repeated with a sample of complete cases only (n = 691). Excluding cases with missing data did not substantively change the results or conclusions, so the following analyses were completed with the larger sample.

Fit with the Pathways Model

Each of the five classes extracted from the LPA was manually assessed for overall fit with one of the Pathways outlined by Ward & Siegert (2002). This was completed with reference to predictions made by the authors prior to data analysis regarding the expected z-scores for each Pathway; these predictions were generated from the hypotheses about the characteristics of individuals within each of the Pathways. A prediction framework was developed such that a z-score of +1 (i.e., one standard deviation above the sample mean) was predicted for measures that represented primary characteristics in the outlined pathways, and a z-score of +0.5 was predicted for secondary characteristics. Figure 1 displays the pairings made between obtained classes and hypothesized pathways graphically, showing the prior predictions for each pathway together with the obtained average z scores on the pre-treatment measures for the matched LPA class. Scores for the ABCS, STAXC and SSEI were reversed such that higher scores were indicative of higher dysfunction in those areas. Figure 1 shows that all pathways were able to be matched to one of the LPA classes, except for the Antisocial Cognitions pathway.

Table 2. Mean unstandardized scores for all measures, by class

Measure	*n*	Class 1: Low Needs Mean (SD)	Class 2: Deviant Sexual Scripts Mean (SD)	Class 3: Intimacy Deficits Mean (SD)	Class 4: Emotional Dysregulation Mean (SD)	Class 5: Multiple Dysfunction Mean (SD)	*F*	η^2
Antisocial cognitions								
Abel-Becker cognition	1125	129.73 (12.50)	121.36^{ab} (15.83)	123.93^{a} (14.97)	116.31^{bc} (17.52)	115.36^{c} (17.88)	30.13	0.10
Hostility toward women	1097	6.91 (4.43)	10.80 (5.08)	12.23 (5.23)	15.74 (5.94)	17.74 (6.09)	127.98	0.32
Rape myth acceptance	1093	38.97 (14.76)	49.07^{a} (18.77)	48.80^{a} (17.43)	55.74^{b} (20.77)	56.62^{b} (16.75)	35.73	0.12
Deviant sexual scripts								
WSFQ - exploratory	1123	6.78^{a} (5.27)	22.19^{b} (7.09)	7.63^{a} (5.35)	15.20 (9.49)	20.91^{b} (7.35)	272.22	0.49
WSFQ - intimate	1131	21.17^{a} (10.23)	34.88 (7.59)	19.86^{a} (9.49)	27.72 (11.40)	31.64 (7.48)	99.53	0.26
WSFQ - impersonal	1131	7.10 (4.75)	22.36 (7.16)	8.66 (4.89)	15.55 (8.05)	19.68 (6.83)	283.48	0.50
WSFQ - sado/masochistic	1108	1.71 (2.88)	10.74^{a} (7.82)	2.77 (3.77)	8.64^{b} (8.04)	9.32^{ab} (7.46)	127.07	0.32
Emotional dysregulation								
State anxiety	1118	30.08 (8.36)	34.11 (8.54)	43.65 (11.47)	50.88^{a} (11.81)	50.28^{a} (11.59)	162.06	0.37
Trait anxiety	1114	33.01 (7.55)	40.13 (7.74)	48.09 (8.36)	53.23 (9.84)	56.71 (7.82)	294.36	0.51
State anger	861	11.31^{a} (2.52)	11.48^{a} (2.47)	12.57 (3.28)	29.50 (4.66)	13.85 (3.59)	566.28	0.73
Trait anger	863	14.59 (3.56)	18.10^{a} (5.12)	18.26^{a} (4.75)	27.56 (6.14)	23.58 (6.11)	140.25	0.40
Anger expression	859	13.61 (3.08)	15.86^{a} (3.68)	15.50^{a} (3.69)	20.98 (4.31)	17.43 (4.80)	65.89	0.24
Anger supression	842	14.40 (3.34)	17.60 (3.90)	18.91 (3.75)	21.69^{a} (3.99)	23.17^{a} (4.08)	128.49	0.38
Anger control	850	25.96 (5.05)	22.00^{a} (5.39)	21.27^{ab} (5.05)	19.37^{c} (5.15)	19.96^{bc} (4.70)	47.86	0.18
Interpersonal deficits								
Social self esteem	1113	132.59^{a} (23.59)	126.66^{a} (22.88)	104.37^{b} (21.28)	105.51^{b} (23.31)	93.40 (22.91)	115.67	0.29
Assertion - response prob.	990	102.55 (20.86)	112.51^{a} (17.44)	118.65^{bc} (19.80)	114.38^{ab} (16.90)	124.22^{c} (18.78)	38.33	0.13
Fear of intimacy	974	79.31 (21.28)	92.04 (17.84)	102.78^{a} (18.14)	104.43^{a} (17.20)	113.20 (20.22)	93.06	0.28
UCLA loneliness	1059	36.10 (7.08)	42.94 (7.98)	49.74^{a} (7.79)	49.18^{a} (7.74)	55.05 (8.00)	204.60	0.44

Note. All *F* values significant at the $p < .001$ level. Groups that share superscripts are not significantly different from one another using Tukey' HSD post-hoc tests ($p < .05$).

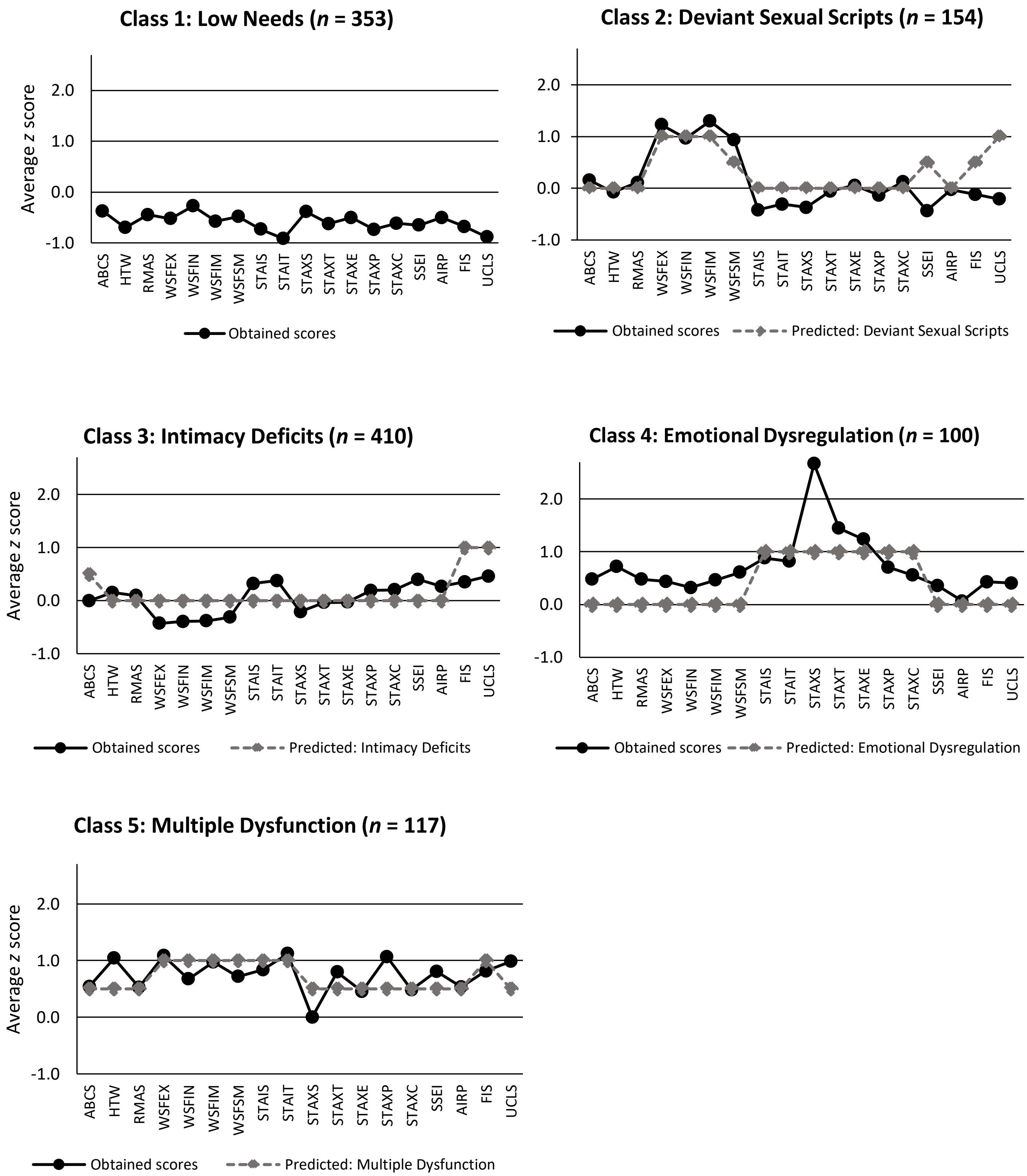

Figure 1. Average *z* scores obtained on each measure for the 5 LPA classes, with theoretical predictions for each pathway

To assess how accurately the LPA classes were matched with pathways, we calculated a measure of goodness of fit as follows. First, the mean squared deviation between the obtained z scores for each participant and the predicted z scores for each pathway was calculated; because there was no match found for the Antisocial Cognitions pathway, predictions for Antisocial Cognitions were not included in the analysis. This resulted in four mean squared deviations obtained for each participant, representing the relative fit of the four different pathways to the obtained psychometric profile for each individual. The pathway that resulted in the lowest mean squared deviation for each case was selected as being the statistical "best" fit for that individual, as this was the pathway in which the differences between obtained and predicted scores were lowest; we labelled this the "Best Fit Pathway". The level of agreement between this "Best Fit Pathway" and the pathway matched to each individual based on their LPA grouping (the "LPA Pathway") was then assessed using Cohen's Kappa. This provided a way of assessing the fit between the hypothesized characteristics of each pathway, and the LPA group that we matched each of these pathways to (as shown in Figure 1). Table 3 shows the number of individuals who were assigned to each LPA Pathway compared with their Best Fit Pathway. The level of agreement between the Best Fit Pathway and the LPA Pathway was found to be Kappa = 0.59, $p < .001$, 95% CI [0.55, 0.64], indicating a moderate level of agreement between the Best Fit Pathway and the LPA Pathway; 49.6% of cases were assigned to pathways that matched their Best Fit Pathway. This suggests a reasonably good fit between the class matched to each pathway and the prior predictions made for these pathways. Thus, our hypothesis was partially supported in that five groups were extracted from the LPA, however only four of these were able to be reliably matched with a hypothesized pathways group.

Table 3. Number of individuals in obtained LPA Pathways compared with predicted Best Fit Pathway

	LPA Pathway				
	Deviant Sexual Scripts	Intimacy Deficits	Emotional Dysregulation	Multiple Dysfunction	**Total**
Best Fit Pathway					
Deviant Sexual Scripts	108	11	1	2	**122**
Intimacy Deficits	17	305	5	10	**337**
Emotional Dysregulation	4	89	59	15	**167**
Multiple Dysfunction	25	5	35	90	**155**
Total	**154**	**410**	**100**	**117**	**781**

Description of Classes

The typical psychometric profile for each of the obtained classes is outlined below.

Class 1: Low Needs (*n* = 353). Individuals who were grouped within Class 1 displayed relatively low scores on all pre-treatment psychometrics, with all scores falling below the mean for the sample as a whole. Particularly low scores (approximately -1.0 SD) were observed on measures of loneliness and trait anxiety. This group was therefore labelled as "Low Needs"; there was no clear pathways match for this class, as all predicted pathways were hypothesized to be linked to at least one primary vulnerability. The Low Needs class comprised the second-largest of all identified classes.

Class 2: Deviant sexual scripts (*n* = 154). Individuals in Class 2 demonstrated elevated scores (approximately +1.0 SD) on all sub-scales of the WSFQ, consistent with predictions made for the Deviant Sexual Scripts pathway. The elevation in scores on the sado-masochistic subscale of the WSFQ was slightly higher than initially predicted. Notable departures from the predictions made for this pathway were scores on the SSEI, FIS and UCLS falling below the average despite being predicted to be either secondary or primary dysfunctions for this group. Slightly lower-than-average scores on measures of anxiety and state anger were also noted for this group.

Class 3: Intimacy deficits (*n* = 410). This class comprised the largest group of the five identified classes, and most psychometric scores for the individuals in this group fell around the mean for the entire sample. There were small elevations noted in measures of interpersonal dysfunction, including the FIS, and this class was therefore matched with the intimacy deficits pathway hypothesized by Ward and Siegert. Notably, scores relating to sexual fantasising were relatively low in this group (around -0.4 SD), and scores on the ABCS were approximately equivalent with the mean of the sample. Individuals in this class also displayed slight elevations (approximately +0.3 SD) in measures of state and trait anxiety.

Class 4: Emotional dysregulation (*n* = 100). Individuals in Class 4 displayed small elevations (approximately +0.5 SD) on all measures aside from assertiveness. This group had the smallest number of cases, and was characterized by extremely high levels (+2.7 SD) of state anger, as well as scores of around +1.0 SD on all other measures of emotional dysfunction (including both anger and anxiety). For this reason, this class was matched with the hypothesized Emotional Dysregulation pathway. Although elevations on other psychometrics were not predicted in the initial description of this pathway, the level of dysfunction in emotional regulation relative to other measures justified this match. As mentioned above, this class was not extracted in the 4-class LPA solution.

Class 5: Multiple dysfunction (*n* = 117). Individuals in Class 5 were characterized by elevations of approximately +0.5 SD to +1.0 SD on all measures aside from state anger, in which they demonstrated average scores. More specifically, individuals in this class demonstrated relatively high levels of dysfunction on measures of anti-social cognition, sexual interest, emotional regulation, and interpersonal functioning. Because of this steady elevation of scores in all domains of dysfunction, this class was matched with the Multiple Dysfunction pathway hypothesized by Ward and Siegert. This class demonstrated a similar profile to those in the Emotional Dysregulation class, however individuals in the Multiple Dysfunction class generally had higher levels of dysfunction on all measures aside from those that reflected poor emotional functioning, for which the Emotional Dysregulation class showed a more prominent elevation in scores.

Analysis of Static Factors

In order to test the further predictions made by Ward and Siegert regarding developmental mechanisms for each of the pathways, as well as providing further meaningful distinctions between the classes, we performed a series of multivariate analyses using static factors that had been scored by therapists upon entry to treatment. These included items relating to offence history (e.g., number of previous offences and number of victims), personal history (e.g. whether abuse had been experienced as a child, and whether a paraphilia had been diagnosed), and victim profiles (e.g., gender and relationship of previous victims). Note that these items were omitted from the LPA used to identify the different classes, and therefore provided a further test of the classes extracted and matched to each pathway. Pairwise deletion was used for cases with missing variables; this meant that the number of cases differed between analyses. The results of these analyses (and the number of cases used for each analysis) are presented in Table 4, and the main findings are summarized below.

Table 4. Comparison of classes on historic and victim-related factors

	n	Statistic	*p*	Class 1: Low Needs	Class 2: Deviant Sexual Scripts	Class 3: Intimacy Deficits	Class 4: Emotional Dysregulation	Class 5: Multiple Dysfunction
				M (SD)	*M* (SD)	*M* (SD)	*M* (SD)	*M* (SD)
M-CSD	1111	$F = 50.22$	<.001	19.64[abc] (5.58)	16.54[ac] (5.38)	16.24[bc] (5.68)	13.55[ab] (5.01)	12.57[ab] (4.81)
Offence history								
No. of prior sex offence	1006	$F = 4.75$	.001	0.60[a] (1.53)	1.16 (3.48)	1.05[b] (2.90)	1.14 (2.55)	2.13[ab] (5.95)
No. of prior major offence	984	$F = 2.19$	.069	3.75 (15.72)	7.82 (24.90)	6.76 (39.38)	12.38 (27.84)	14.79 (73.44)
No. of victims	1007	$F = 8.71$	<.001	3.75[a] (5.40)	8.40[b] (20.29)	5.75[c] (17.75)	9.72 (19.81)	15.60[abc] (36.99)
Personal history				% (*n*)	% (*n*)	% (*n*)	% (*n*)	% (*n*)
Was sexually abused	1005	$\chi^2 = 23.24$	<.001					
Yes				48.9 (150)	56.8 (79)	61.6 (226)	74.7 (68)	60.4 (61)
No				51.1 (157)	43.2 (60)	38.4 (141)	25.3 (23)	39.6 (40)
Offending pre-adulthood	1010	$\chi^2 = 29.10$	<.001					
Yes				30.2 (93)	50.4 (70)	37.1 (137)	48.4 (44)	52.4 (54)
No				69.8 (215)	49.6 (69)	62.9 (232)	51.6 (47)	47.6 (49)
Has a paraphilia	1020	$\chi^2 = 38.75$	<.001					
Yes				22.3 (69)	45.4 (64)	23.8 (89)	42.4 (39)	32.7 (34)
No				77.7 (240)	54.6 (77)	76.2 (285)	57.6 (53)	67.3 (70)
Victim profiles								
Relationship	923	$\chi^2 = 24.14$	.002					
Related only				55.4 (155)	48.1 (63)	63.6 (211)	49.4 (41)	42.3 (41)
Extrafamilial only				32.1 (90)	34.4 (45)	25.9 (86)	42.2 (35)	41.2 (40)
Mixed				12.5 (35)	17.6 (23)	10.5 (35)	8.4 (7)	16.5 (16)
Gender	1009	$\chi^2 = 18.40$	.018					
Male				16.3 (50)	10.2 (14)	20.3 (75)	20.7 (19)	22.3 (23)
Female				76.9 (236)	82.5 (113)	73.0 (270)	64.1 (59)	68.9 (71)
Both				6.8 (21)	7.3 (10)	6.8 (25)	15.2 (14)	8.7 (9)

Note: M-CSD = Marlowe-Crowne Social Desirability Scale. Groups that share superscripts are significantly different from each other using Tukey's HSD post hoc tests ($p < .05$)

Offense history

As seen in Table 4, the average number of convictions for prior sexual offences was significantly different across classes, F (4, 1006) = 4.75, p = .001, $\eta 2$ = 0.02, although the size of this difference was relatively small. Individuals in the Multiple Dysfunction class had the highest average number of prior convictions for sexual offences (M = 2.13, SD = 5.95), which was significantly greater than for the Low Needs class (with the lowest number of prior sexual convictions; M = 0.60, SD = 1.53, p <.001), and Intimacy Deficits class (M = 1.05, SD = 2.90, p = .017). However, differences in numbers of previous convictions for other major offences between clusters (defined as violent offences excluding sexual offences) only approached significance, F (4, 985) = 2.19, p = .069, $\eta 2$ = 0.01.

The average number of reported victims significantly differed across clusters, F (4, 1008) = 8.71, p < .001, $\eta 2$ = 0.04. The highest average number of victims was found in the Multiple Dysfunction class (M = 15.60, SD = 36.99), with this number being significantly higher than the average for Low Needs class (M = 3.75, SD = 5.40, p <.001), the Deviant Sexual Scripts class (M = 8.40, SD = 20.29, p = .028), and the Intimacy Deficits class (M = 5.75, SD = 17.75, p <.001).

Personal history

As shown in Table 4, there was a significant association between class membership and reports of childhood sexual abuse, $\chi 2$ (4, n = 1005) = 23.24, p <.001, φ = 0.15. The Emotional Dysregulation class had the highest proportion of individuals who reported being sexually abused as a child (74.7%), with the Low Needs class having the lowest proportion (48.9%). The remaining classes all reported similar proportions of around 60% of individuals reporting childhood sexual abuse. Examination of the standardized residuals indicated that the largest cell discrepancies related to significantly lower proportions of reported abuse than expected for the Low Needs and Deviant Sexual Scripts classes, and a significantly higher than expected proportion for the Emotional Dysregulation class.

A significant association was found between class membership and the onset of sexual offending occurring pre-adulthood, $\chi 2$ (4, n = 1010) = 29.10, p <.001, φ = 0.17. The Multiple Dysfunction class had the largest proportion of individuals who commenced offending pre-adulthood (52.4%), although both the Deviant Sexual Scripts (50.4%) and Emotional Dysregulation (48.4%) classes also had close to half their members falling into this category. The lowest proportion of pre-adulthood offending was found for the Low Needs class (30.2%). Standardized residuals indicated that this class also had a significantly smaller proportion of pre-adulthood offenders than expected.

Class membership was also found to be significantly associated with evidence of a paraphilia, $\chi 2$ (4, n = 1020) = 38.75, p <.001, φ = 0.20. The highest proportion of individuals with a paraphilic interest was found for the Deviant Sexual Scripts class (45.4%), followed closely by the Emotional Dysregulation class (42.4%). The Low Needs class had the lowest rate of paraphilia (22.3%), with this percentage being significantly lower than the expected proportion. Higher than expected proportions of individuals with paraphilias were found for the Deviant Sexual Scripts and Emotional Dysregulation classes.

Victim profiles

Analyses were run to identify any significant associations between class membership and the relationship between victim and offender. Individuals were classified as having related victims only (defined as any relationship in which marriage would typically be outlawed, including step-parents), having extrafamilial victims only, or as having both related and extrafamilial victims. A significant association was found between class membership and relationship with victim, $\chi 2$ (8, n = 923) = 24.14, p = .002, φ = 0.16. The Intimacy

Deficits class had the highest proportion of men with a related victims only (63.6%), followed by the Low Needs class (55.4%). The lowest proportion of individuals with related victims only was found for the Multiple Dysfunction class (42.3%). Relatively large proportions of the Emotional Dysregulation (42.2%) and Multiple Dysfunction (41.2%) classes had extrafamilial victims only, with the Intimacy Deficits class (25.9%) having the lowest proportion. The Deviant Sexual Scripts class had the largest proportion of men with both related and extrafamilial victims (17.6%), followed by the Multiple Dysfunction class (16.5%). The Emotional Dysregulation class (8.4%) had the lowest proportion of men with both related and unrelated victims. Examination of standardized residuals showed that the proportion for the Intimacy Deficits class had a larger proportion of men with related victims only than expected, and a lower proportion of men with extrafamilial victims only.

Lastly, there was a significant association between preferred victim gender and class membership, $\chi 2$ (8, n = 1009) = 18.40, p = .018, φ = 0.14. Although having girl victims only was more prevalent for all classes than having boy victims only or both girl and boy victims, there were a number of key findings from this analysis. In brief, the Emotional Dysregulation class had the highest proportion of individuals with both girl and boy victims (15.2%), with all other classes falling between 6.8% and 8.7%. The Multiple Dysfunction class had the highest proportion of individuals with boy victims only (22.3%), and the Deviant Sexual Scripts class had the lowest proportion (10.2%). The remaining three classes all had similar proportions of men with boy victims only, falling between 16.3% and 20.7%. The Deviant Sexual Scripts class had the highest proportion of men with girl victims only (82.5%), followed by the Low Needs (76.9%), and the Intimacy Deficits (73.0%) classes. Both the Emotional Dysregulation and Multiple Dysfunction classes had around 65% of men with girl victims only. Standardized residuals indicated that the Emotional Dysregulation class had a significantly higher proportion of men with both girl and boy victims than expected and the

Deviant Sexual Scripts class had a significantly lower than expected proportion of men with male victims only.

Discussion

The major aim of the current study was to investigate whether the five subgroups (or ‘pathways’) of sex offenders against children hypothesized by Ward & Siegert's (2002) Pathways Model could be identified in a large sample of 1,134 incarcerated men entering treatment. Pre-treatment scores from a psychometric battery completed by the men were analysed using LPA in order to identify latent classes within the data. Our analysis showed that five classes provided the best overall fit for the data, which were labelled: Low Needs (characterized by scores lower than the overall average on all measures), Deviant Sexual Scripts (characterized by elevated scores on measures of sexual fantasising), Intimacy Deficits (characterized by elevated scores on measures of interpersonal difficulties), Emotional Dysregulation (characterized by scores indicating dysfunction on measures of emotional control and expression), and Multiple Dysfunction (characterized by scores that were above the overall average across all measures apart from state anger). There was no match found for the hypothesized Antisocial Cognitions pathway, and the Low Needs class was not predicted in the original formulation of the Pathways Model.

Comparison with Pathways Model Predictions and Previous Studies

As mentioned above, the main aim of the current study was to test the validity of the distinct etiological pathways proposed in the Pathways Model of sexual offending against children (Ward & Siegert, 2002). One strength of this model is its potential to improve on the utility and explanatory depth of alternative etiological theories by specifying developmental mechanisms which cause the primary vulnerabilities hypothesized to drive offending (Ward et al., 2006). Thus, an additional aim of the current study was to assess whether the hypothesized developmental mechanisms proposed for each pathway were empirically supported.

Because LPA is a "top-down" approach that assigns individuals to classes on the basis of model fit rather than arbitrary distance between individual scores (Hagenaars & McCutcheon, 2009), we can be confident that the clusters extracted in our study identify qualitatively meaningful groups. Our results can therefore provide useful guidance for revisions to the Pathways Model. Below we discuss the fit between our results and the hypotheses of the model, and where required, explore possible amendments to the underlying theory of the Pathways Model.

Deviant Sexual Scripts

Overall, the Deviant Sexual Scripts class showed a profile that was a relatively good fit to the predictions of the Pathways Model, although there were some notable differences. The Pathways Model proposed that individuals in this pathway would have distortions in their sexual scripts related to a preference for impersonal sex; a fear of intimacy was the proposed driver of this preference, with corresponding experiences of loneliness and low social self-esteem. Contrary to this prediction, the Deviant Sexual Scripts class were found to have relatively low levels of dysfunction in social self-esteem, fear of intimacy, and feelings of loneliness. Notably, the group identified in the study by Gannon and colleagues (2012) as being closest to the Deviant Sexual Scripts pathway (labelled 'boy predators') also showed relatively low levels of dysfunction in social intimacy and loneliness. In addition, rates of sexual fantasising for this class in the current study were elevated across all sub-scales of the WSFQ - including exploratory and sado/masochistic – rather than for the impersonal sub-scale alone. Furthermore, individuals in this class reported relatively low rates of sexual victimisation as children; this finding is contrary to the initial theory that the distortions in sexual scripts were a result of experiences of sexual abuse.

These findings show that although individuals in this group might have a primary vulnerability relating to distortions in sexual scripts, it is likely that these distortions relate to more than just the context in which sex takes place (i.e. a preference for impersonal sex). This indication is further supported by the relatively high rates of paraphilia evident for individuals in this class. This would suggest that there may be additional distortions in what Gagnon (1990) referred to as the cultural level of a sexual script, that is, distortions in understanding of cultural norms around sexual behavior, or what is considered culturally permissible. It is also possible that the primary vulnerability for this class relates more to high levels of sexual preoccupation than to distortions in sexual scripts; this is supported by the relatively high proportion of men with both related and unrelated victims, possibly indicating a level of impulsivity relating to offending. Previous research has highlighted the potential role of hypersexuality in sexual offending (Kafka, 2003), and sexual preoccupation has been identified as a dynamic risk factor related to sexual reoffending in previous research (Hanson & Harris, 2000). It would be helpful to identify whether individuals in this class also report excessive masturbation or other compulsive sexual behavior in order to clarify the role of sexual preoccupation in the offending behavior of men in this class.

Intimacy Deficits

The profile of the Intimacy Deficits class was a good fit with the initial predictions made for this pathway. Although these individuals had elevations in anxiety that were not specifically predicted by the Pathways Model, it was suggested that problems with mood regulation could be present; this related to the insecure attachment patterns hypothesized to be the major cause of the intimacy problems for these men. Although the Pathways Model hypothesized that individuals in this group would possess attitudes and beliefs that were supportive of relationships between adults and children, the extracted class did not display corresponding elevations on the ABCS. The Intimacy Deficits group identified by Gannon and colleagues (2012) also showed relatively low levels of support for sexual offending against children. This lack of pro-offending attitudes is also arguably displayed in the relatively low number of prior sexual offenses committed by the men in this class.

This finding does not exclude the possibility that men in this class have cognitive distortions that are specific to a perceived relationship between themselves and their own victim(s); previous research has found that sex offenders report more cognitive distortions on average that relate to their own offending than they endorse on a generic paper-based cognitions measure (Neidigh & Krop, 1992). Indeed, there was a higher proportion of men with related victims only in the Intimacy Deficits class compared to other classes, and the lowest proportion of men with extrafamilial victims only. A closer relationship between offender and victim is congruent with the idea that offending in this pathway occurs after a period of grooming in which the victim is substituted in the place of an adult in a pseudo-relationship. The relatively low scores on measures of sexual fantasising in this class, as well as the relatively low rates of paraphilia, also suggest that the mechanism behind offending for these men relates more strongly to the desire for intimacy than any kind of sexual motivation, thereby supporting the initial hypotheses of the Pathways Model.

Emotional Dysregulation

The Emotional Dysregulation class extracted in the current study displayed the marked dysfunction in emotional competence predicted by the Pathways Model. However, individuals in this class also displayed low-level deficits in almost all other areas measured, something not predicted by the model. Ward and Siegert (2002) proposed that offending in this pathway could either be the result of (a) under-regulation or (b) the use of sex as a way of coping with negative emotions. The current study did find some support for the under-regulation mechanism (specifically limited ability to control anger), however, the average number of prior major offenses was not significantly higher than other classes; high rates of previous offending might be expected from individuals who lose control over their behavior in general. Previous research has, however, identified several deficits related to the use of sex as a coping strategy, including higher sexual preoccupation, and intimacy deficits and loneliness (Cortoni & Marshall, 2001). It is therefore possible that the widespread deficits in this class, including fear of intimacy, loneliness, and high levels of sexual fantasising, could be related to the use of sex to regulate mood. The high proportion of men with signs of a paraphilia potentially also indicates a level of sexual preoccupation in this group. Thus, it is possible that using sex as coping may be the primary mechanism leading to offending in this group, rather than an under-regulation of emotions. Future research is required in order to determine the exact mechanism by which emotional dysregulation could be causally linked to offending.

Multiple Dysfunction

The Multiple Dysfunction class extracted in the current study showed a good fit with the predictions made in the Pathways Model. The individuals in this class showed relatively high rates of dysfunction (approximately 0.5-1.0 SD) in all areas measured aside from State Anger, in which they showed scores that were around the average for the sample as a whole. Furthermore, a Multiple Dysfunction group has been identified in all previous validations of the model. Although the Emotional Dysregulation class also demonstrated relatively high levels of dysfunction in multiple areas, the degree of dysfunction displayed by the individuals in the Multiple Dysfunction class was more consistent across all domains of functioning; for example, scores on measures of intimacy deficits and deviant sexual scripts were much lower than scores on measures of emotional control for the Emotional Dysregulation class, whereas elevations in scores were relatively equal on all measures for the Multiple Dysfunction class.

Although Ward and Siegert initially proposed that the issues related to the Multiple Dysfunction pathway were caused by experiences of childhood sexual victimisation, individuals in the Multiple Dysfunction class that we extracted did not report high levels of this compared to other classes; rates of childhood sexual victimisation were higher for both the Emotional Dysregulation and Intimacy Deficits classes. Additionally, men in the Multiple Dysfunction class did not demonstrate the comparatively higher rates of paraphilia expected by the original Pathways Model. Both of these findings suggest that a distorted sexual script leading to “pure paedophilia” may not be the primary mechanism of dysfunction for individuals in this pathway. Instead, the comparatively greater level of risk and need displayed by these men across a range of indicators is indicative of there being a complex network of mechanisms that contribute to the dysfunction displayed by these men, rather than a single mechanism driving the dysfunction seen in other domains. Further research is required to develop a richer picture of the characteristics of men in this pathway (including developmental factors), in the aim of developing insights into the causal mechanism driving the multiple dysfunction for this group.

Low Needs

As mentioned above, the current study identified a relatively large group of individuals characterized by below average scores across all domains measured. This group was not predicted by Ward and Siegert (2002) in their formulation of the the Pathways Model, and seems contrary to the main hypothesis that sexual abuse of children is driven by some kind of primary vulnerability or dysfunction. Instead, the individuals in this class had raw scores that fell within normative levels of functioning on each of the measures used in this study. Although this group had not previously been predicted, it has been found in previous studies in this area; Gannon et al. (2012) identified a large group of men that appeared to have relatively low levels of dysfunction across all measures (which they labelled 'impulsive' due to a relatively small elevation on a measure of impulsivity). Futhermore, a recent study by Seto and Fernandez (2011) also identified a 'low needs' group when performing a cluster analysis on Stable-2000 scores from 419 men who had been convicted of a sexual offense; these men scored below the overall sample mean on all 16 Stable-2000 items. This suggests that the identification of this low needs class is not the result of random permutations in our data set, but represents a meaningful category of individuals who sexually offend against children.

Whereas it is possible that offending is related to a primary vulnerabilty in a domain that has not yet been identified, the causes of offending for these individuals may relate to idiosyncratic factors that are individual-specific and not able to be identified at an aggregate level. An alternate explanation for the lack of dysfunction found for this group relates to the possibility of the results being biased by socially desirable responding (Gannon et al., 2012). This relates to the tendency for some individuals to respond in ways that they anticipate will be met with approval from others, or to refrain from responding in ways that will be met with disapproval. Consistent with this suggestion, the Low Needs class had an average M-CSD score that was significantly higher than the average for all other classes (as shown in Table 4). This indicates that the level of dysfunction was under-reported by men in the Low Needs class. However it is important to note that although the difference is statistically significant, the mean M-CSD score for the Low Needs class is still within the normative range of functioning found during the scale validation (Marlowe & Crowne, 1964). Thus, it is uncertain whether this slightly higher tendency towards socially desirable responding can fully account for the large differences seen in the levels of dysfunction between this class and the other classes found. Future research is required to ascertain the extent to which socially desirable responding can account for the low levels of need reported, and to identify the primary causes of offending for these individuals; this is particularly important given that optimum treatment targets are not immediately apparent for this relatively large proportion of the offending population. The current study did not include a measure of impulsivity and could therefore not validate the elevations in this found by Gannon et al. (2012), however this is a possible treatment target that should be investigated by future studies.

Antisocial Cognitions

The current study failed to identify a group of men characterized by levels of antisocial cognitions that were higher than average, as predicted by the Pathways Model. A 'generally antisocial' group was identified by Gannon et al. (2012) in their previous validation, which was characterized by high levels of endorsement of cognitions supporting general criminal activity. Notably, this group did not show high levels of endorsement of cognitions supportive of sexual offending in particular. This could be a possible explanation of our failure to find such a group; the current study contained measures of cognitions specific to sexual offending only, rather than including measures of general pro-criminal beliefs. However, given that previous tests of the model have also failed to find an antisocial cognitions group (Connolly, 2004), our results may indicate that antisocial cognitions might not have the primary role that was hypothesized. Indeed, there has been some discussion in the literature that antisocial cognitions may relate more to post-hoc rationalisations of offending rather than playing a causal role in offending pathways (Ward et al., 2006). According to this view, offense-supportive cognitions are generated post-offence to deflect internal or external criticism and therefore maintain self-esteem. Although this coping strategy may have an impact on treatment and the likelihood of future offending, one would therefore not expect to find that antisocial cognitions constitute a primary mechanism of offending. Future research is needed in order to further our understanding of the nature of offense-supportive cognitions, and to identify whether an antisocial cognitions pathway can be identified.

Applications and Limitations

The results of the current study contribute to a growing body of literature that highlights the need for treatment programmes to incorporate a level of flexibility in their delivery, so as to allow for the targeting of treatment to individual needs. This is contrary to the modular delivery of treatment that currently dominates the treatment of sexual offending internationally (Gannon et al., 2012). If offending is indeed driven by distinct mechanisms as posed by the Pathways Model, then it is imperative that we modify treatment programmes to reflect this. For example, individuals who demonstrate high levels of dysfunction in terms of their ability to develop intimate adult relationships (the Intimacy Deficits pathway) are more likely to benefit from treatment that aims to improve their ability to develop meaningful romantic relationships with appropriate partners, rather than treatment modules that address the regulation of anger or hostility. Future research could potentially assess whether treatment gains made in areas directly related to the etiological pathway of a given offender leads to greater reductions in recidivism than general treatment gains. This would provide support for the idea that particular deficits are key causal mechanisms in offending for different individuals.

The current study has also provided valuable guidance for future attempts to modify the Pathways Model to incorporate empirical evidence of distinct mechanisms of offending. Having a clearer understanding of the causes and maintaining factors of sexual offending is essential for identifying promising targets for prevention efforts, which an exclusive focus on risk prediction cannot provide (Ward, 2014). Although Ward and Siegert (2002) provided some suggestions as to the mechanisms behind the primary vulnerabilities in each pathway, our results largely failed to support these predictions. It is important that future research (both theoretical and empirical) focuses on identifying more promising ideas of the developmental and environmental mechanisms that lead to the development of these vulnerabilties; some possible directions for this have been outlined above.

Although the findings of the current study are useful in terms of guiding both research and theory development, there are a number of limitations that must be considered. Perhaps the most important limitation is our study's correlational, retrospective design, and therefore an inability to draw any strong conclusions about causality. Although we were able to identify classes of offenders with profiles that largely matched the causal pathways hypothesised by Pathways Model, we are unable to directly test whether the primary vulnerabilities identified for each class actually lead individuals to sexually offend, or whether they were individually predictive of sexual offending. Additionally, because our psychometrics were applied post-offence, it is not possible to ascertain at what point the offenders gained these characteristics (e.g., the problems in emotional dysregulation could have only become present post-offence or post-arrest). We can therefore only provide corroboration of the hypotheses of the model (as well as highlight areas that require revision), rather than provide a true test of its assertions around causation. It is therefore uncertain whether the primary vulnerabilities for each pathway are best considered as risk factors, or whether they are true causal mechanisms. Because everyone in our sample had committed at least one sexual offence, it is possible that they are causal mechanisms. However further research is needed to determine this with greater certainty.

This limitation regarding causality is not unique to our study, of course; establishing causality is difficult within the area of sexual offending due to ethical and practical considerations, as well as the multi-factorial nature of many key related constructs. However, we believe that it may be possible for future studies to provide a stronger test of the causal theories of the Pathways Model, using our study as a guide. One example would be a study utilising a prospective, longitudinal design that collected psychometric information from a community sample over a number of years. Given a large enough sample, it is likely that it would contain individuals who went on to sexually offend part-way through the study. The psychometric information could then be analysed pre- and post-offence for this group of individuals and compared to the population as a whole in order to determine: a) whether the psychological profiles identified for the classes in our study could also be found pre-offence for the offending sub-sample; and, b) whether these psychological profiles were absent in the non-offending sample. This would provide stronger evidence of causality that would not be subject to the limitations of a retrospective design.

An additional limitation is that the measures used in the study were self-report only, which allows for the possibility of reporting bias to affect the results. One method that has been used in the past to address this concern has been to statistically control for deceptive responding using a measure of socially desirable responding (SDR; e.g., the M-CSD). However, a growing body of research is finding that statistically controlling for SDR actually decreases the accuracy of dynamic risk measures (Mills & Kroner, 2006; Stevens et al., 2016). The results from these studies suggest that SDR should be considered more of a pro-social personality trait rather than a bias towards deceptive responding that needs to be controlled for. In support of this alternate interpretation of SDR, note that a large number of men in our current sample responded in ways that were openly antisocial or abnormal (e.g. men in the Deviant Sexual Interests class). However, it may be useful for future studies should include clinician-rated measures in their analyses to ensure that deceptive responding is not an issue.

A further limitation of the current study is that the measures used were not selected specifically to test the Pathways Model, but were instead chosen from a pre-treatment battery that had been delivered prior to treatment for all men entering the programme. This meant that the measures used were not always perfectly matched with the type of dysfunction hypothesized; for instance, we did not have any measures of the attitudes supportive of general criminality suggested to be present for individuals in the antisocial cognitions pathway. It also meant that there was limited variation in the measures used to assess a particular domain, such as the deviant sexual scripts domain being assessed using only measures of sexual fantasising. It is promising that even given these limitations in the measures our results were still largely consistent with the Pathways Model predictions and the findings of previous empirical studies, however it would be an improvement if future studies were able to deliberately select the measures to be used in the analysis. This would also allow for the measurement of additional domains of functioning that might prove to be important mechanisms of dysfunction for individuals that we identified as being low needs on the basis of our current measures.

Conclusion

The current study's findings corroborated many key hypotheses of the Pathways Model. Our substantially larger sample size allowed for greater confidence in the extracted classes than previous tests of the model, and also allowed for post-hoc analyses to test the developmental mechanisms suggested by Ward and Siegert (2002). Five classes were extracted from our sample, each representing distinct groups of men with differing psychological profiles. Although four of these groups matched pathways originally predicted by Ward and Siegert (the Deviant Sexual Scripts, Intimacy Deficits, Emotional Dysregulation and Multiple Deficits classes), we identified one group of men that was not predicted by the model (a Low Needs class) and we failed to identify a further pathway that was present in the original model (an Antisocial Cognitions pathway). We additionally found a number of discrepancies in our attempt to test the hypothesized mechanisms behind the primary dysfunction found for each class of men. While these results largely provide support for the Pathways Model, they suggest that substantial amendments are required in terms of describing the developmental factors contributing to the development of the four primary vulnerabilities, as well as further explanation of the drivers of offending for men in the Low Needs and Multiple Dysfunction classes. It is hoped that a greater understanding of the etiology of sexual offending may help us in our prevention efforts regarding both re-offending post-release, and first-time sexual offending.

Chapter Four/Study Two: A Difference in Degree or Kind? A Taxometric Analysis of Treatment Change

The previous study provided tentative support for Ward and Siegert's (2002) Pathways Model of sexual offending against children. This model is grounded in the theory that sexual offending is caused by a number of distinct psychological mechanisms, with these distinct mechanisms underlying similar types of harmful sexual behaviour for different groups of individuals. This finding contributes not only to our understanding of the aetiology of child sexual offending, but also indicates the importance of providing individualised treatment for sexual offenders based on their particular needs. It seems to make intuitive sense that if distinct vulnerabilities and mechanisms are driving offending for different individuals, then the needs targeted during treatment should also differ between individuals. This is indeed a concept that is prominent in many influential sexual offending treatment frameworks, including in the Need principle of the Risk-Need-Responsivity model (RNR; Andrews & Bonta, 2010), and the Good Lives Model (GLM; Ward, 2002).

If qualitatively distinct groupings of offenders can be identified on the basis of causal mechanisms or criminogenic needs, and if these distinct needs are targeted in treatment on an individual basis, then it is plausible that the ways in which offenders change over the course of treatment may meaningfully differ as well. For example, some individuals may make large changes in terms of their emotional regulation and smaller changes in terms of their intimacy skills over the course of treatment; others might display large amounts of change only in reducing their sexual compulsion; while others still may make poor amounts of change in all areas.

As discussed in Chapter Two, most current approaches to the measurement of treatment change conceptualise treatment change as a dimensional construct. Because of this, treatment change is generally measured as a continuous variable whereby individuals can make anywhere from small to large amounts of change (or even no or negative change). In other words, differences between individuals in terms of the change they make over the course of treatment is generally considered a difference in the *amount* of change taking place overall, rather than a difference in the *kind* of change being made. However, as mentioned above, the findings of the previous study point towards this latter possibility, that individuals change in different ways over the course of treatment, dependent on the unique mechanisms causing their offending and the criminogenic needs targeted in their treatment programme.

The suggestion that individuals might change in qualitatively different ways rather than only in different amounts would require a shift in our conceptualisation of treatment change, from that of a dimensional construct to a categorical construct. Identifying which of these two options is the more accurate conceptualisation requires furthering our understanding of the latent (or underlying) nature of treatment change itself. Once we know more about the underlying structure of treatment change, we will have a more solid foundation from which to advance our understanding of the mechanisms of change, including the possibility that there are meaningful differences between individuals in terms of the kinds of change they make during treatment.

The Implications of Latent Structure

Uncovering the latent structure of treatment change is a vital step in developing a strong conceptualisation of what it means for an offender to make meaningful change, and therefore how change can best be assessed and applied to offender treatment and management decisions. This is because an understanding of the latent structure of a construct provides a foundation upon which to interpret the different manifestations of the construct. As explained by Walters, Knight and Thornton (2009), if a construct is dimensional in nature then differences between individuals related to that construct are likely to be quantitative in nature. That is, the differences are a matter of the degree to which a certain individual possesses a given characteristic or demonstrates a particular behaviour. Conversely, if a construct is categorical in structure then it suggests that differences between individuals are likely to be qualitative in nature. That is, there are likely to be meaningful, non-arbitrary differences between individuals grouped according to construct parameters. In terms of treatment change, if change is found to be dimensional then it would suggest that individuals differ only in the degree to which they were able to change; from making negative change to making a lot of positive change, and everything in between. However, if treatment change were found to be categorical, then it would suggest that there is something qualitatively different between different groups of offenders and the type of change they make.

The distinction between dimensional and categorical underlying structures also has implications for the aetiology of the construct in question (Meehl, 1973, 1992). Dimensional constructs are typically caused by mechanisms that are multi-factorial and complex in nature, leading to the wide spectrum of construct levels that constitute dimensionality. The aetiology of a categorical construct, however, is typically much simpler in nature, such as a single distinct causal factor or an interaction between a small number of factors. Identifying the latent structure of treatment change will therefore contribute significantly to our understanding of the mechanisms of change and how this might be better promoted in treatment.

A better understanding of the latent structure of change would also have practical applications for the measurement and monitoring of change over treatment. If offender change was found to be dimensional it would suggest that longer, more complex measures of change would be required in order to comprehensively capture the possible spectrum of behaviours displayed at different levels along the change continuum (J. Ruscio et al., 2006).

Conversely, the rejection of dimensionality would indicate that shorter assessment tools would be sufficient, with a focus on maximising discrimination between groups making different kinds of change at the points where these groups might overlap.

Understanding the true latent nature of treatment change is therefore an important scientific question that has major implications for how we understand and measure change, rather than being a purely mathematical or academic inquiry. There are currently no published studies that have attempted to identify whether treatment change is categorical or dimensional in nature. However, one approach to answering this question is through the use of a statistical technique called taxomectric analysis, with Meehl's taxometric method being most commonly used in the literature (Meehl, 1995, 2004; Meehl & Yonce, 1994, 1996; Ruscio, Haslam, & Ruscio, 2006; Waller & Meehl, 1998). The taxometric method has been used in a number of related forensic areas in the past, including investigating the latent structure of: psychopathy (Edens, Marcus, Lilienfeld, & Poythress, 2006; Walters, Duncan, & Mitchell-Perez, 2007); antisocial features (Walters, Diamond, Magaletta, Geyer, & Duncan, 2007); intermittent explosive disorder (Ahmed, Green, McCloskey, & Berman, 2010); and alcohol misuse (Green, Ahmed, Marcus, & Walters, 2011; Slade, Grove, & Teeson, 2009). Perhaps most relevant to the current study, taxometric analysis has also been used to assess the latent structure of sex offender risk (Walters et al., 2009).

Walters and colleagues (2009) used actuarial scores from a variety of sex-offender risk tools in order to investigate the underlying structure of sexual offending risk. Six risk tools were used in the final analysis, including the Static-99, the RM2000, and the SVR-20; overall, a mixture of both static and dynamic measures of risk were used. Total risk scores obtained from a sample of 503 men convicted of sexual offences against adults and/or children were subjected to a series of three taxometric procedures. Each procedure displayed results more consistent with a dimensional latent structure than with a categorical latent structure. These results indicated strongly that sexual offending risk is best conceptualised as a dimensional construct, in that differences in risk between different offenders are likely to be a difference in degree of risk, rather than a difference in the type or nature of this risk. This study also highlighted the importance of recognising that the risk categories currently being used to communicate risk (e.g. low, moderate, and high risk) are likely to be assigned using arbitrary cut-offs between risk levels, rather than being categories that reflect natural or non-arbitrary groupings of offenders based on risk.

The Current Study

The aim of the current study was to investigate the latent structure of treatment change for sexual offenders against children. Treatment change was measured by comparing pre- and post-treatment scores on a psychometric battery completed by men who had completed an in-prison treatment programme for sexual offending against children. This psychometric battery was the same as that described in Study One. These treatment change scores were then submitted to a taxometric analysis in order to provide evidence of the underlying structure of the change demonstrated by the sample over the course of their treatment.

Although there have been no published taxometric analyses of sexual offending treatment change, the results from Walters and colleagues (2009) can be used to infer what the latent structure of treatment change might be. As outlined in Chapter Two, the most common method for measuring treatment change is to compare pre- and post-treatment levels of criminogenic needs using either specific psychometrics (as was used in the current study), or through the use of dynamic risk assessment tools, such as the Violence Risk Scale: Sexual Offense Version (VRS-SO; Wong, Olver, Nicholaichuk, & Gordon, 2003). Risk assessment tools are used in the measurement of treatment change because our current conceptualisation of risk factors and criminogenic needs often overlap to a high degree (Ward & Fortune, 2016). Dynamic risk assessment tools were used in the taxometric analysis in the Walters et al. paper, which found that risk was dimensional in nature. It can therefore be inferred that change in these risk scores (and related criminogenic needs) might also be dimensional. This conclusion is contrary to the possible categorical nature of treatment change suggested by Study One, however there is currently no empirical evidence to support this theoretical assumption. Indeed, as mentioned above, treatment change is widely conceptualised as being dimensional in the way that it is currently measured and applied to decision-making.

Therefore, the hypothesis for the current study was that the taxometric analysis would indicate that treatment change has a dimensional underlying structure, based on existing empirical evidence.

Method

Participants

All men (N = 1,474) who had participated in a high-intensity, prison-based treatment programme for sexual offenders against children in New Zealand between 1990 and 2007 were identified. For 1,128 men, pre- and/or post-treatment scores were incomplete for the psychometric battery, and they were therefore removed from the sample for this study due to the need for complete data to conduct the taxometric analysis. This resulted in a final sample of N = 346, which is a subset of the sample used in Study One. The majority of the sample (64.7%) identified as being of European descent, with 22.3% identifying as New Zealand Māori, 4.6% as Pasifika, and 2.9% as other ethnicities. Most men in the sample had attended the Kia Marama Special Treatment Unit in Rolleston, New Zealand (*n* = 291; 84.1%), with the remaining 55 men (15.9%) having attended the Te Piriti Special Treatment Unit in Auckland, New Zealand.

Both of these treatment units run as therapeutic communities providing intensive cognitive-behavioural treatment programmes, and can house up to 60 men each. Participation in both units is voluntary, with entry requirements including having no more than minimum or medium security classification, and having a conviction for (or admission of) sexual offences against a person under 16 years of age. Men with an IQ of less than 70 were ineligible to attend Kia Marama at the time of data collection due to the cognitive content of the programme, however a modified treatment programme was available for these men at Te Piriti (and is now also available at Kia Marama). For further information on programme content, see Hudson, Wales and Ward (1998).

Because the current study utilises a psychometric battery administered pre- and post-treatment to assess treatment change, programme non-completers were ineligible for inclusion in the final sample. All men had provided written consent for their information to be used for research and evaluation purposes prior to the commencement of assessment and treatment. Ethics approval was obtained from the University of Canterbury Human Ethics Committee prior to commencing the study.

Psychometric battery

The measures used in this study are largely the same as those described in Study One, however both the pre- and post-treatment completed measures were utilised in the measurement of treatment change (further details below). Two additional measures were used in this study: the Beck Depression Inventory II (Beck, Steer, & Brown, 1996) and the Adult Nowicki-Strickland Internal-External Control Scale (Nowicki & Duke, 1983). Descriptions of these measures are provided below, along with a brief reminder of the measures already introduced in the previous chapter. The measures have been grouped according to the overarching psychological construct they relate to.

Anti-social cognitions

The Abel-Becker Cognitions Scale (ABCS; Abel et al., 1989) measures distorted attitudes and beliefs about sexual offending against children.

The Hostility Towards Women scale (HTW; Check, 1985) measures negative beliefs about women, including the acceptance of aggressive motivations and behaviours directed at women.

The Rape Myth Acceptance Scale (RMAS; Burt, 1980) measures attitudes supportive of sexual violence and aggression.

Deviant sexual scripts

Wilson's Sex Fantasy Questionnaire (WSFQ; Wilson, 1978) measures the frequency or strength of different types of sexual fantasies, including intimate themes (e.g. sex with a partner), exploratory themes (e.g. group sex), impersonal themes, (e.g. sex with a stranger), and sado-masochistic themes (e.g. sex involving pain or use of force).

Emotional Dysregulation

The Beck Depression Inventory-II (BDI-II) measures depressive symptoms. The inventory comprises 21 items that are scored from 0 to 3. Standardised cutoff scores are used to indicate minimal depression (0-13), mild depression (14-19), moderate depression (20-28) or severe depression (29-63).

The State-Trait Anxiety Inventory (STAI; Spielberger, 1983) measures general anxiety (T-scale) and current anxiety (S-scale).

The State-Trait Anger Expression Inventory (STAXI; Spielberger, 1988) measures several aspects of anger and anger expression, including state anger, trait anger, anger suppression, anger expression, and anger control. As mentioned in the previous chapter, both the STAXI and the STAXI-2 were used in the assessment battery over the period of time the treatment units have been running; the two versions of the measure are largely similar, however the STAXI-2 splits the anger control scale into two different subscales (Anger Control-Out and Anger Control-In) and has slightly more items overall.

Intimacy Deficits

The Revised UCLA Loneliness Scale (UCLS; Russell et al., 1980) measures experiences of loneliness.

The Fear of Intimacy Scale (FIS; Descutner & Thelen, 1991) measures anxiety about intimate dating relationships.

The Assertion Inventory (AI; Gambrill & Richey, 1975) measures degree of discomfort in situations requiring assertiveness (e.g. turning down a request for a meeting or date), and an individual's likelihood of making an assertive response in these situations.

The Social Self-Esteem Inventory (SSEI; Lawson et al., 1979) measures self-esteem in social situations.

The Adult Nowicki-Strickland Internal-External Control Scale (ANSIE) measures locus of control. This refers to whether an individual perceives events or outcomes as dependent on their own behaviour (internal) or as a result of independent forces, such as other people, luck or fate (external). Forty yes/no items are used to create a total score that ranges along a continuum from internal (0) to external (40) locus of control.

Procedure

Data for the current study was extracted from a database managed by the New Zealand Department of Corrections. This database holds the pre- and post-treatment total scores for each psychometric used in the assessment battery, as well as demographic and offence history information collected prior to programme admission. As mentioned above, there were initially 1,474 cases extracted from the database. The data was cleaned and checked for errors; any entered scores that fell outside of the minimum and maximum scores possible for a given measure were deleted and thereafter considered missing for that individual. Any case that had missing pre- and/or post-treatment scores were then excluded from the study. This resulted in a final sample of 346 men.

Scores for the two different versions of the STAXI were dealt with in the same manner as described in the previous study. For individuals who completed the STAXI-2 (n = 245) rather than the STAXI (n = 101), totals for the two Anger Control subscales were added together to create one overall Anger Control score. This was to ensure that the number of subscales was equivalent between the two versions. Because these subscales had differing numbers of items between the two versions, all scores on the individual subscales were then standardised. To retain consistency across all measures, raw scores on all remaining psychometrics were also standardised before being used in the subsequent analyses described below.

Planned Data Analysis

Treatment change

Treatment change was calculated for each individual using pre- and post-treatment scores on the 19 psychometrics outlined above. Raw change scores were calculated as the difference between pre- and post-treatment scores, with some measures reversed so that positive change scores always indicated change in a pro-social direction. As explained by

Beggs and Grace (2011), raw change scores are problematic because of the confounding effect of pre-treatment scores on overall raw change. Because psychometrics have a minimum or maximum score, the maximum amount of raw change that a given individual can make is restricted by their pre-treatment score. In other words, individuals who have more extreme or deviant scores pre-treatment have greater opportunity to demonstrate larger amounts of raw change (i.e. have a larger raw difference between their initial score and the maximum/minimum score possible).

In order to control for the confounding effect of pre-treatment scores, standardised residual change scores were calculated using the method described in Beggs and Grace (2011). Raw change scores for each of the psychometrics were regressed onto pre-treatment scores. Residuals from the regressions (i.e. obtained change score – predicted change score) were calculated, with these residuals then being standardised for each psychometric.

Taxometric analysis

Once treatment change had been calculated, a taxometric analysis was performed on these standardised residual change scores using the TaxProg programme developed for the R computing environment by J. Ruscio (2014). R Studio version 3.2.2 was used to run the programme.

The term ‘taxometric analysis’ encompasses a series of statistical techniques used to identify whether the latent structure of a given construct is better considered dimensional or taxonic (i.e. categorical). Unlike most statistical methods used in psychological research, taxometric analysis uses a consistency testing approach rather than a significance testing approach (Walters et al., 2009). This means that multiple mathematically-independent techniques are applied to the data, with consistency between the findings of these analyses indicating the latent structure of a given construct. In the current study four taxometric techniques were used: mean above minus below a cut (MAMBAC; Meehl & Yonce, 1994); maximum eigenvalue (MAXEIG; Waller & Meehl, 1998); maximum slope (MAXSLOPE; Grove & Meehl, 1993); and, latent mode (L-Mode; Waller & Meehl, 1998).

MAMBAC. The MAMBAC procedure operates from the premise that for categorical constructs, there will be an optimum "cutting point" on a given indicator whereby individuals can be most accurately separated into each category of the construct. If a construct is dimensional, however, there are no categories to be "correctly" assigned to, and therefore no cutting point. If this cutting score can be found in the data, it is therefore considered to be suggestive of a categorical construct.

To test this, a series of cuts are made at regular intervals along an input indicator (50 cuts were used in the current study). For each of these cuts, the mean scores on an output indicator are calculated for all cases falling above, and all cases falling below, the cut. The difference between means are then plotted along the *y* axis of a graph, with the value of the cut on the input indicator being plotted along the *x* axis. The shape made by these plotted points is then examined. If a cutting point exists in the data then the mean differences should be largest near this cut score, and should decline as the groups are assigned according to cuts further away from this optimum cutting point. Therefore, a peaked MAMBAC curve is indicative of a categorical construct. Dimensional constructs, on the other hand, will not demonstrate this peaked curve and instead often show a concave curve, with mean differences increasing at either extreme of the input indicator. Because the dataset in the current study comprised more than two indicators, this procedure was replicated using two variables (i.e. change scores on a particular psychometric) at a time as the input and output indicators, until all possible pair combinations had been exhausted.

MAXEIG. The MAXEIG taxometric procedure uses associations between indicators as a measure of latent structure, with eigenvalues as the measure of association. In MAXEIG, one variable is selected as the input variable and all other variables are used as output indicators. Cases are sorted according to their score on the input indicator, and are then divided to form a series of subsamples. Within each sample, the first eigenvalue of the output indicator covariance matrix is noted, with high eigenvalues suggesting that variables are strongly related to one another, and low eigenvalues indicating low levels of association between indicators. These eigenvalues are then plotted against the input indicator used to determine the subsamples. The shape of the plotted eigenvalues can then be used to infer latent structure. Constructs that are categorical in nature will show peaked graphs, reflecting the increase in eigenvalues when subsamples contain closer to 50% of taxon and 50% of complement members. Dimensional constructs tend to show relatively flat, irregular or concave graphs because the level of association between variables should stay relatively consistent across subsamples.

In the current study MAXEIG was calculated using 25 windows (i.e. subsamples) with 90% overlap between each successive window. The use of a small number of overlapping windows rather than non-overlapping intervals when using the MAXEIG procedure was supported in a study of the accuracy of different MAXCOV and MAXEIG implementation options (Walters & Ruscio, 2010). Cases were separated into subsamples using the base rate procedure.

MAXSLOPE. The MAXSLOPE procedure identifies latent structure by assessing the shape of the line of best fit through a scatterplot displaying the relationship between two indicators. After plotting each indicator on an *x and y* axis, categorical constructs will generally show two clusters of points located at the upper right and lower left of the graph. Dimensional constructs will instead show a relatively homogenous cloud of points stretching from the lower left to the upper right of the graph. To more accurately gauge this pattern, a local regression curve is generated using scatterplot smoothing techniques that limit the regression to particular points of the graph. This is to allow curved lines of best fit, rather than forcing the solution into a linear form. The graphs of the slopes in the current study were calculated using a locally weighted scatterplot smoother (LOWESS; Cleveland, 1979). For categorical constructs this line of best fit typically displays an S-shaped curve, whereas dimensional constructs produce relatively linear lines of best fit.

L-Mode. L-Mode was the final taxometric procedure used in this study. The L-Mode procedure utilises a starkly different approach to the three procedures introduced above, which is useful in taxometric analysis given the consistency-testing approach to structure identification; the goal is to use non-redundant and relatively independent techniques in order to provide the strongest test of the structure of the research data (Walters et al., 2009). In addition, L-Mode has been found to be increasingly accurate as the number of indicators included in the procedure increases (J. Ruscio et al., 2006).

In L-Mode, indicators are first submitted to a factor analysis. The first (and largest) latent factor is identified, and the scores for each indicator on this latent factor are calculated and graphed. Because these factor scores are composites of indicators that (at least theoretically) are in themselves valid measures of a common latent construct, they should identify categories of the construct more clearly, where these categories exist. Thus, the factor curve resulting from a categorical construct should show a bi-modal shape, whereas a dimensional construct typically displays a unimodal factor curve.

Evaluating consistency. Since the development of taxometric analysis and related procedures, there has been ongoing discussion and debate in the literature regarding the type of data required for robust results using the procedures outlined above. A number of different data requirements have been identified by different groups of researchers, including normality (A. M. Ruscio & Ruscio, 2002), continuity (J. Ruscio, 2000), low levels of within-group correlation (Meehl, 1995), and indicator ability to distinguish between potential categories (also known as 'validity'; Meehl, 1995). That said, a growing number of studies are finding that current taxometric procedures are relatively robust to deviations from these previously identified "requirements" (J. Ruscio et al., 2006). In addition, it can be difficult to accurately assess whether a particular dataset meets these requirements without having prior knowledge of the criteria for identifying latent groups in the data. Given that a number of studies are also finding exceptions to what had previously been considered requirements of the research data (J. Ruscio et al., 2006; Walters et al., 2009), it is also important that the requirements are considered within the context of the unique characteristics of each dataset.

In recognition of this complexity and growing need for a more flexible and applied way to assess the appropriateness of data for taxometric analysis, J. Ruscio, A. M. Ruscio and colleagues developed a novel approach for determining whether a given set of indicators are appropriate for planned taxometric procedures (A. M. Ruscio, Ruscio, & Keane, 2002; J. Ruscio, Ruscio, & Meron, 2007). This approach involves the use of a bootstrapping technique to create comparison sets of data that retain the characteristics of the research data but differ in terms of their latent structure (i.e. one comparison set is forced into a dimensional structure and the other a categorical structure). Each of the comparison sets of data are then subjected to the same taxometric procedures that the research data is to be submitted to, providing an indication of expected results if the research data were categorical or dimensional.

In the current study, 100 sets of dimensional comparison data and 100 sets of categorical comparison data were generated and submitted to each taxometric procedure, with results being combined to produce averaged curves for each procedure. These simulated results demonstrate what results are likely to look like for dimensional data and categorical data sharing the same characteristics as the research data. If the planned taxometric procedures are appropriate for the research data, then the simulated averaged curves should be relatively distinct from one another; if the averaged curves look similar for both types of data, taxometric procedures are unlikely to be able to accurately identify which structure better matches the research data. The curves generated in the current study were visually assessed to determine whether the results were distinguishable from one another, and therefore whether each taxometric procedure was appropriate for the data available.

These comparison curves are also able to be utilised to objectively determine fit between the results obtained from the research data, and results expected from dimensional and categorical data. Prior to the development of this comparison technique, taxometric results could only be visually inspected to identify best fit. This introduced bias and inaccuracy into the process, with the shape of outputs often being subtle and easily mislabelled. Other less subjective techniques for identifying latent structure, such as investigating the consistency of base rate estimates between different procedures or assessing Bayesian probability distributions, have been found to be unreliable indicators of underlying structure (Walters et al 2010).

The Comparison Curve Fit Index (CCFI; J. Ruscio et al., 2006) provides an empirically-based alternative to these more subjective assessments of fit. The CCFI measures residuals from the fit between research data curves and the simulated comparison curves to identify how well the research data fits what would be expected from dimensional and categorical data. The CCFI ranges from 0 to 1, with 0 indicating a perfect fit between research data curves and comparison dimensional curves, and 1 indicating a perfect fit with comparison categorical curves. A CCFI of 0.5 indicates equal fit between the dimensional and categorical comparison curves, indicating indeterminate results. Recent studies have suggested that treating CCFIs that fall between 0.4 and 0.6 as indeterminate leads to the most accurate identification of latent structure (Walters et al., 2010).

The CCFI has been supported as a robust measure of relative fit by a number of Monte Carlo studies (J. Ruscio & Kaczetow, 2009; J. Ruscio & Marcus, 2007; J. Ruscio et al., 2007; Walters & Ruscio, 2009). The use of simulated comparison data combined with the CCFI has also been found to produce accurate results with non-normal data, and can identify categorical latent factors with more than two classes (Walters et al., 2010). In addition, this method of assessing results is compatible with the idea of consistency-testing that lies at the heart of the taxometric procedure, with CCFIs able to be averaged across procedures if desired. For these reasons, and because of the uncertainty regarding the accuracy of alternative methods, comparison data coupled with the CCFI was used in the current study to assess the appropriateness of data and the latent structure indicated by the results.

Results

Raw change

Table 5 below provides the means and standard deviations for raw change scores on each of the measures used. Mean raw change was generally found to be in the prosocial direction across all measures, with the exception of the intimate subscale of the WSFQ. Average change on this subscale indicated a small increase in the amount of sexual fantasising with intimate themes, which could be related to the inclusion of a behavioural arousal reconditioning component of the treatment programme the men attended.

Table 5. Means and standard deviations for raw change scores

Measure	Mean raw change	SD	Mean raw change	SD
ABCS	10.15	13.43		
HTW	3.09	5.97		
RMAS	10.47	15.35		
WSFEX	2.84	8.52		
WSFIN	-0.97	10.56		
WSFIM	3.19	7.92		
WSFSM	1.95	6.32		
BDI-II	6.11	9.52		
STAIS	6.87	13.57		
STAIT	6.08	11.52		
	STAXI (n = 245)		*STAXI-2 (n = 101)*	
STAXS	1.78	6.00	1.10	5.76
STAXT	1.14	5.70	1.74	5.08
STAXE	0.08	3.62	0.42	3.67
STAXP	1.87	4.91	3.29	5.53
STAXC	0.78	5.79	4.24	10.27
SSEI	8.64	25.49		
AIRP	12.61	22.57		
FIS	8.10	22.68		
UCLS	6.54	10.52		
NSIES	3.06	4.79		

MAMBAC

A mean taxon base rate of 0.61 (*SD* = 0.27) was found for the 380 summed input MAMBAC curves. The mean MAMBAC curve produced a CCFI of 0.806, which is consistent with a categorical structure. The mean MAMBAC and simulated comparison curves are displayed in Figure 2; the thick black line represents the mean MAMBAC curve, and the thick grey line represents the averaged curve from the simulated categorical (left) and dimensional (right) data.

The simulated comparison curves are visually distinguishable from each other in terms of their shape, with the dimensional curve showing the characteristic U-shaped curve. The distinctiveness of the two curves confirms the suitability of the data for the MAMBAC procedure. Visual inspection of the curves verifies the CCFI results, with the mean MAMBAC curve more closely resembling the shape of the simulated categorical curve.

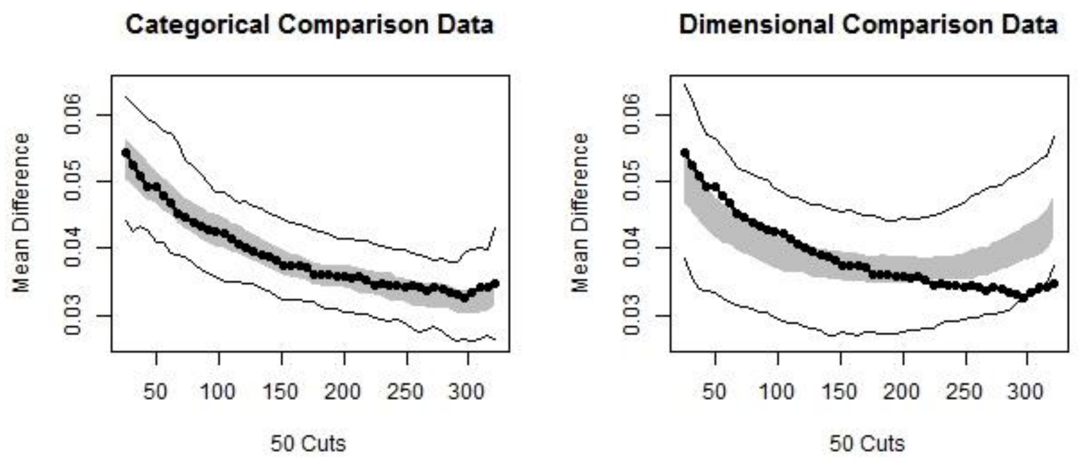

Figure 2. Average MAMBAC curve (darker line) in comparison to simulated taxonic and dimensional data (grey line).

MAXEIG

The average base rate found across the 20 summed MAXEIG curves was 0.70 (*SD* = 0.09). The averaged curve had a CCFI of 0.492, which falls within the 0.4 - 0.6 range indicating an inconclusive result. Figure 3 shows the average MAXEIG curve from the research data, with the average curves from the simulated categorical and dimensional comparison data. The categorical comparison data does demonstrate a slight peaked shape that is expected of categorical constructs, indicating that the parameters of the research data are appropriate for the MAXEIG procedure. However, the shape of the average MAXEIG curve for the research data is neither sufficiently peaked (categorical) nor flat (dimensional) to make a clear determination of latent structure.

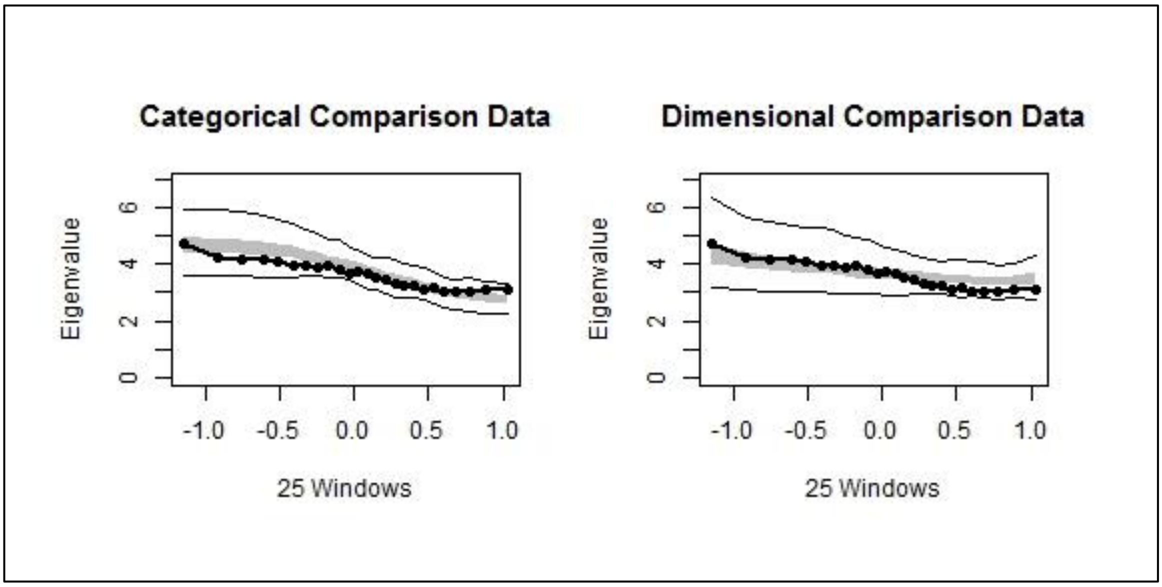

Figure 3. Average maximum eigenvalue (dark line) in comparison to simulated maximum eigenvalues (grey line) for categorical and dimensional data.

MAXSLOPE

The smoothed line of best fit generated by the MAXSLOPE procedure is displayed in Figure 4. The patterns of data points for the dimensional and categorical comparisons (grey line) are noticeably different, with the dimensional comparison data displaying a flatter trendline than the categorical comparison data; this difference confirms the suitability of the research data for the MAXSLOPE procedure.

The research data visually displays a peaked shape that is closer to the categorical comparison data than to the dimensional comparator. This visual similarity is confirmed by a CCFI of 0.639, which is more consistent with a categorical latent structure.

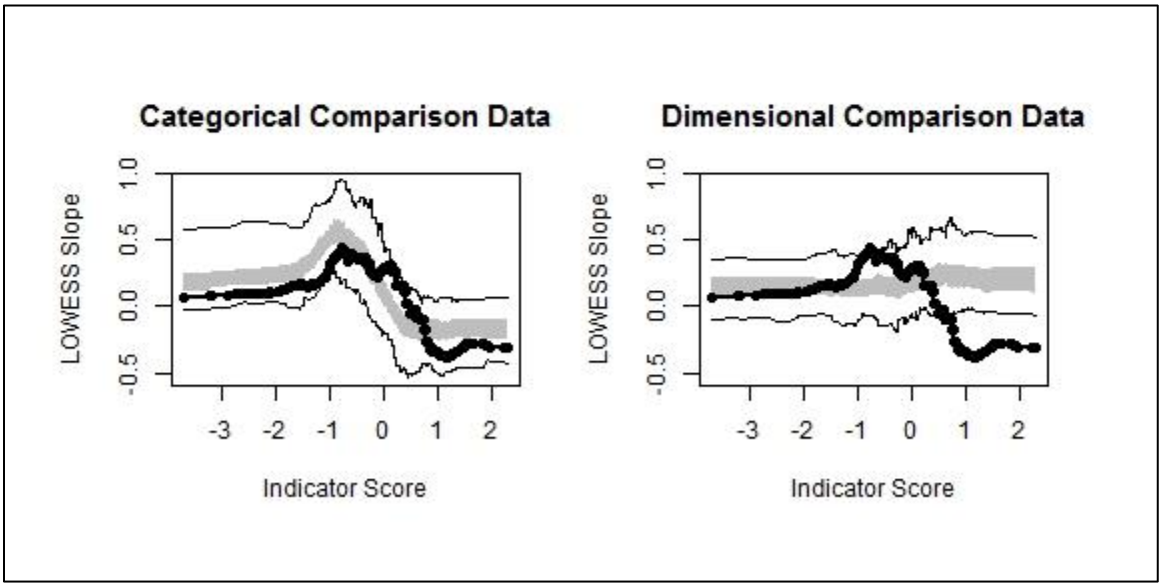

Figure 4. Smoothed MAXSLOPE line of best fit (dark line) in comparison to simulated lines of best fit (grey line) for categorical and dimensional data.

L-Mode

As with the MAXEIG procedure, results from the L-Mode procedure were inconclusive in terms of support for either a categorical or dimensional latent structure. The data curve from the research data produced a CCFI of 0.405, which falls within the 0.4 to 0.6 range indicative of an inconclusive result. Visual inspection of the simulated data curves displayed in Figure 5 shows that anticipated data curves are relatively similar for both categorical and dimensional comparison data. This suggests that the research data may not be particularly suitable for the L-Mode procedure, which is causing problems with successfully interpreting the results.

Consistency testing

Overall, four taxometric procedures were conducted to assess the latent structure of treatment change in our research sample. Two of these procedures (MAMBAC and MAXSLOPE) produced results consistent with categorical comparison data, whereas two of the procedures (MAXEIG and L-Mode) produced results that were inconclusive. Based on the consistency-testing approach of taxometric analysis, these results can be cautiously interpreted as indicating that treatment change is best conceptualised as a categorical rather than a dimensional construct.

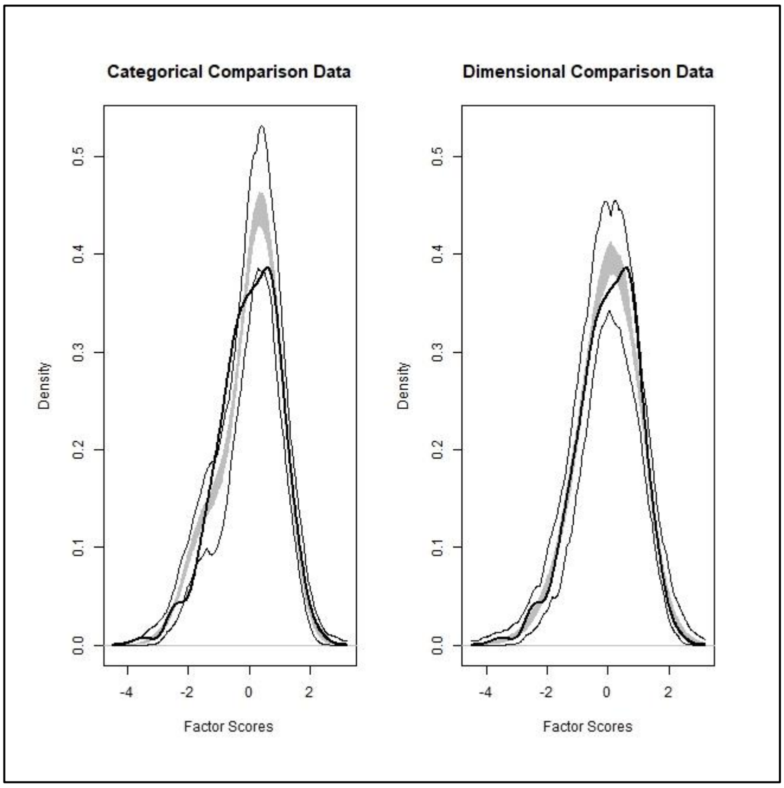

Figure 5. Data curve for the research data (black line) in comparison to simulated categorical and dimensional data (grey lines).

Discussion

Together, the results from this study suggest that treatment change for child sexual offenders is best conceptualised as a categorical construct rather than a dimensional construct. That is, differences between individuals in terms of the change they make over the course of treatment are best understood as differences in the kind of change made rather than simply the amount of change made. This contradicts our original hypothesis, that treatment change is best conceptualised as a dimensional construct. This finding is unexpected given that treatment change amongst child sexual offenders is generally treated as a dimensional construct in current theory and practice. Conceptualising treatment change as categorical rather than dimensional therefore has large implications for our theoretical understanding of the mechanisms involved in change, and the practical applications of this knowledge in terms of monitoring and assessing change over the course of treatment.

As previously discussed, tools that are designed to measure dimensional constructs should differ in important ways from tools designed to measure categorical constructs (J. Ruscio et al., 2006). Tools that measure categorical constructs aim to classify individuals into distinct groups rather than determining a precise level of a construct. This means that rather than providing a numerical total score as an output, categorical measures need to instead provide an indication of class assignment. This goal - to maximally distinguish between different groups of individuals rather than accurately identifying the precise level of a construct – also means that categorical measures often contain a smaller number of items and are scored in ways that differ from a straightforward totalling of individual item scores. It also means that cutting scores or thresholds in measures are non-arbitrary and infer meaningful distinctions between individuals assigned to different groups.

The finding the treatment change is best conceptualised as a categorical construct therefore has important implications for how we should be measuring change. Currently treatment change is often measured by calculating the raw amount of change made according to existing measures of dynamic risk, either in the form of individual risk factors or as a total risk score. However, these measures of risk have been developed to assess dimensional constructs rather than categorical, outputting a numerical score that places individuals along a risk continuum. This makes sense in terms of their primary purpose, given that previous research has identified risk as a dimensional construct (Walters et al., 2009). This is not, however, the way in which we should be assessing a categorical construct such as treatment change. Assessment of treatment change should be aiming to maximally discriminate between different groups of offenders making different kinds – rather than levels – of change. This could provide therapists, judges, parole boards, and other interested parties with a non-arbitrary indication of the type of change demonstrated by a given individual, and the implications this might have for their risk of reoffending.

The results of the current study therefore suggest that we need to develop measures that are specifically designed to measure treatment change, that are distinct from measures developed to primarily assess risk. These measures would only need to consist of a small number of items that allow for the accurate classification of individuals in terms of the amount of change made, with the final assignment of class membership based on a system that may not just be the simple tallying of individual item scores. One example of an existing measure that displays most of these properties is the Violence Risk Scale – Sexual Offense Version (VRS-SO; Wong et al., 2003). Although the VRS-SO is predominately a measure of risk, it also incorporates a separate way in which to measure treatment change. This assessment method is based on Prochaska and Diclemente's Stages of Change Model (Prochaska et al., 1992), which conceptualises change as a process that occurs over five distinct stages: Precontemplation, Contemplation, Preparation, Action, and Maintenance. Change is therefore measured by determining the number of stages that an individual has progressed though over the course of treatment, rather than by subtracting post-treatment scores from pre-treatment scores. The VRS-SO contains a large number of items in order to accurately assess level of risk, and it is unlikely that all of these items would be required to maximally assign individuals to groups based on the type of change made. However, this measure provides a useful starting point for developing a tool specifically designed for the assessment of treatment change. Indeed, it is one of the few tools that has been able to produce measures of treatment change that validly predict sexual reoffending rates (Beggs & Grace, 2011).

Before we can develop tools specifically to measure treatment change – or adapt existing measures that show promise in this domain – we need to further our understanding of exactly how change occurs over the course of treatment, what types of change are displayed by different groups, and the implications this change has for future behaviour. The current study provides an incremental step towards the first of these goals, as the finding that change is best conceptualised as categorical suggests that only a small number of mechanisms are responsible for determining the type of change an individual makes. This means that the causes of change are unlikely to be multifactorial and complex, with change instead being driven by a discrete etiological source, such as a particular cognitive framework or environmental factor, or a specific interaction between these factors. The driver of offender change therefore might be as simple as a "switch" turning on for individuals, whether this is an internal (e.g., a desire to change) or external (e.g., marriage) "switch". Further investigation of the types of change demonstrated by individuals is required before theories of change can be explored further, however the Stages of Change model again provides a promising direction to explore in this area, explaining change as a relatively simple progression in the attitude of individuals towards the possibility and need for change.

Although the current study adhered to current best practice in terms of conducting taxometric analyses, there are a number of limitations that need to be considered when interpreting its results. The main limitation is the relatively small sample size (346 men). Previous taxometric research has indicated that the procedures are most reliable when conducted using sample sizes exceeding 300 (Beauchaine, 2007), a guideline that our current sample meets. However, it would be preferable to conduct this analysis with a larger sample of individuals. This could perhaps lead to more conclusive results with the MAXEIG and L-Mode procedures, thereby increasing the ability to assess procedure consistency and therefore the reliability of the analysis as a whole. As it stands, it is important that further replications of this study are conducted, due to the lack of consistency across all taxometric procedures conducted.

Additionally, the current study uses a measure of change derived from self-report psychometrics. This introduces the possibility of bias in terms of the true extent of change exhibited by individuals; given the importance of treatment progress in the determination of future custody decisions (including release), individuals may have a large incentive to inflate their level of change. It is also preferable to use a range of different measures of a construct when conducting taxometric analysis. This allows for a full assessment of the range of ways in which a given construct can manifest (Broman-Fulks, Hill, & Green, 2008). However, a growing body of research suggests that the artificial inflation of prosocial characteristics by offenders (also known as socially desirable responding) may not be as large a problem as first thought (see Stevens, Tan, & Grace, 2016, for review). Furthermore, the psychometrics included in the current study measure treatment change in a wide range of psychological functioning, despite their all being self-reports. This lessens the potential adverse effects that our method of measuring change has on the final results. That said, future research should incorporate other measures of treatment change, including clinician-rated measures.

The current study is helpful in laying groundwork that will lead to a deeper understanding of the nature and causes of child sexual offender treatment change. We now know that change is likely to have relatively simple etiological roots, with only a small number of factors or interactions driving individual change over the course of treatment. Additionally, our finding that treatment change is categorical has important implications for the measurement of change, and for how change should best be incorporated into future offender management decisions. This includes decisions around further treatment and or relating to offender release and management in the community.

However, before these implications can be fully understood we need to identify what these different classifications of treatment change look like, and the characteristics of individuals who display different kinds of change. The next study addresses this need by conducting a latent class analysis of individuals based on their treatment change. This analysis will help to identify meaningfully distinct groups of individuals who have changed in similar ways over the course of treatment. This is the first step towards identifying how treatment change is best classified, and whether these classes have implications for future reoffending rates (and therefore whether they should be factored in to management decisions).

Chapter Five/Study Three: The Classification of Treatment Change

The results of Study Two indicated that the change that offenders made in treatment across a range of psychological factors was best described as categorical rather than dimensional. That is, some groups of offenders changed in meaningfully different kinds of ways from other groups, rather than merely changing to a different degree. As discussed, this finding has important implications both for the measurement of change and for the way in which change information is used to inform decisions such as the determination and communication of risk, or parole decisions. The obvious next step is therefore to identify the characteristics associated with each of these change categories, and the association between different types of change and sexual recidivism.

A majority of the existing literature assessing change across treatment for sexual offending has measured treatment change as if it were a dimensional construct (see Beggs, 2010, for review). For example, one common method of measuring treatment change is to calculate the quantitative difference between pre- and post-treatment scores across a range of psychometric measures. Differences in the amount of change made are then used to test associations between change and outcomes of interest (such as recidivism; e.g., Allan, Grace, Rutherford, & Hudson, 2007; Beech & Ford, 2006) or factors thought to influence the amount of change made (such as therapist features; e.g., Marshall et al., 2002, 2003). Studies utilising this method to measure treatment change have typically found that individuals make pro-social change over the course of treatment on average, however there is a lack of reliable and consistent findings linking this pro-social treatment change to reduced recidivism (Beggs, 2010).

A similar method of measuring treatment change that is used in the literature is through assessing changes in dynamic risk based on risk assessment tools that incorporate some measure of treatment progress (Beggs, 2010). Some guided clinical judgement tools (such as the Structured Anchored Clinical Judgment [Thornton, 1997] and the Multifactorial Assessment of Sex Offender Risk for Recidivism [Barbaree, Seto, Langton, & Peacock, 2001]) incorporate items relating to treatment progress into their post-treatment assessments of risk; however, these tools are designed more for an overall assessment of post-treatment risk rather than a discrete measure of overall treatment progress. Treatment change is sometimes measured using these risk assessment tools by calculating change as the difference between pre- and post-treatment risk scores. A recent meta-analysis used this approach to assess the predictive validity of change scores derived from dynamic risk assessment tools across nine studies and six unique samples (van den Berg et al., 2018). Overall, the meta-analysis found that change scores based on dynamic risk assessments were able to successfully predict sexual recidivism (d = .26, 95% CI [.10 - .42]), and added incremental predictive validity beyond that provided by static risk and pre-treatment dynamic risk alone.

Although most dynamic risk assessment tools do not explicitly incorporate measurement of treatment change, there is one exception: the Violence Risk Scale - Sexual Offense version (VRS-SO; Olver, Wong, Nicholaichuk, & Gordon, 2007), which includes a structured method for scoring overall treatment change. The VRS-SO calculates post-treatment risk across a number of dynamic risk factors by adjusting pre-treatment scores based on progression through a series of "stages of change". These stages of change are drawn from the Transtheoretical Model (Prochaska et al., 1992), and represent the internal change process occurring in individuals during treatment, providing a way to categorise individuals based on their intentions or demonstrated efforts to make change. The VRS-SO reduces each dynamic risk factor score by 0.5 points for each stage progressed through past the point of "Contemplation"; therefore, the method by which the VRS-SO calculates change is based on change categories, but is translated into a dimensional overall change score.

Notably, there have been a number of studies that have shown total VRS-SO change scores to be a reliable and valid predictor of recidivism after controlling for static and pre-treatment dynamic risk (Beggs & Grace, 2011; Olver, Sowden, et al., 2018; Olver et al., 2007).

A third way in which treatment change can be measured is through the use of tools specifically developed to measure change (Beggs, 2010). Examples of these tools include the Standard Goal Attainment Scaling (SGAS) for sexual offenders (Hogue, 1994) and the Sex Offender Treatment Intervention and Progress Scale (SOTIPS; McGrath, Lasher, & Cumming, 2012). Change is typically measured on these tools using structured scoring frameworks across items that encompass both common dynamic risk factors and treatment-specific motivation or engagement items. Additionally, these measures can be scored multiple times across treatment to track progress and inform treatment targets. For example, the SOTIPS has previously been used to successfully guide collaborative treatment planning with offenders (Lasher et al., 2015). Importantly, change measured using both the SGAS and the SOTIPS has been found to significantly predict reoffending (Beggs & Grace, 2011; R. J. McGrath et al., 2012).

Although there are multiple ways in which treatment change is measured in the literature, there is one shared feature of these approaches that is pertinent in light of the findings from Study Two: treatment change is measured as a dimensional construct rather than as categorical. In each of these approaches change is represented on a continuous scale, calculated through some method of collating differences between pre-and post-treatment scores across a series of dynamic risk factors or engagement/motivation-related factors. This runs counter to the findings of Study Two, which suggested that treatment change is better conceptualised as a categorical construct rather than as dimensional. The fact that previous studies have been able to identify significant associations between treatment change (measured on a continuum) and recidivism may suggest that this distinction is not important in practical terms. However, the measurement of change using methods outlined above has recently come under question due to the potentially unreliable nature of change measured in this way, leading to the utilisation of "clinically significant change" as a way of measuring treatment change (Nunes et al., 2011). Clinically significant change involves categorising change based on two factors: whether post-treatment scores are within levels found in the "functional" population, and whether the amount of change exceeds the margin of error. When the clinically significant change requirements are applied to change measured in the ways outlined above, studies typically find no significant association between change and recidivism (Barnett et al., 2013; Wakeling, Beech, & Freemantle, 2013). A study by Olver and colleagues (Olver et al., 2015) provides a notable exception to this, finding that clinically significant change as measured by the VRS-SO was significantly associated with recidivism; however, their results also suggested that the clinically significant change categories may be partially conflating treatment change with pre-treatment risk.

The Current Study

The literature is currently unclear as to the most appropriate method of measuring treatment change, and regarding the association between measured treatment change and recidivism. As already mentioned, the results of Study Two suggest that we need to adapt our method of measuring treatment change to better reflect its categorical nature. This would involve first developing an understanding of what these categories of change are, and next on this basis, how we would best go about grouping individuals according to these categories. Identifying the most appropriate way of measuring and conceptualising treatment change may lead to more consistent findings regarding the association between change and recidivism, and would have important implications for how change should be communicated and incorporated into risk assessment.

The purpose of the current study was therefore to identify whether the change made by individuals over the course of treatment can be grouped into distinct and meaningful categories, and then to explore the patterns of change within these categories. To achieve this, latent profile analysis was used to explore whether change made across a pre- and post-treatment psychometric battery can be categorised into meaningfully distinct groupings of individuals. Patterns of change within these groups are then explored to identify any overt characteristics of change that differ between groups, with the aim of increasing our understanding of the unique categories of treatment change suggested by the results of Study 2.

Method

Participants

All men (N = 1,474) who had participated in a high-intensity, prison-based treatment programme for sexual offending against children in New Zealand between 1990 and 2007 were identified; this was the same sample originally identified for Study Two. All men had provided written consent for their information to be used for research and evaluation purposes, prior to their commencement of assessment and treatment.

Participants were excluded from the current study if, due to missing data, they did not have raw pre-post change scores for at least one of the four factors assessed by the psychometric battery (more on the factor structure of the psychometric battery below). This was the case for 304 men, leaving a final sample of 1,170 men for inclusion in this study. This sample includes all men from the sample for Study Two, and is a subset of the sample used in Study One.

The majority of the final sample (69.5%) identified as being of European ethnicity. Just under a quarter (23.4%) identified as NZ Māori, with 5.2% identifying as Pasifika and

1.9% as other ethnicities. Most men in the sample had attended the treatment programme at Kia Marama Special Treatment Unit in Rolleston, New Zealand (*n* =760, or 65.0%), with the remaining 410 men (35.0%) having attended Te Piriti Special Treatment Unit in Auckland, New Zealand.

Psychometric battery

The measures used in this study are the same measures that were used in Study Two, with scores from pre- and post-treatment self-report psychometrics being used to assess treatment change. Below a brief reminder of each of these measures is provided, grouped according to the overarching psychological construct that they relate to.

Anti-social cognitions

The Abel-Becker Cognitions Scale (ABCS; Abel et al., 1989) measures distorted attitudes and beliefs about sexual offending against children.

The Hostility Toward Women scale (HTW; Check, 1985) measures negative beliefs about women, including the acceptance of aggressive motivations and behaviours directed at women.

The Rape Myth Acceptance Scale (RMAS; Burt, 1980) measures attitudes supportive of sexual violence and aggression.

Deviant sexual scripts

Wilson's Sex Fantasy Questionnaire (WSFQ; Wilson, 1978) measures the frequency or strength of different types of sexual fantasies, including intimate themes (e.g. sex with a partner), exploratory themes (e.g. group sex), impersonal themes, (e.g. sex with a stranger), and sado-masochistic themes (e.g. sex involving pain or use of force).

Emotional Dysregulation

The Beck Depression Inventory-II (BDI-II) measures depressive symptoms.

The State-Trait Anxiety Inventory (STAI; Spielberger, 1983) measures general anxiety (T-scale) and current anxiety (S-scale).

The State-Trait Anger Expression Inventory (STAXI; Spielberger, 1988) measures several aspects of anger and anger expression, including state anger, trait anger, anger suppression, anger expression, and anger control. As with Studies One and Two, both the STAXI and the STAXI-2 were used in the assessment battery at different times over the period the study sample was collected from. The two versions of the measure are largely similar, however the STAXI-2 splits the anger control scale into two different subscales (Anger Control-Out and Anger Control-In) and has slightly more items overall (57 items compared with 44 items in the original measure).

Intimacy Deficits

The Revised UCLA Loneliness Scale (UCLS; Russell et al., 1980) measures experiences of loneliness.

The Fear of Intimacy Scale (FIS; Descutner & Thelen, 1991) measures anxiety about intimate dating relationships.

The Assertion Inventory (AI; Gambrill & Richey, 1975) measures degree of discomfort in situations requiring assertiveness (e.g. turning down a request for a meeting or date), and an individual’s likelihood of making an assertive response in these situations.

The Social Self-Esteem Inventory (SSEI; Lawson et al., 1979) measures self-esteem in social situations.

The Adult Nowicki-Strickland Internal-External Control Scale (ANSIE) measures locus of control. This refers to whether an individual perceives events or outcomes as dependent on their own behaviour (internal) or as a result of independent forces, such as other people, luck or fate (external).

Sexual recidivism

Criminal history information was obtained from the National Intelligence Application (NIA) database maintained by the NZ Police. Details of all criminal convictions were obtained, including the type of offence, hearing and offence dates, and prison release dates. Convictions coded by NZ Police as relating to sexual offending were counted as sexual convictions, including contact offences such as sexual assault and non-contact offences such as exhibitionism and pornography-related offending. Time at large was also obtained from this database, with the follow-up period starting when the individual was released from prison and continuing until the offence histories were obtained (31 December 2008). Reconviction information was able to be obtained for 1,037 of the 1,170 men included in this study.

Static risk

To measure the static risk of the men in our sample, Static-99 (Hanson & Thornton, 1999) scores were obtained from offender records; these had been previously rated according to file information for a subset of men in our sample (*n* = 218). The Static-99 is a well-validated actuarial scale for predicting sexual recidivism based on demographic and offence-history information. The measure comprises 10 items scored either 0 or 1, with the exception of one item (prior sex offences) that is rated on a 0-3 scale. Item scores are summed, giving a maximum total score of 12; higher scores are indicative of higher levels of recidivism risk.

Procedure

Data for the current study was extracted from the New Zealand Department of Corrections database used to obtain information for the sample in Study Two. This database holds pre- and post-treatment total scores for each measure used in the psychometric battery, as well as demographic information and prior offence histories collected at programme admission. Initially, 1,474 cases were extracted from the database. The data was then cleaned and checked for errors; any psychometric scores that fell outside of the minimum and maximum scores possible for a given measure were deleted and thereafter considered missing for that individual.

Scores for the two different versions of the STAXI were dealt with in the same manner as described in the previous study. For individuals who completed the STAXI-2 (n = 245) rather than the STAXI (n = 101), totals for the two Anger Control subscales were added together to create one overall Anger Control score. This was to ensure that the number of subscales was equivalent between the two versions. Because these subscales had differing numbers of items between the two versions, all scores on the individual subscales were then standardised. To retain consistency across all measures, raw scores on all remaining psychometrics were also standardised before being used in the subsequent analyses described below.

Planned Data Analysis

To reduce the complexity of the results, a factor analysis was conducted using standardised pre-treatment scores on the measures included in the psychometric battery. The goal of this analysis was to identify whether the large number of measures could be adequately and reliably captured by a smaller number of dimensions that could then be used in further analyses.

Treatment change was calculated as described in Study Two, with residual change scores being calculated from the raw pre- and post-treatment change to control for the confounding effect of pre-treatment status. Raw change scores for each of the psychometrics were regressed onto pre-treatment scores, and residuals from the regressions were calculated and then standardised for each measure. Change was calculated in such a way that positive levels of change on each measure represented change in a pro-social direction. Using the results of the factor analysis, residual change scores for each individual were then averaged across the resultant factors, generating an average residual change score for each of the factors identified.

A latent profile analysis (LPA) was then conducted to identify meaningfully distinct groupings of individuals based on the change they made during treatment; a full description of the LPA procedure can be found in Study 1. The average residual change scores across each factor were used in the analysis, which was conducted with MPlus 7.4 software. Full information maximum likelihood estimation was used to impute data for cases with missing factor change scores.

Men were assigned to groups according to their highest level of membership probability suggested by the LPA model. A series of ANOVAs were then run to identify differences between the groups in terms of their average factor change scores (and therefore the type of change made during treatment). To further investigate the implications of group membership, a Kaplan-Meier survival analysis was conducted to compare the survival rates of the groups identified. These additional analyses were conducted using SPSS 23 software.

Results

Factor analysis

A series of exploratory factor analyses were conducted to determine whether the change scores for individual psychometrics could be adequately represented by one or more dimensions that combined change across similar kinds of measures. A principle components analysis (PCA) was conducted using standardised pre-treatment scores for each of the psychometrics. Cases that did not have pre-treatment scores for all measures were excluded from this analysis, resulting in a sample of 558 men.

Inspection of the scree plot and eigenvalues produced by the PCA indicated that a four-factor solution was optimal, with the four factors together accounting for 61.5% of the variance in the pre-treatment scores. To improve the interpretability of results, factor loadings were rotated by the varimax normalised method and items with loadings of 0.35 and above were retained in each factor. The results of this analysis are shown in Table 6.

Table 6. Factor loadings for psychometric battery obtained from principal components analysis with varimax normalised rotation

Measure	**F1:** *Social inadequacy*	**F2:** *Sexual interests*	**F3:** *Anger /hostility*	**F4:** *Pro-offending attitudes*
UCLA loneliness	0.83			
STAI - state	0.64			
STAI - trait	0.75			
Social self-esteem	-0.73			
Fear of intimacy	0.67			
AI – response probability	0.61			
Beck depression inventory	0.58			
Internal-external control	0.48			-0.47
WSFQ - exploratory		0.92		
WSFQ - impersonal		0.91		
WSFQ - intimate		0.87		
WSFQ – sado/masochistic		0.75		
STAXI - state			-0.60	
STAXI - trait			-0.81	
STAXI - suppression	0.58			
STAXI - expression			-0.88	
STAXI - control			0.51	
Rape myth acceptance				-0.90
Abel-Becker cognitions				0.80
Hostility toward women	0.36			-0.45

Individual psychometrics loaded onto the four factors in an almost identical structure as that identified in Allan et al.’s (2007) study, however in the current sample the Hostility Toward Women scale loaded onto both the Social Inadequacy and the Pro-offending Attitudes dimensions. Because of the similarity in results, the factor nomenclature developed by Allan et al. was also used in the current study. The first of these factors, termed *Social Inadequacy*, contained measures relating to poor social skills (loneliness, low social self-esteem, fear of intimacy, low assertiveness, external locus of control, hostility toward women), anxiety and depression. The second factor, *Sexual Interests*, contained all four sub-scales of the Wilson Sexual Fantasy Questionnaire. The third factor contained measures associated with *Anger/Hostility*, including state and trait anger, anger expression, and low control of anger. The final factor, *Pro-offending Attitudes*, included measures of external locus of control, and distorted attitudes and beliefs about sex (rape myth acceptance, distorted cognitions) and women (hostility toward women).

Latent Profile Analysis (LPA)

Following the factor analysis, an LPA was conducted to identify meaningfully distinct groups of men based on the change they made over the course of treatment. First, treatment change across each of the four psychometric factors were derived for each individual by averaging their standardised residual change scores across the measures captured by each factor.

These average factor change scores were then used to test a series of two- to six-class models using LPA. As previously explained in Study 1, there is currently no consensus on the best approach for quantifying the optimum number of classes using LPA and most researchers therefore use a variety of different fit indicators in their analysis (Nylund et al., 2007; Tein et al., 2013). These quantitative fit criteria are then considered in conjunction with several other factors – including parsimony, interpretability of results, and theoretical expectations – to determine the optimum number of classes; the optimum number is where all groups are distinct and the addition of an extra class does not provide additional explanatory power (Fox & Farrington, 2016). For the current study, the Akaike information criterion

(AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and sample-size adjusted BIC (SBIC; Sclove, 1987) were used as initial indicators of model fit, with lower values indicating a better fit.

As shown by Table 7, AIC, BIC and SBIC values continued to drop as the number of classes included in the LPA model increased. Because results from these fit criteria were inconclusive, additional indicators were utilised to determine the optimum class solution, including entropy and the Lo-Mendell-Rubin test (LMRT; Lo, Mendell, & Rubin, 2001). As explained in Study One, the LMRT is a likelihood ratio test that compares the fit of a model with k class to a model with k-1 classes. The test returns a statistically significant result if the model with the greater number of classes provides a significantly better fit with the observed data.

Table 7. Fit Indices and Entropy for all Class Solutions

Class Solution	Logliklihood	AIC	BIC	SBIC	Entropy	LMRT
2 classes	-4236.47	8498.94	8564.78	8523.49	0.70	<.001
3 classes	-4167.82	8371.64	8462.81	8405.64	0.72	.018
4 classes	-4116.74	8279.48	8395.97	8322.92	0.66	.401
5 classes	-4080.77	8217.55	8359.36	8270.42	0.71	.128
6 classes	-4056.82	8179.63	8346.77	8241.95	0.74	.646

Note: AIC = Akaike information criterion; BIC = Bayesian information criterion; SBIC = sample-size-adjusted BIC; LMRT = Lo-Mendell-Rubin test.

Table 7 shows that the LMRT was significant at the p<.05 level for the 2- and 3-class solutions, but not significant for the subsequent models tested. This indicates that the 3-class solution was a better fit to the observed data than the 2-class solution, but that subsequent models did not provide a significant improvement to model fit. Entropy also reduced for the 4- and 5-class solutions compared with the 3-class solution, indicating that the 3-class solution resulted in groups that were more distinct from one another than the 4- or 5-class solutions; however, entropy was slightly higher for the 6-class solution than for the 3-class solution.

Together, the results from the quantitative fit indicators suggested that the 3-class solution was the optimal fit for our data. This was confirmed by assessing the average factor change scores for each group obtained from the different models tested. Results were most parsimonious and easily interpreted for the 3-class solution, with the patterns of change demonstrated by the three groups aligning with existing literature and theories of the mechanisms of offender change (including the relatively straightforward change mechanism indicated by results from Study 2). For these reasons, it was decided that the 3-class solution was the best fit for our data.

To assess the potential impact of imputing missing data on the results of our analysis, the LPA was repeated using two further sub-samples. The first was a sub-sample of 906 men with change scores available for at least one of the measures included in each of the four factors (meaning that there were no missing factor scores), and the second was a sub-sample of 346 men who had complete change scores across all of the psychometrics used in the current study. Results from both of these LPAs were not substantively different from that obtained using imputed data, with both concluding that the 3-class model was the optimum fit to the data. For this reason, the following analyses were completed using the original larger sample.

Treatment change by group membership

Individuals were assigned to groups according to their highest group membership probability indicated by the LPA. The average standardised factor change scores found for each of these three groups is displayed in Figure 6 below.

A series of one-way ANOVAs were conducted to assess the difference in average standardised change between the three groups for each of the four factors. All ANOVAs were significant (*F* values ranged from 72.63 to 773.63, *p*<.001); post-hoc Tukey HSD tests showed that each group significantly differed from the other two groups across all factor scores (*p*<.001 for all comparisons).

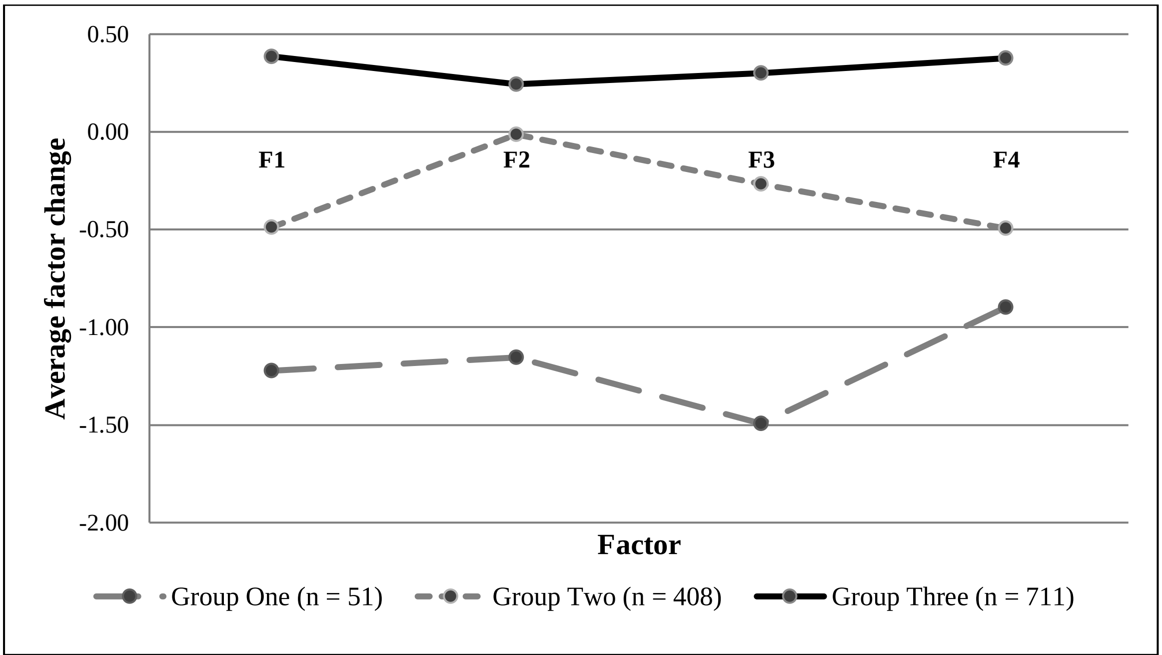

Figure 6. Average standardised factor change, by LPA group

Note: F1 = Social Inadequacy; F2 = Sexual Interests; F3 = Anger/Hostility; F4 = Pro-Offending Attitudes

The average standardised factor change scores displayed in Figure 6 were able to provide a general indication of the differences between groups in terms of the change made during treatment. However, because the scores are standardised, it is not possible to directly identify from these figures whether each group made positive or negative change overall (i.e., whether the groups showed pro-social change on average, or whether raw change was minimal or even negative across measures). To provide an understanding of the type of change made by each of the groups, the average raw change made on each of the psychometrics is presented in Table 8 below. One-way ANOVAs were conducted to assess the difference in raw change across groups; a significant difference was found for all measures apart from antisocial cognitions, with effect sizes (eta squared) generally falling in the medium to high range.

The first group identified in the LPA was the smallest group extracted, representing 51 (4.4%) men from the sample. This group displayed the lowest amounts of average change across all factors compared with the other two groups, with the average levels of change across factors ranging from 0.9 to 1.5 standard deviations below the average change for the entire sample. As demonstrated in Table 8, this group made negative change (i.e., change in the antisocial direction) on average across all psychological factors measured, aside from antisocial cognitions, rape myth acceptance, and having an external locus of control. For this reason, this group was named the *Poor Change* group, representing a group of individuals who on average did not appear to respond in the desired way to the treatment programme.

The second group identified in the LPA captured 408 (34.9%) men from the sample. As demonstrated in Figure 6, this group also displayed average factor change scores that were lower than the average for the entire group, with scores ranging from 0.5 standard deviations below the mean to sitting just below the mean (-0.01 for Sexual Interests). These negative standardised factor change scores are largely the result of the considerably higher amounts of change made by the larger Group Three, however; as Table 8 shows, individuals in Group Two demonstrated change in a pro-social direction across most of the psychological factors measured.

Table 8. Mean unstandardised change scores for all measures, by LPA class

		Group One: Poor change		Group Two: Moderate change		Group Three: Good change			
Measure	*n*	Mean	SD	Mean	SD	Mean	SD	*F*	*η2*
F1: Social inadequacy									
Social self esteem	995	-8.53 [a]	24.20	0.16 [a]	22.62	13.12	27.93	37.23**	0.07
Assertion - response prob.	847	-2.89 [a]	24.77	1.20 [a]	20.80	18.27	23.22	64.20**	0.13
Fear of intimacy	870	-7.05 [a]	23.19	-0.08 [a]	20.72	13.55	20.46	52.69**	0.11
UCLA loneliness	904	-4.21	9.52	1.72	8.80	9.65	10.47	88.20**	0.16
Depressive symptoms	668	-1.70	12.72	3.44	9.69	8.01	9.20	26.29**	0.07
External locus of control	929	0.45 [a]	5.70	0.74 [a]	4.16	4.30	4.83	67.61**	0.13
State anxiety	991	-5.71	17.78	3.01	12.89	10.35	13.08	54.74**	0.10
Trait anxiety	988	-5.28	9.74	1.82	8.71	9.47	11.26	87.10**	0.15
Anger supression	908	-2.79	4.39	0.38	4.22	3.66	4.92	76.92**	0.15
F2: Sexual interests									
WSFQ - exploratory	1122	-2.60	8.70	3.08	8.92	4.67	8.14	18.58**	0.03
WSFQ - intimate	1002	-3.70 [a]	10.02	0.30 [b]	10.04	-0.42 [ab]	10.10	3.38*	0.01
WSFQ - impersonal	998	-3.13	7.63	2.53	7.60	3.86	7.36	20.37**	0.04
WSFQ - sado/masochistic	965	-5.00	8.92	1.76 [a]	6.79	1.76 [a]	5.40	26.86**	0.05
F3: Anger/hostility									
State anger	933	-4.87	7.68	1.50 [a]	6.51	1.98 [a]	5.26	28.52**	0.06
Trait anger	935	-4.46	6.39	-0.07	5.09	2.70	4.93	62.98**	0.12
Anger expression	932	-2.33	4.21	-0.26	3.94	0.78	3.60	19.19**	0.04
Anger control	915	-1.44 [a]	5.05	-1.21 [a]	5.84	2.78	6.87	43.79**	0.09
F4: Pro-offending attitudes									
Abel-Becker cognition	1019	12.00 [a]	13.40	12.34 [a]	17.08	12.18 [a]	13.21	0.02	0.00
Hostility toward women	963	-1.54	7.37	0.70	5.19	4.57	5.35	72.02**	0.13
Rape myth acceptance	1005	5.11 [a]	18.88	6.85 [a]	15.23	12.89	13.58	22.95**	0.04

Note. **p*<.05, ***p*<.001. Groups that share superscripts are not significantly different from one another using Tukey's HSD post-hoc tests (*p*<.05)

Notably, scores on a number of measures of anger (trait anger, anger expression and anger control) showed average changes in an antisocial direction, as did fear of intimacy scores. Additionally, although many of the measures showed change in a pro-social direction for this group, post-hoc Tukey's HSD tests found that the average amount of change was not significantly different to levels of change made by the Poor Change group across many measures. For these reasons, this group was named the *Moderate Change* group, representing a group of individuals who appeared to have responded to some degree to treatment (or perhaps who had begun to respond).

The third group extracted by the LPA was the largest group, with 711 (60.8%) men from the sample being assigned to this group. These individuals showed higher average change than the other two groups across all factors, ranging from 0.2 to 0.4 standard deviations above the sample average. As shown in Table 3, the average raw change across the measures used in the study was also in a pro-social direction for all psychological factors except the frequency of sexual fantasising with intimate themes. Furthermore, the average raw change demonstrated was significantly higher than the average change made by the other two groups across all factors measured apart from four (sexual fantasising with intimate and sado/masochistic themes, state anger, and pro-offending attitudes). For these reasons, this group was named the *Good Change* group, representing a group of individuals who appeared to make good amounts of change across all factors over the course of treatment.

Group recidivism

To assess whether group membership had an association with outcomes post-treatment, sexual recidivism information was obtained for 1,037 men in the sample. Average follow-up time was 15 years 3 months, and ranged from 15 days to 25 years 4 months. A one-way ANOVA found that average follow-up times were not significantly different between the three groups (F = 0.30, p = .742). As shown in

Table **9**, 154 men (14.9%) received a new conviction for a sexual offence during the follow-up period. The proportion of men receiving a new conviction for a sexual offence was highest for the Poor Change group (25.0%) followed by the Moderate Change group (17.7%), with the Good Change group showing the lowest rates of sexual recidivism (12.4%).

Table 9. Sexual recidivism outcomes by change group

Group	*n*	Average follow-up (days)	Number of recidivists	Proportion recidivist
Poor Change	48	5416	12	25.0%
Moderate Change	362	5640	64	17.7%
Good Change	627	5555	78	12.4%
Total	**1037**	**5578**	**154**	**14.9%**

To account for differences in follow-up times, a Kaplan-Meier survival analysis was conducted to confirm these differences in recidivism rates between groups after controlling for time at risk. As shown in Figure 7, there were significant between-group differences in the rate of sexual recidivism (using Generalised Wilcoxon; χ^2 = 7.56, p = .023). Mantel-Cox pairwise comparisons revealed that individuals in the Good Change group reoffended at a significantly slower rate than those in the Moderate Change (χ^2 = 4.53, p = .033) and the Poor Change (χ^2 = 6.09, p = .014) groups; there was no significant difference in recidivism rates between the Poor and Moderate change groups (χ^2 = 1.36, p = .244), potentially because of the small sample size for the Poor Change group (n = 48).

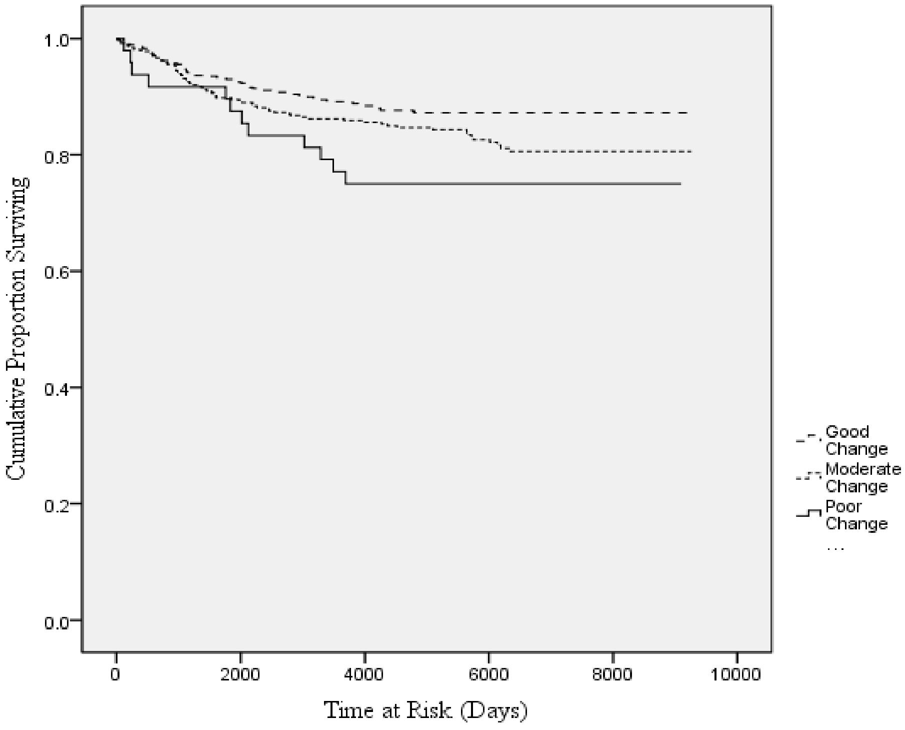

Figure 7. Kaplan-Meier survival plot for sexual recidivism by treatment change group

To assess whether change group membership provided additional predictive validity beyond static risk, a hierarchical Cox regression was run in which Static-99 scores were included as a predictor of sexual recidivism in the first step, and change group membership was added as a covariate in the second step (group membership was dummy-coded). This analysis was conducted using a sub-sample of the entire study sample for whom Static-99 scores were available (*n* = 218). As shown in Table 10, change group membership was significantly associated with sexual recidivism after controlling for static risk.

Table 10. Incremental predictive validity of change group membership for sexual recidivism

- Loglikelihood	Chi-square change
Step One - Static risk	
391.35	
Step Two - Static risk and change group	
379.03	12.32 (*p* = .002)

Discussion

The analyses for the current study began by exploring the factor structure of the psychometric battery used in the study, finding that the individual psychometrics could be adequately represented by the four factors identified in previous research by (Allan et al., 2007): Social Inadequacy, Sexual Interests, Anger/Hostility, and Pro-Offending Attitudes. A Latent Profile Analysis (LPA) was then conducted using average standardised residual change scores across these four factors, which found that a three-class model provided the best fit to the available data. Inspection of the average factor change and raw psychometric change across these three groups led to the identification of *Poor Change* (*n* = 51), *Moderate Change* (*n* = 408) and *Good Change* (*n* = 711) groups of offenders, based on the overall levels of change they made over the course of treatment. Furthermore, a survival analysis and Cox regression found that group membership was significantly predictive of the rate of sexual recidivism after controlling for static risk, with individuals in the Good Change group reoffending at a significantly slower rate than individuals in the Moderate Change and Poor Change groups.

Together, these results indicate that there are meaningful distinctions that can be made between the change that different individuals make over the course of treatment. The differences in types of change between groups related more to the overall amount of change made across factors, rather than differences in the relative amount of change made between factors. For example, the Poor Change group did not make noticeably less change in a particular factor relative to other factors, but instead made less change than other groups across all four factors. That said, there was a slight tendency for the Moderate and Poor Change groups to show greater standardised pro-social change on measures of sexual interests relative to other factors than the Good Change group. Previous research has found that both sexual offending desisters and persisters show positive change across measures of sexual deviance but differ in terms of change across other risk factors (Lasher & McGrath, 2017); it is therefore possible that the patterns in standardised change seen across groups can be explained by more similar levels of change in sexual deviance across the sample (i.e., standardised levels of change in sexual deviance are more similar across groups because there is less variation in scores than for other factors measured).

The finding that there are distinct categories of change made by individuals over the course of treatment has important implications for the measurement of treatment change. As discussed in Study Two, measures that are developed to assess categorical constructs typically differ from measures developed to assess dimensional constructs in important ways. Namely, measures of categorical constructs typically contain fewer items that relate to the key characteristics of each category, with the goal to assign individuals to categories with maximal discriminatory accuracy; fewer items are required because there is no need to distinguish between individuals across a continuum of the construct. However, as discussed in the introduction to this study, most current approaches to the measurement of treatment change treat change as a continuous construct, by calculating change across a dimensional scoring framework.

Using measures with continuous total or scale scores may not be the optimum way for treatment change to be measured; however, the nature of the treatment change groups identified in the current study suggests that these measures may still be able to be applied in a way that is congruent with the categorical nature of treatment change. As shown in the current study, offenders can be meaningfully grouped according to the change that they have made, ranging from poor, moderate or good change. Because of the apparent sequential nature of these categories, it may be possible to identify optimum "cutting points" across treatment change scores that most accurately distinguish between the three change groups identified. It is important to explore whether purpose-built categorical treatment change measures provide incremental predictive power to the current measures, however in the meantime valid and reliable dynamic risk assessment tools may be able to continue to be used to assess change over treatment.

One benefit of using dynamic risk tools to assess both risk and treatment change is the ease of combining the two factors for risk communication. Namely, the communication of treatment change to decision-makers (e.g., judges, parole boards) could become simpler based on the findings of the current study, with levels of change being able to be conveyed using simple and non-arbitrary language. The treatment change groups identified in the current research also lend themselves well to the amendment of post-treatment risk based on change made during treatment. There are now several risk assessment tools for which estimates of recidivism have been developed by risk category (determined by identifying cut-off scores across the total score continuum; (Hanson, Babchishin, Helmus, Thornton, & Phenix, 2017; Hanson, Helmus, & Thornton, 2010; Helmus, Hanson, Thornton, Babchishin, & Harris, 2012; Olver, Mundt, et al., 2018; Olver, Beggs Christofferson, Grace, et al., 2014). Notably, the recidivism estimates calculated for the VRS-SO (Olver, Mundt, et al., 2018; Olver, Beggs Christofferson, Grace, et al., 2014) take into account both risk category and change category. These change categories (low, moderate and good change) were identified by differentiating between individuals based on their change score relative to other individuals in the sample (i.e. based on standard deviation from the sample mean). Given that this is a relatively arbitrary method of categorising change, the change categories identified in the current study could provide more meaningful distinctions between the change made by different individuals. If we can develop a reliable way of identifying group membership for different individuals (e.g., identifying cut-off scores on common measures of change, or developing specialised categorical change measures), these categories could potentially provide greater levels of predictive accuracy to recidivism estimates; the predictive power of the current study's change categories was demonstrated by their significant association with recidivism rates.

In order to develop reliable methods of assigning class membership, we first need to understand the characteristics of individuals who belong to each group, and the mechanisms that lie behind group membership (i.e., the mechanisms of treatment change). Because the change groups identified are distinguished by the overall level of change made (from poor to medium to high), it is possible that individuals move from one group to another as they continue to progress through treatment and demonstrate greater levels of change. Another possibility is that there is a factor not measured in the current study (e.g., motivation to change) that groups these individuals together for reasons other than merely the amount of change made, and which drives the patterns seen across overall treatment change at the end of treatment. The first possibility could be tested using repeated measures of treatment progress, using a tool designed for this purpose such as the SOTIPS. To assess the second possibility, we require a tool that comprehensively assesses both change in dynamic risk across multiple offence-related domains as well as factors that may contribute to the overall levels of change made across the course of treatment (i.e., a tool that captures possible mechanisms of or contributors to change); the VRS-SO is a tool that may be able to fill this gap, a possibility that is explored in the next study, Study Four.

The following study will also address one of the primary limitations of the current study: the use of self-reported information to assess treatment change. There is a notable amount of scepticism in the literature regarding the use of self-reported information to assess offender risk and change, largely because of the relative transparency of some measures (e.g., measures of sexual deviancy and cognitive distortions) and the potential benefits to "faking good" for the offender (Beech, 1998; Gannon & Polaschek, 2005). Indeed, in the current study the only measure that did not show significantly different amounts of change between the three groups was the ABCS, which measures distorted beliefs relating to sexual offending against children. As this measure is relatively transparent in what it is measuring, it is possible that individuals may have been responding in ways that indicated that they had made more prosocial change than the reality. That said, it is notable that despite the possibility of individuals “faking good”, distinct groups of individuals making different levels of change were still able to be found. Perhaps even more importantly, the recidivism rates across these groups were significantly different and in the direction that would be expected from the change profile of each group. This indicates that perhaps self-reported change is not as problematic when it comes to measuring risk and change as first thought, a finding that aligns with increasing evidence that socially desirable responding might be better considered as a factor that decreases risk of reoffending than as a potential limitation to research (Mills & Kroner, 2006; Stevens et al., 2016).

The current study extended the results of Study Two by investigating the categorical nature of change indicated in the former study, finding that change is best captured by three distinct categories: poor, medium and high change. This finding has important implications for the conceptualisation, measurement, and communication of treatment change and its relationship to risk. In the following study the validity and reliability of this finding is replicated using a clinician-rated measure of treatment change, the VRS-SO. The nature and characteristics of the change groups are also explored further to enhance our understanding of the types of individuals belonging to each change category, and therefore our understanding of the mechanisms that drive differences in treatment change.

Chapter Six/Study Four: Exploring the Relationship Between Pre-Treatment Risk, Needs and Characteristics, and Treatment Change

Results from the previous two studies suggest that treatment change displayed by men who have sexually offended against children is best conceptualised as a categorical construct, with men potentially able to be classified into one of three change groups. Findings from Study Three indicate that these groups represent men who made 'poor', 'moderate' and 'good' amounts of change across the course of treatment. Furthermore, Study Three provided preliminary evidence that these groups are significantly predictive of sexual recidivism, with men in the Good Change group being significantly less likely to reoffend sexually than men in the Poor or Moderate Change groups.

The first purpose of the current study was to validate the three treatment groups identified in Study Three by using a different measure of treatment change and observing whether the same change patterns are found between groups. This was important because as noted in Study Three, the measures used to identify the three change groups were not specifically designed to accurately and validly capture treatment change. Replication of these results using an independent and specialised measure of treatment change was therefore required.

The Violence Risk Scale – Sexual Offense Version (VRS-SO; Wong, Olver, Nicholaichuk, & Gordon, 2003) was used as the additional measure of treatment change in the current study. As explained previously, the VRS-SO is a measure that was intentionally designed to track treatment change and as such incorporates a theoretical framework for measuring treatment change: the Transtheoretical Model of Behaviour Change (Prochaska et al., 1992). More information about the VRS-SO is provided in the Method section, however there are a few notable features of the measure that are important to highlight, given the value they contribute to the current study.

First, because the VRS-SO is strongly grounded in theory and theoretical frameworks, it provides a valid and robust way of measuring meaningful characteristics related to offending both pre- and post-treatment, and overall treatment change. This provides a stronger foundation for subsequent theory generation than information provided by measures of characteristics that merely correlate with offending behaviour and/or were not specifically designed to measure treatment change, such as the psychometric battery used in Study Three. Second, the VRS-SO captures meaningful criminogenic needs across multiple offence-related domains as well as factors that may contribute to overall levels of change across treatment. As mentioned in the previous study, the comprehensiveness of the VRS-SO increases the likelihood of detecting potential change mechanisms from the exploratory analysis compared to using measures of offence-correlated characteristics only, particularly if the mechanisms are driven by characteristics that sit outside criminogenic needs themselves (e.g., motivation to change).

Linking Data to Theory

The findings from Studies Two and Three provide important information for determining the nature of treatment change and potential mechanisms underlying this change. That said, the previous two studies focus largely on describing phenomena within the data (i.e. the existence of three distinct groups of individuals who meaningfully differ from one another in terms of change made over the course of treatment) without providing strong guidance as to theories that might explain this phenomenon. As explained by Haig (2014), the abductive approach to psychological research - whereby data is used to detect phenomena which are in turn used to generate explanatory theories about the phenomena of interest -

may provide a more meaningful and valid approach to theory generation than purely inductive approaches that merely describe data (see also Borsboom, Mellenbergh, & van Heerden, 2004). For this reason, inductive approaches to theory generation can lead to the development of theories that have little value beyond describing or predicting patterns within data. For example, this criticism has been levelled at dynamic risk factors (or criminogenic needs), which in the past were commonly identified purely by their statistical correlation with recidivism outcomes (Cording, Beggs Christofferson, & Grace, 2016; Heffernan & Ward, 2015).

Altering one's research approach from inductive inference to abductive inference requires using the data and identified phenomena to generate explanatory theories; that is, abductively inferring the underlying causes of the phenomena (Haig, 2014). This is most validly achieved by using transparent and robust reasoning as a foundation for these theorised mechanisms, which includes incorporating links between the proposed mechanisms and other relevant existing theories that are well understood and widely accepted. Studies Two and Three have provided the groundwork for this process by using existing data on treatment change to identify relevant phenomena, through a largely exploratory analytical process (which Haig [2014] posits as a strong analytical approach for abuctive theory generation). However, more information about each of the three groups identified in Study Three is required to support the generation of possible causal mechanisms underlying group membership (and therefore treatment change).

The Current Study

The purpose of the current study was therefore to assess the validity of the change groups identified in Study Three, and to conduct an exploratory investigation of the pre-treatment risk, needs and demographics of individuals in each of the three change groups.

Identifying the psychological characteristics (or phenomena) displayed by individuals before they enter treatment may provide an indication of the circumstances under which change is most, or least, likely to occur. This therefore provides a useful starting point for inferences about possible casual mechanisms that are linked to these characteristics or circumstances, and which may be causing the differences in treatment change (and eventual recidivism) previously observed.

To achieve these purposes, the VRS-SO was scored for a sub-sample of the individuals used in Study Three, both pre- and post-treatment. Change across the VRS-SO total score and three sub-scale scores was then compared across groups to identify whether the same change patterns found in Study Three (poor, moderate and good change) were able to be replicated with the VRS-SO. Following this analysis, patterns of pre- and post-treatment risk, pre-treatment needs, and demographics (including offence-related histories) were compared between groups to investigate whether there were any significant differences between groups in terms of their characteristic profiles.

It was hoped that the results of this study would provide relevant and important information about the three change groups that could be used to link the findings from Studies Two and Three to possible explanatory theories of underlying causal mechanisms involved in treatment change.

Method

Participants

The sample used for the current study was the same sample as that used in Study Two. The sample comprised 1,170 men who had participated in a high-intensity, prison-based treatment programme for sexual offending against children in New Zealand between

1990 and 2007. All men had provided written consent for their information to be used for research and evaluation purposes, prior to their commencement of assessment and treatment.

The majority of the sample (69.5%) identified as being of European ethnicity. Just under a quarter (23.4%) identified as NZ Māori, with 5.2% identifying as Pasifika and 1.9% as other ethnicities. Most men in the sample had attended the treatment programme at Kia Marama Special Treatment Unit in Rolleston, New Zealand (*n* =760, or 65.0%), with the remaining 410 men (35.0%) having attended Te Piriti Special Treatment Unit in Auckland, New Zealand.

Violence Risk Scale – Sexual Offense Version (VRS-SO)

The Violence Risk Scale – Sexual Offense Version (VRS-SO; Wong, Olver, Nicholaichuk, & Gordon, 2003) is a 24-item sexual offence risk assessment and treatment planning tool. The measure includes seven static items (addressing criminal history, and victim and offender demographics) and 17 dynamic items that encompass three factors: Sexual Deviancy (sexually deviant lifestyle, sexual compulsivity, offence planning, sexual offending cycle, and deviant sexual preference); Criminality (criminal personality, interpersonal aggression, substance abuse, community support, impulsivity, and compliance with community supervision); and Treatment Responsivity (cognitive distortions, insight, release to high risk situations, and treatment compliance). There are also two items that do not load onto any factor: emotional control and intimacy deficits.

Each static and dynamic item is rated pre-treatment on a 4-point scale ranging from 0 to 3, with higher scores representing a higher level of risk/need. An offender's stage of change (Precontemplation, Contemplation, Preparation, Action, or Maintenance) is also rated for each dynamic item rated a 2 or 3, reflecting the individual's motivation and readiness to change in relation to their treatment targets (i.e. dynamic items rated a 2 or 3 pre-treatment).

Stage of change on each treatment target is then re-assessed at the end of treatment. For each stage of change that the individual has progressed through on a given item, 0.5 is subtracted from the pre-treatment score for that item (although no subtraction is made for progression from Precontemplation to Contemplation, reflecting the lack of observable behaviour change). For example, if an individual scored 3 on a dynamic item pre-treatment and progressed from the Precontemplation or Contemplation stage to the Action stage, the individual would score 2 on this item post-treatment. VRS-SO change scores are calculated by subtracting pre-treatment dynamic scores from post-treatment dynamic scores for each individual.

For this study, the VRS-SO was rated from file information for a sub-sample of 292 men, including treatment reports, case notes, and offence history documents by two independent coders who were blind as to recidivism outcome. The VRS-SO scores for 218 of these men were collected and reported on for previous research (Beggs & Grace, 2010, 2011); the coder who scored the additional 74 cases for the current study also scored the VRS-SO for 10 of the men in the original sample to test for inter-rater reliability. Where items were not able to be scored due to a lack of relevant information in the file, factor and dynamic item totals for that individual were pro-rated based on scores that were available for the remaining dynamic items, as per scoring manual protocol. Good inter-rater reliability was found for dynamic scores both pre-treatment, $r_{ICC} = .83$, $p = .011$, and post-treatment, $r_{ICC} = .79$, $p = .021$, and for static risk items, $r_{ICC} = .97$, $p < .001$.

Psychometric battery

Scores from a self-report psychometric battery completed prior to treatment were used to measure pre-treatment need in the current study; these psychometrics are the same measures used in Studies Two and Three. Below is a brief reminder of each of these measures, grouped according to the overarching psychological construct that they relate to.

Anti-social cognitions

The Abel-Becker Cognitions Scale (ABCS; Abel et al., 1989) measures distorted attitudes and beliefs about sexual offending against children.

The Hostility Toward Women scale (HTW; Check, 1985) measures negative beliefs about women, including the acceptance of aggressive motivations and behaviours directed at women.

The Rape Myth Acceptance Scale (RMAS; Burt, 1980) measures attitudes supportive of sexual violence and aggression.

Deviant sexual scripts

Wilson's Sex Fantasy Questionnaire (WSFQ; Wilson, 1978) measures the frequency or strength of different types of sexual fantasies, including intimate themes (e.g. sex with a partner), exploratory themes (e.g. group sex), impersonal themes, (e.g. sex with a stranger), and sado-masochistic themes (e.g. sex involving pain or use of force).

Emotional Dysregulation

The Beck Depression Inventory-II (BDI-II) measures depressive symptoms.

The State-Trait Anxiety Inventory (STAI; Spielberger, 1983) measures general anxiety (T-scale) and current anxiety (S-scale).

The State-Trait Anger Expression Inventory (STAXI; Spielberger, 1988) measures several aspects of anger and anger expression, including state anger, trait anger, anger suppression, anger expression, and anger control. As with Studies One and Two, both the

STAXI and the STAXI-2 were used in the assessment battery at different times over the period the study sample was collected from. The two versions of the measure are largely similar, however the STAXI-2 splits the anger control scale into two different subscales (Anger Control-Out and Anger Control-In) and has slightly more items overall (57 items compared with 44 items in the original measure).

Intimacy Deficits

The Revised UCLA Loneliness Scale (UCLS; Russell et al., 1980) measures experiences of loneliness.

The Fear of Intimacy Scale (FIS; Descutner & Thelen, 1991) measures anxiety about intimate dating relationships.

The Assertion Inventory (AI; Gambrill & Richey, 1975) measures degree of discomfort in situations requiring assertiveness (e.g. turning down a request for a meeting or date), and an individual's likelihood of making an assertive response in these situations.

The Social Self-Esteem Inventory (SSEI; Lawson et al., 1979) measures self-esteem in social situations.

The Adult Nowicki-Strickland Internal-External Control Scale (ANSIE) measures locus of control. This refers to whether an individual perceives events or outcomes as dependent on their own behaviour (internal) or as a result of independent forces, such as other people, luck or fate (external).

Procedure

Data for the current study was extracted from the New Zealand Department of Corrections database used to obtain information for the sample in Studies Two and Three. This database holds pre- and post-treatment total scores for each measure used in the psychometric battery, as well as demographic and historical information collected from participants at programme admission. Because the primary focus of the current study was validating and further exploring the results of Study Three, the current study used the same sample of 1,170 men from Study Three.

Scores for the two different versions of the STAXI were dealt with in the same manner as described in the previous two studies. For individuals who completed the STAXI-2 (n = 245) rather than the STAXI (n = 101), totals for the two Anger Control subscales were added together to create one overall Anger Control score. This was to ensure that the number of subscales was equivalent between the two versions. Because these subscales had differing numbers of items between the two versions, all scores on the individual subscales were then standardised. To retain consistency across all measures, raw scores on all remaining psychometrics were also standardised before being used in the subsequent analyses described below.

Planned Data Analysis

The first step of the analysis used a sub-sample of 292 men with completed VRS-SO measures to validate the change groups identified in Study Three. This was achieved by using one-way and repeated measures ANOVAs to compare change made on the VRS-SO across the three groups identified in Study Three. If the change groups identified in Study Three are valid, we would expect a similar change profile to be found using VRS-SO change (i.e. the three groups making poor, moderate and good amounts of change relative to one another, with significant differences in mean change between the three groups).

To increase understanding of the profile of individuals categorised in each of the change groups, a series of ANOVAs and crosstabulations was then conducted to compare risk, criminogenic needs and demographics across the three change groups. These analyses used the full sample of 1,170 men.

All analyses for the current study were conducted using SPSS 23 software.

Results

Comparison with Study Three Change Groups

From the 292 men who were included in the Study Three sample and who had completed VRS-SO measures, 14 had been categorised in Study Three as belonging to the Poor Change group, 109 in the Moderate Change group, and 169 in the Good Change group. This was approximately proportional to the relative size of the groups from Study Three, however it meant that the sample was relatively small for the Poor Change group. This means that the following analyses involving the VRS-SO were relatively low-powered for this group and results should therefore be interpreted with caution.

On average, men in the Poor Change group made the highest amount of average change between the pre- and post-treatment VRS-SO dynamic scores (M = 4.73, SD = 1.84), with the Good Change group having the second-highest average change (M = 4.62, SD = 1.83) and the Moderate Change group the lowest (M = 3.60, SD = 1.66). A one-way ANOVA showed that the difference between groups was significant (F (2,289) = 11.50, p < .001, η^2 = 0.07). Post-hoc Tukey HSD tests indicated that the Good Change group showed significantly higher overall change than the Moderate Change group (p <.001), but other group comparisons failed to reach significance.

A similar pattern was seen for change across the three VRS-SO factors, as displayed in Figure 8. Overall, men in the Poor Change group showed the highest average change for the Criminality and Treatment Responsivity factors, with men in the Good Change group showing the highest average change for the Sexual Deviance factor. Men in the Moderate

Change group had the lowest average change across all three factors. A one-way repeated measures ANOVA was conducted to assess whether the differences in factor change were significant between groups. The results showed that there was a significant difference in factor change across groups, $F(2, 289) = 9.03$, p <.001, $\eta^2 = 0.06$). Post-hoc Tukey HSD tests found that the Good Change group showed significantly higher average change than the Moderate change group for Factor 1 (p < .001) and Factor 3 (p = .032), and approached significance for Factor 2 (p = .056), but there was no significant difference between the Poor Change group and other change groups. The most consistent difference found between groups was that the Good Change group consistently made greater change than the Moderate Change group; as mentioned above, results for the Poor Change group were limited by the small sample size.

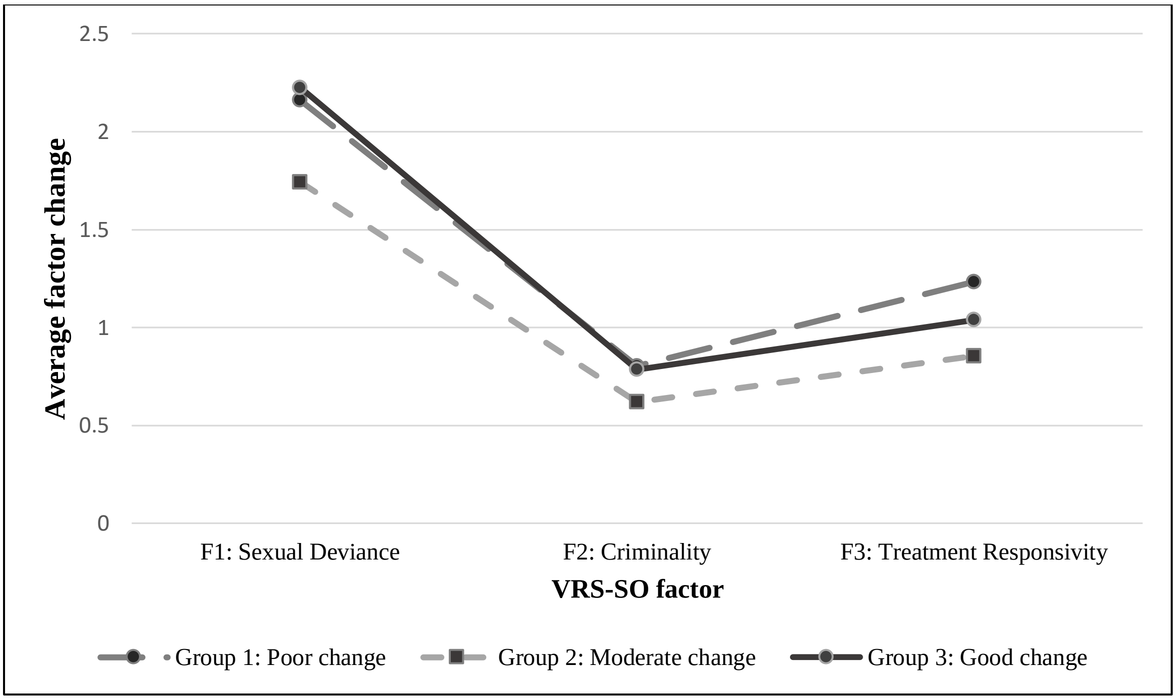

Figure 8. Average pre-post change across the VRS-SO factors, by treatment change group

Group characteristics

In order to understand more about the pre-treatment characteristics of individuals who were classified into each Study Three change group, a series of ANOVAs and crosstabulations was conducted using information about static risk, criminogenic need, and demographic characteristics. Because the different measures were not always available for the entire sample, the following analyses have been conducted using sub-samples of varying sizes. The size of each sub-sample has been indicated alongside the results of each analysis.

Pre-treatment needs

Differences in pre-treatment needs were assessed using the self-reported psychometrics completed by men prior to treatment entry. Table 11 displays pre-treatment scores on the psychometric battery, by change group. In general, the Good Change group had average scores that indicated more pro-social functioning at pre-treatment than the Moderate Change group, which in turn had average scores that were indicative of more pro-social functioning than the Poor Change group. A series of one-way ANOVAs were conducted which showed that differences across groups were significant for all psychological factors measured by the battery except social self-esteem (F(2, 1,119) = 1.90, p >.05, η^2 = 0.00) and probability of responding assertively to situations (F(2, 997) = 2.17, p >.05, η^2 = 0.00).

Post-hoc Tukey HSD tests were then conducted for all measures with significant ANOVAs. All three groups were significantly different from one another across several factors, including: having an external locus of control; anger expression; anger control; and hostility toward women. Average scores for the Good Change group were also significantly more pro-social than the other two groups for depressive symptoms, sado-masochistic fantasising, offence-supportive cognitions, and rape myth acceptance. This group also showed significantly more pro-social scores than the Poor Change group (but not the

Moderate Change group) for all remaining sub-scales on the WSFQ, state and trait anxiety, and state anger. The Moderate Change group showed significantly more pro-social scores than the Poor Change group for trait anger, but did not show significant differences across a number of other factors, including: fear of intimacy; loneliness; depressive symptoms; all subscales of the WSFQ; state and trait anxiety; state anger; anger suppression; offence-supportive cognitions; and rape myth acceptance.

Pre- and post-treatment risk

Pre- and post-treatment risk was assessed for each of the groups using scores on the VRS-SO subscales. Average scores on each of the VRS-SO subscales by group are presented in Table 12. One-way ANOVAs indicated that the change groups did not significantly differ from one another in terms of static risk; using the recent five-level risk category framework for the VRS-SO developed to align with the Council of State Governments non-arbitrary risk framework (Olver et al., 2018), all groups showed Level III (Average) static risk. There were significant differences in pre- and post-treatment dynamic risk, and pre- and post-treatment total risk across groups, however. Post-hoc Tukey HSD tests indicated that the Good Change group had significantly lower pre- and post-treatment dynamic and total risk than the Moderate Change group. Note that although the average subscale scores were significantly different, the two groups both had average pre- and post-treatment scores that placed them in the Level III (Average) risk category (although average post-treatment scores for the Good Change group were bordering on Level II [Below Average]). There were no significant differences between the Poor Change group and the other two change groups in terms of pre- or post-treatment dynamic or total risk, however the small sample size for the Poor Change group reduces the reliability of this finding.

Table 11. Pre-treatment psychometric scores, by change group

		Group One: Poor change		Group Two: Moderate change		Group Three: Good change			
Measure	*n*	Mean	SD	Mean	SD	Mean	SD	*F*	*η2*
F1: Social inadequacy									
Social self-esteem	1122	110.10	28.26	113.92	25.17	116.31	27.54	1.90	0.00
Assertion - response prob.	1000	117.30	21.78	114.24	20.40	112.08	21.12	2.17	0.00
Fear of intimacy	973	99.49[ab]	22.02	96.75[a]	20.69	93.04[b]	23.62	3.84*	0.01
UCLA loneliness	1058	45.80[ab]	9.08	46.38[a]	9.12	44.37[b]	10.64	4.71**	0.01
Depressive symptoms	818	19.21[a]	11.74	17.22[a]	10.25	14.68	9.35	8.27***	0.02
External locus of control	1082	18.57	6.66	16.33	5.29	14.27	6.03	22.96***	0.04
F2: Sexual interests									
WSFQ - exploratory	1134	14.48[a]	10.51	11.91[ab]	9.12	10.92[b]	8.54	4.59**	0.01
WSFQ - intimate	1143	27.90[a]	11.99	24.50[ab]	10.20	23.85[b]	11.44	3.16*	0.01
WSFQ - impersonal	1140	15.04[a]	8.69	12.42[ab]	8.10	11.22[b]	8.06	6.76***	0.01
WSFQ - sado/masochistic	1118	6.48[a]	5.79	5.48[a]	6.76	4.15	6.04	7.52***	0.01
F3: Anger/hostility									
State anxiety	1115	44.21[a]	14.28	40.11[ab]	12.81	38.79[b]	12.93	4.58**	0.01
Trait anxiety	1112	47.98[a]	11.73	44.44[ab]	10.86	42.87[b]	12.02	5.63**	0.01
State anger	1048	16.33[a]	7.19	14.36[ab]	6.27	14.07[b]	5.59	3.34*	0.01
Trait anger	1051	22.65	7.98	18.49[a]	6.10	17.80[a]	5.65	15.58***	0.03
Anger expression	1046	18.49	4.77	15.94	4.40	15.13	3.87	17.41***	0.03
Anger suppression	1026	19.28[ab]	4.57	18.59[a]	4.47	17.64[b]	4.73	6.51**	0.01
Anger control	1031	21.00	9.09	25.26	10.26	28.14	12.09	13.67***	0.03
F4: Pro-offending attitudes									
Abel-Becker cognition	1157	118.61[a]	16.45	119.62[a]	17.19	127.05	13.92	33.76***	0.06
Hostility toward women	1111	15.04	6.57	12.29	5.87	10.44	6.24	20.49***	0.04
Rape myth acceptance	1126	56.87[a]	21.52	51.97[a]	18.93	43.50	16.55	36.41***	0.06

* *p* <.05; ** *p* <.01; *** *p* <.001. Figures that share superscripts are not significantly different at *p* <.05 level.

Table 12. Average VRS-SO subscale scores, by group

	Group One: Poor Change		Group Two: Moderate Change		Group Three: Good Change			
n	14		109		169			
VRS-SO sub scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	*η2*
Static	7.71	4.46	8.05	4.67	7.56	4.73	0.35	0.00
Pre-treatment dynamic	24.32[ab]	5.92	24.06[a]	6.23	21.57[b]	5.62	6.58**	0.04
Pre-treatment total	32.04[ab]	9.31	32.10[a]	9.39	29.12[b]	8.83	3.84*	0.03
Post-treatment dynamic	19.59[ab]	6.19	20.46[a]	6.93	16.95[b]	6.21	9.94***	0.06
Post-treatment total	27.30[ab]	9.48	28.51[a]	9.98	24.57[b]	9.33	5.7**	0.04

* p <.05; ** p <.01; *** p <.001. Figures that share superscripts are not significantly different at p <.05 level.

Demographic and historical characteristics

The demographic and historical characteristics of group members were also compared across groups; results are displayed in Table 13. Overall, the three groups were relatively similar to one another in terms of demographic and background characteristics; one-way ANOVAs and crosstabulations found no significant differences between the groups based on: number of prior sexual offences; number of other prior serious offences; sexual abuse in childhood; physical abuse in childhood; onset of offending prior to adulthood; and relationship to victims.

A significant difference was found between the groups in terms of the average number of victims ($F(2,996) = 4.07$, $p = .017$, $\eta^2 = 0.01$). A post-hoc Tukey HSD test found that men in the Poor Change group had a significantly higher number of average victims ($M = 15.33$, $SD = 39.23$), compared with men in the Moderate Change ($M = 6.30$, $SD = 14.82$) and Good Change ($M = 6.72$, $SD = 19.38$) groups.

A significant difference was also found in the preferred victim gender between groups ($\chi 2$ (4, n = 1001) = 10.96, $p = .027$, $\varphi = 0.11$). Men in the Poor Change group were significantly less likely to report an interest in males only (9.5%) compared to expected rates, but were significantly more likely to report an interest in both males and females (14.3%).

Men in the Moderate Change group were significantly more likely to report an interest in females only (78.0%) and significantly less likely to report an interest in males only (14.3%) compared to expected rates, whereas men in the Good Change group were significantly more likely to report an interest in males only (21.0%) and significantly less likely to report an interest in females only (71.4%) than expected rates.

Table 13. Demographic characteristics by change group

	n	Statistic	*p*	Group 1: Poor Change	Group 2: Moderate Change	Class 3: Good Change
				M (*SD*)	*M* (*SD*)	*M* (*SD*)
Offence history						
No. of prior sex offence	997	$F = 0.10$	.903	1.26 (2.45)	1.03 (2.85)	1.05 (3.32)
No. of prior major offence	977	$F = 2.20$	.111	12.53 (36.84)	10.18 (57.61)	5.40 (16.39)
No. of victims	999	$F = 4.07$	.017	15.33 (39.23)	6.30[a] (14.82)	6.72[a] (19.38)
Personal history				% (*n*)	% (*n*)	% (*n*)
Was sexually abused	998	$\chi^2 = 3.13$	.209			
Yes				61.9 (26)	54.5 (189)	60.1 (366)
No				38.1 (16)	45.5 (158)	39.9 (243)
Was physically abused	1001	$\chi^2 = 3.63$	.163			
Yes				61.9 (26)	47.1 (164)	46.8 (286)
No				38.1 (16)	52.9 (184)	53.2 (325)
Offending pre-adulthood	1002	$\chi^2 = 2.08$	.353			
Yes				47.6 (20)	37.1 (130)	40.2 (245)
No				52.4 (22)	62.9 (220)	59.8 (365)
Victim profiles						
Relationship	911	$\chi^2 = 2.76$	.598			
Related only				47.5 (19)	57.1 (185)	54.5 (298)
Extrafamilial only				40.0 (16)	29.3 (95)	33.3 (182)
Mixed				12.5 (5)	13.6 (44)	12.2 (67)
Gender	1001	$\chi^2 = 10.96$	.027			
Male				9.5 (4)	14.3 (50)	21.0 (128)
Female				76.2 (32)	78.0 (273)	71.4 (435)
Both				14.3 (6)	7.7 (27)	7.6 (46)

Note: Figures that share superscripts are not significantly different at p <.05 level

Discussion

The results of the current study provided tentative support for the change groups identified in Study Three. Individuals in the Good Change group displayed significantly greater average change than the Moderate Change group for VRS-SO total and for each of the three VRS-SO subscales. However, neither the Good Change nor the Moderate Change groups made greater amounts of change than the Poor Change group on any of the VRS-SO subscales or the total VRS-SO score. Instead, average change scores were higher for the Poor Change group than the other two groups, though not significantly so.

That the Poor Change group showed the highest average change on the VRS-SO total and subscale scores was counter to expectations based on the results of Study Three. That said, the difference in change between the Poor Change group and other groups was not significant. This indicates that the small sample size could potentially be creating misleading results for the Poor Change group. The fact that the Good Change group showed significantly greater average change than the Moderate Change group across all scales, however, is what would be expected from the results of Study Three. Taken together, these results provide tentative support for Study Three's findings, however the analysis should be repeated with a larger sample in order to obtain more conclusive results.

Group Characteristics

Overall, analysis of the pre-treatment psychometric score patterns suggested that there were significant differences between change groups across a number of domains, including sexual interests, anger/hostility, pro-offending attitudes and social inadequacy. Social self-esteem and likelihood of responding assertively were the only two characteristics measured that did not show significant differences between groups. In general, individuals in the Good Change group were significantly more pro-social across a number of areas than the Poor

Change group. The Moderate Change group also tended towards being more pro-social than the Poor Change group, but was generally less pro-social than the Good Change group; however, these differences between the Moderate Change group and the other two groups were often not significant.

This finding that the Good Change group had the lowest pre-treatment level of need was also supported by results from the analysis of average VRS-SO dynamic scores, which found that the Good Change group had significantly lower pre- and post-treatment dynamic and total risk than the Moderate Change group. There was no significant difference in pre- or post-treatment dynamic or total VRS-SO scores between the Poor Change group and the other two groups; again, this could potentially be due to of the small sample size for the Poor Change group (n = 14).

To some extent, the finding that the Good Change group had lower pre-treatment needs and dynamic risk than other change groups may be surprising given that previous studies have found that individuals who are at higher risk pre-treatment tend to make larger gains during treatment (e.g., Beggs & Grace, 2011); this was theorised to be caused by higher risk/needs individuals having the opportunity to make greater amounts of change in order to reach normative levels of functioning compared to individuals who were lower risk to begin with. This theory is supported by other studies that have found that individuals who "improved" on dynamic risk factors over the course of treatment had the highest recidivism rates compared with individuals who "deteriorated" or made "no change" over treatment, using clinically significant change methodology (Wakeling et al., 2013). It was this phenomena that led to the use of standardised residual change scores when assessing the relationship between change and recidivism (Beggs & Grace, 2011; Olver, Kingston, Nicholaichuk, et al., 2014).

However, previous research has found results that are consistent with the findings of the current study. For example, Stirpe, Wilson, & Long (2001) used Standard Goal Attainment Scaling for sexual offenders (SGAS) to rate clinical treatment change demonstrated by a group of 28 low-moderate risk and 20 high-risk sexual offenders on conditional release in Canada. The researchers found that the low-moderate risk group improved significantly more than high risk offenders on GAS total scores and across multiple areas of offence-related functioning during treatment. Furthermore, the low-moderate risk group was found to have continued to change in a pro-social direction at a 3 month post-treatment follow-up, whereas the high risk group had only maintained improvements (with non-significant decreases in pro-social gains made during treatment).

One notable difference between Stirpe and colleagues' (2001) study and previous studies finding a link between high risk and high overall change is that Stirpe and colleagues captured progress using a tool specifically designed to measure treatment outcome (the SGAS), whereas others used change across psychometric batteries. Other previous studies have also found that it is not necessary to adjust raw change to control for pre-treatment risk when using specifically-designed change measures to predict sexual recidivism (e.g., Beggs & Grace, 2011, for the VRS-SO and SGAS). This suggests that these tools incorporate an important factor or mechanism of change that is not captured by merely assessing change across various psychometrics (even if these psychometrics are predictive of recidivism at pre- and/or post-treatment). This speaks to the importance of understanding the mechanisms underlying treatment change when attempting to accurately measure change and incorporate the information into decision-making and risk assessment. Such an approach provides a more robust and accurate measurement of offence-related change than relying on change across individual dynamic risk items that may be predictive of risk but lack clear causal links to offending behaviour (Heffernan & Ward, 2015).

Findings from the current study are able to provide some tentative guidance as to potential mechanisms underlying change. Importantly, the current study found that although individuals in the Good Change group were generally lower risk at pre- and post-treatment than individuals in other change groups, they did not differ significantly from the other groups in terms of static risk or most demographic or historical characteristics (apart from number of victims and preferred victim gender). This suggests that the Good Change group were not significantly different from other groups in terms of their offence-related histories or demographic backgrounds, but that they did differ in terms of their pre-treatment needs (or level of pro-sociality). This finding - that individuals in the Good Change group were significantly more pro-social prior to entering treatment despite having similar backgrounds and static risk to other groups - could potentially indicate that the Good Change group represents a group of individuals who are further along their journey of desistance than other groups, due to the influence of internal factors that support this desistence.

Findings from Stirpe and colleagues' (2001) study provide some support for this conclusion. They found that individuals who had low-moderate levels of dynamic risk prior to entering treatment (and who made greater levels of overall change across treatment) displayed similar increases in motivation to change to individuals in the high risk group. However, individuals in the low-moderate risk group maintained these increases in motivation 3 months after treatment completed, whereas the high risk group did not. Motivation to change therefore appeared to be a stronger and more permanent drive for individuals who had made the most improvement over treatment and after treatment had finished, compared with those who had made smaller and less sustained change. Other studies have also found that increased motivation to engage in treatment, manage risk factors, and exhibit a change in behaviour is linked with lower sexual reoffending rates (Olver, Beggs Christofferson, Grace, et al., 2014), with one study finding that motivation to change behaviour was the only one of the six goals included in the SGAS to significantly predict sexual reoffending after controlling for static risk (Beggs, 2008).

In terms of what might be driving this motivation to change, research and theory has suggest that “cognitive transformation” is a key mechanism underlying desistance (Giordano, Cernkovich, & Rudolph, 2002). Cognitive transformation represents a shift in the way that individuals conceptualise themselves and how their previous behaviour fits in with this self-conceptualisation. It can include changes in: openness to change; the meaning or salience of pro-social reinforcers; pro-social identity; and views of the antisocial behaviour previously engaged in. The importance of cognitive transformation has also been highlighted in the narratives of individuals who have desisted from sexual offending (Harris, 2014). Because cognitive transformation is an internal phenomena that then drives the behavioural shifts measured by treatment change tools, it could potentially be the underlying mechanism driving the high amounts of change demonstrated by the Good Change group.

That cognitive transformation is a primary driver of motivation, and therefore change, could also explain why the Good Change group does not differ from the other groups with regard to static risk or demographic characteristics – it is not differences in offence history or backgrounds that is driving the change, but instead an internal shift in self-conceptualisation toward a more pro-social identity. It could potentially also be this internal level of motivation that drives these individuals to display greater amounts of change over treatment, despite having less “room” to make change than the individuals higher in dynamic risk (although the LPA in Study Three was conducted using standardised residual change scores that control for this issue, assessment of raw change scores still showed that the Good Change group made greater amounts of raw change than other groups). Furthermore, the lower rates of dynamic risk prior to treatment entry for the Good Change group suggests that this cognitive transformation process may have begun prior to entering treatment. The concept of cognitive transformation as one of the key underlying mechanisms of treatment change also helps to explain why specifically-designed change measures such as the VRS-SO or SGAS generally perform better than other risk tools when using change to predict recidivism: these tools are grounded in change theory that captures internal motivation and stages of change, and therefore provide a better measure of potential change mechanisms than tools that rely on raw change across items that merely correlate with recidivism, or represent proxies for desistance-related factors.

The current study provides preliminary evidence to support the three change groups identified in Study Three, and for the importance of cognitive transformation as an underlying mechanism of treatment change. That said, there are a number of limitations that must be noted when interpreting these results. First, the strength of the findings is limited by the relatively small sample size for the VRS-SO, and in particular, the small number of cases in the Poor Change group. The small sample size made drawing conclusions about the validity of Study Three's findings, and the link between pre- and post-treatment risk and change, difficult. It is therefore recommended that the study is replicated using a larger sample of individuals with VRS-SO information. This would allow for a full replication of both Studies Two and Three (i.e. a taxometric analysis and latent class analysis of VRS-SO change scores), providing a strong test of the findings of these two studies.

Second, the VRS-SO scores were obtained by retrospectively scoring file information about individuals. Although raters were blind to recidivism outcomes, coding from file information means that raters were limited to the information that had been collected and retained about individuals when making their assessments. This meant that information was sometimes missing, and items at times had to be omitted. This approach also relied on the accuracy and depth of information that was previously collected. Being able to code the VRS-SO from in-depth clinical assessments including interviews would be a more robust and rich source of information for scoring the tool. That said, previous studies have still been able to demonstrate the predictive validity of VRS-SO scores coded from file information (Beggs & Grace, 2011; Olver, Beggs Christofferson, Grace, et al., 2014).

Future research could extend the findings of the current study by incorporating measures of protective factors into the assessment of change. Tools such as the SAPROF-SO (currently being piloted; Willis, Thornton, Kelley, & de Vries Robbé, 2018) include measures relevant to cognitive transformation, such as life goals, attitudes toward rules and regulations, and motivation for managing risk. It could prove useful to assess the pre- and post-treatment level of protective factors in the different change groups identified in the current research, to identify whether the Good Change group do indeed display higher levels of cognitive transformation than other groups. Research could further test the importance of cognitive transformation by exploring whether measures of cognitive transformation provide incremental predictive validity for recidivism beyond that provided by change scores alone.

Future research could also further our understanding of the importance of internal mechanisms in treatment change and desistance from offending by further exploring the stages of change captured by tools such as the VRS-SO. If internal motivation prior to entering treatment is a key component of treatment progress, this may be identified by assessing whether stage of change at pre-treatment is predictive of total change, after controlling for pre-treatment risk. Furthermore, the role of internalised stage of change in desistence could be explored by assessing whether the stage of change at pre- or post-treatment provides incremental predictive validity for recidivism, beyond that provided by risk or overall treatment change.

Overall, the results of the current study provide tentative validation of the findings from Study Three that suggested that individuals who have engaged in sexual offending treatment can be categorised into three meaningful groups: those who made Good Change, those who made Moderate Change, and those who made Poor Change over the course of treatment. Combined with the findings from Study Two, it can be concluded that these groupings are meaningfully distinct from one another, and that group membership provides important information relevant to future offending (or desistence) and to the mechanisms that underlie treatment change. The further exploration of risk, needs and demographic characteristics of these groups in the current study also provided some important directions as to what these underlying mechanisms might be, and how they might best be measured. The implications of this study, and those of the other studies presented in this dissertation, are further explored in the discussion that follows in the next chapter.

Chapter Seven: General Discussion

The previous four studies collectively provide an important contribution to the sexual offending literature by first aiming to validate an influential theory of the aetiology of sexually harmful behaviour (thereby identifying psychological factors or criminogenic needs that are meaningful in terms of prevention and treatment), and then by conducting an in-depth exploration of the nature of treatment change and the mechanisms that underlie this change. Together, this research represents a notable attempt at shifting away from a purely data-driven approach to research, and instead employing an abductive scientific approach by first using data to identify phenomena, and then exploring identified phenomena in a way that informs the generation or validation of causal theories regarding underlying mechanisms (Haig, 2005).

In this final chapter I briefly touch upon the findings and implications of the four studies included in this thesis, and then conclude with an overview of the limitations of the research and an indication of the future research directions suggested by the studies.

Overview of Research Findings and Implications

The primary aim of Study One was to attempt to validate Ward and Siegert's (2002) Pathways Model of sexual offending against children. One of the strengths of this model is the inclusion of causal mechanisms associated with each of the pathways, which provides important guidance for intervention and prevention. The study used pre-treatment scores on a psychometric battery completed by 1,134 male sexual offenders against children to conduct a Latent Profile Analysis (LPA) that identified meaningful latent classes of individuals within the sample. Results found that the sample was best captured by five classes: Low Needs (individuals with scores lower than the average across all measures); Deviant Sexual Scripts (individuals with elevated scores on measures of sexual fantasising); Intimacy Deficits (individuals who displayed elevated scores on measures of interpersonal difficulties); Emotional Dysregulation (individuals demonstrating issues with emotional control and expression); and Multiple Dysfunction (individuals with scores above the average across all measures, aside from state anger). These groups provided a good fit with the hypothesised pathways from the original model, with two notable exceptions: the originally-hypothesised Antisocial Cognitions pathway was not identified in the study sample, and the Low Needs group that was identified was not originally hypothesised in the model.

Overall, the study provided tentative support for the pathways and related causal mechanisms originally hypothesised in the model. Where findings differed from that hypothesised, the study provided valuable information for further theory development and refinement. This included suggesting further causal mechanisms not included in the original model (e.g., hypersexuality for the Deviant Sexual Scripts group), or an adjustment of the proposed mechanisms (e.g., greater emphasis on sex as coping as the primary mechanism for the Emotional Dysregulation group).

The study also contributes to a growing body of literature that stresses the importance of targeting and individualising treatment for sexual offending, in direct comparison to the currently more common approach of modular delivery of treatment (T. Gannon et al., 2012). If the offending of different individuals can be explained by distinct causal mechanisms, then it makes intuitive sense that the targets of intervention should also be tailored to address these distinct mechanisms. Results from Study One suggest that the Pathways Model provides a valuable and reliable guide for practitioners and researchers attempting to individualise treatment in this way, providing promising targets for prevention and intervention efforts. Further developing and revising the model based on results from existing validations is therefore an important focus for future research.

The results of Study One suggested that sexual offending against children is caused by a number of distinct mechanisms, which differ between individual offenders. Extrapolating these results further suggests that because of these differing mechanisms, different individuals may also vary in terms of appropriate or effective treatment targets, and therefore in terms of the kind of change that they make over the course of treatment. Furthermore, as discussed in Chapter Two of this thesis, most existing measures of treatment change treat change as dimensional (see Beggs, 2010, for review), despite no existing studies exploring or confirming this assumption.

The aim of Study Two was therefore to explore the latent structure of treatment change for sexual offenders against children, by identifying whether treatment change is best conceptualised as categorical or dimensional in nature. Results from a taxometric analysis, conducted using standardised residual change scores from 346 individuals who had completed treatment, suggested that treatment change is best conceptualised as a categorical, rather than dimensional, construct. That is, differences in treatment change between individuals are best understood as differences in the type of change made, rather than simply the amount of change made.

Conceptualising change as categorical has important implications for our theoretical understanding of the mechanisms involved in change, and for how change is most appropriately measured and communicated. Tools that are designed to measure categorical constructs differ in important ways from tools designed to measure dimensional constructs (J. Ruscio et al., 2006). These differences include measures of categorical constructs generally being shorter and less complex, and are scored in ways that help to distinguish between members of different groups in areas where there might be some overlap. Perhaps the most obvious requirement of a categorical measure of treatment change is the need for non-arbitrary assignment of individuals to different change groups, to replace the current common approach of summing numerical item scores for a continuous overall change score. Results of Study Two suggest that assessments of treatment change should be aiming to maximally discriminate between different groups of offenders making different kinds – rather than different amounts – of change. This is an approach that has been partially adopted by some dynamic risk assessment measures (e.g., the VRS-SO and its use of stages of change to measure treatment progress); however, the current research suggests that these tools will need to be revised to ensure that all items are contributing meaningfully to the measurement of change groups. As previous research has found that risk of offending is best conceptualised as dimensional (Walters et al., 2009), it is unlikely that dynamic risk measures represent the most appropriate or accurate measures of categorical treatment change.

Study Two provides an important contribution to the literature by representing the first study of the latent structure of treatment change, thereby laying the groundwork for a deeper understanding of treatment change as a phenomenon. We know from previous research that categorical constructs typically have more simple underlying mechanisms than dimensional constructs (Meehl, 1973, 1992), suggesting that the drivers of prosocial change are similarly non-complex and limited in number. As discussed above, the findings of Study Two also have important implications for the ongoing development of treatment change measures. The next obvious step to further explore the ramifications of these findings was to explore what a typology of treatment change might look like, and to identify the key characteristics or factors associated with the structure of this typology.

Thus, the aim of Study Three was to identify, based on the findings from Study Two, whether the change made by individuals over the course of treatment can be classified into meaningful and distinct categories. LPA was used to achieve this aim, using standardised residual change scores from the psychometric battery utilised for Studies One and Two. From the sample of 1,170 men convicted of sexual offences against children, three classes were extracted. Average change across the four factors identified in the psychometric battery was assessed for each group, leading to the following labels being used to describe the change patterns of each group: Poor Change (a group of individuals who did not appear to respond to treatment, and in fact displayed raw change in an antisocial direction for most measures); Moderate Change (a group of individuals who responded to some degree to treatment, or who perhaps had just begun to make change); and Good Change (a group of individuals who appeared to make good progress over the course of treatment across all factors). Furthermore, the change groups identified in the LPA demonstrated a significant association with recidivism after controlling for static risk, with individuals in the Good Change group reoffending at a significantly slower rate than individuals in the Moderate and Poor Change groups. Individuals in the Moderate Change group also tended towards reoffending at a lower rate than individuals in the Poor Change group, however this difference was not significant (possibly because of the small sample size for the Poor Change group).

Overall the findings from Study Three suggested that meaningful distinctions can be made between different kinds of change made over the course of treatment, and that these differences in change are linked to future rates of recidivism. The findings also support the conclusions suggested by the results of Study Two, that some existing measures of treatment change that incorporate some degree of treatment categorisation (e.g., the VRS-SO) may be suitable for continued application for measuring change, with some slight modifications to final scoring. Because of the apparent sequential nature of change made across groups (from poor to moderate to good amounts of change), it may be possible to identify optimum "cutting points" across existing change scales that accurately discriminate between groups. The categories of treatment change identified in the current research also fit well with existing methods of incorporating change with risk assessment and communication used for some measures. For example, the VRS-SO already utilises a three-change scale to calculate normed recidivism estimates by change and risk scores (Olver, Mundt, et al., 2018; Olver, Beggs Christofferson, Grace, et al., 2014). The change groups identified in the current research could provide a more meaningful way of distinguishing between the change groups used for recidivism estimates, compared to the current relatively arbitrary approach of using deviations from the mean to identify cut points.

Although the results of Study Three are encouraging regarding the use of some existing measures of change, it is also important to explore whether purpose-built measures will provide incremental predictive validity over modified existing measures of change. Although the change groups identified in the current research appear to be linked closely to the overall amounts of change made by individuals, it is not clear from Studies Two and Three the factors that determine group membership. On the one hand, it may be that individuals move between groups (from poor to moderate to good change) as they continue to make progress over the course of treatment. That is, it is possible that group membership is only determined by the amount of change made at a given point in time. However, the negative (i.e. antisocial) change made by individuals in the Poor Change group raises questions about this conclusion, suggesting that other factors might be driving response to treatment. An alternative explanation is that members in each group share a common characteristic (or characteristics) not measured in Study Three (e.g., motivation for change) that both serves as the primary factor linking individuals together, and influences the overall patters of change displayed across groups. Conducting further research to tease apart these two possible explanations was important both for guidance around the development of specific measures of change, and for informing theories about the mechanisms underlying the change displayed by each group. It was also important that the results of Study Three were validated using a separate measure of change, to increase confidence in the findings and provide a more robust foundation for future research and theorising building on these findings.

The aim of Study Four was therefore to provide a validation of the change groups identified in Study Three, and to conduct an exploratory investigation of the pre-treatment risk, needs and demographics of individuals in each of the three change groups. Exploring the characteristics associated with members of each group was important for providing additional understanding about the phenomenon of treatment change; a comprehensive understanding of a given phenomenon is an important step in abductively developing meaningful aetiological theories of change (Haig, 2005, 2013). The VRS-SO was used as the additional measure of change in Study Four because of the benefits of the measure already noted above, including: the incorporation of a categorical assessment of change into its ratings of change (i.e. the stages of change assessment); the comprehensive assessment of multiple offence-related items with strong theoretical underpinnings (which would allow for a wider assessment of potential factors linked to change); and previous research demonstrating the predictive validity of the VRS-SO change scores (Beggs & Grace, 2011; Olver et al., 2015). These benefits made the VRS-SO a useful tool for both validating the findings of Study Three, and providing more information that could feed into theory development.

To achieve these aims, Study Four utilised VRS-SO scores for a sub-sample of 292 men from Study Three to assess change profiles on the VRS-SO change scales for each group, and compared these with the change profiles obtained from the psychometric battery in Study Three. Overall, the results provided tentative support for the results of Study Three, finding that individuals in the Good Change group displayed significantly greater average change than the Moderate Change group for VRS-SO total and factor change scores. Contrary to expectations, the Poor Change group showed higher average change on the VRS-SO compared to the Moderate Change and Good Change groups, however this difference was not significant. The small sample size for the Poor Change group in this analysis (n = 14) may have affected the reliability of this finding, however. It is important that this analysis is replicated with a larger sample to provide a more robust test of the validity of Study Three's findings.

Study Four also included an analysis of differences in pre-treatment needs, static risk, and historical or static factors between change groups. Overall, significant differences between change groups were identified across several domains of pre-treatment need, including sexual interests, anger/hostility, pro-offending attitudes and social inadequacy, and for VRS-SO dynamic scores; in general, individuals in the Good Change group were significantly more pro-social across needs domains than the Poor Change group, with the Moderate Change group scores falling between these two other groups (although often with non-significant differences to both groups). Importantly, despite these differences in pre-treatment needs between groups, no significant differences were found in static risk or across most historical or demographic characteristics between groups.

The finding that individuals were similar in terms of their backgrounds and static level of risk prior to entering treatment, but that the groups differed in terms of their dynamic or criminogenic needs provided some potentially important information about what might be driving the patterns of change seen across groups. Given that dynamic risk factors typically provide a more recent assessment of individual characteristics and environments, the lower level of pre-treatment dynamic needs for individuals in the Good Change group (despite the similar levels of static risk) indicates that these individuals may already have been on a pathway to desistance prior to entering treatment. Additionally, we know that measuring change itself provides incremental predictive validity beyond that provided by pre- or post-treatment dynamic risk (Beggs & Grace, 2011), indicating that measuring change is capturing some important characteristic or process beyond just reducing dynamic risk factors.

Furthermore, as mentioned previously, the finding that change is categorical suggests that the mechanisms underlying change are relatively straightforward and involve only a small number of external factors (Meehl, 1973, 1992).

These considerations together provide an indication of potential mechanisms underlying change. Previous research has demonstrated that motivation to change appears to be a key component of the amounts of change made by individuals over the course of treatment (Olver, Beggs Christofferson, Grace, et al., 2014; Stirpe et al., 2001), with "cognitive transformation" potentially underlying this level of motivation (Giordano et al., 2002). Cognitive transformation represents a shift in the way that individuals conceptualise themselves and the level of congruence between their previous behaviour and this self-conceptualisation. This is an internal transformation that is not explicitly measured by existing measures of treatment change, but which would have a noticeable impact on other behaviours measured by these tools.

As the Good Change group demonstrated lower levels of dynamic risk prior to entering treatment, it is possible that this transformation had begun taking place prior to entering treatment, driving the higher levels of change displayed despite having "less room" to make this change compared to higher-risk individuals. According to frameworks such as the Good Lives Model (Ward, 2002) or Risk-Needs-Responsivity (Bonta & Andrews, 2016), this greater level of change could be explained by intervention providing individuals with the tools to successfully achieve the prosocial goals that they have developed for themselves through their transformation. Individuals in the Moderate Change group could therefore represent those individuals who required slightly more work to begin or complete their cognitive transformation, beginning to make the corresponding behavioural changes at a later stage than the individuals who were further along the process prior to entering treatment. The concept of cognitive transformation as one of the key underlying mechanisms of change also explains why change measured by tools such as the VRS-SO or SGAS are typically found to have higher predictive validity: these tools are grounded in frameworks that incorporate internal motivation and stages of change, thereby providing closer measurement of potential change mechanisms than tools that merely assess behavioural proxies for these internal motivation states. Focus on measuring internal motivation states rather than behaviours that correlate with reoffending may therefore be key in predicting response to treatment and in accurately capturing factors that are directly relevant to changes in risk or future behaviour.

Limitations

Although the research presented in the current thesis contributes information that can help fill central gaps in the current literature, there are a number of limitations associated with the research that are important to note. The first of these limitations is that all studies used data that was obtained retrospectively about individuals in each sample. The use of retrospective information is subject to a greater level of bias and confounding effects than prospectively collected information. The collection of largely administrative data in the current research also had an impact on the type of information that was available for each study. There may have been other factors or variables that it would have been valuable to include in the research, but which were not available in the existing administrative datasets. Of particular note is the lack of protective factors in the data used to explore treatment change. Although there is ongoing debate about whether protective factors represent characteristics that are meaningfully distinct from dynamic risk factors (Serin, Chadwick, & Lloyd, 2016), an increasing focus of current research involves the investigation and measurement of protective factors to identify their utility in assessment, treatment and risk prediction (Fortune & Ward, 2017). This includes the development of a sexual offending-specific protective factor assessment tool, the SAPROF-SO (Willis et al., 2018). Including protective factors in the available data might have provided a more comprehensive picture of the internal characteristics and external environments associated with individuals in each aetiological pathway (in Study One) or each change group (in Study Four), including an assessment of prosocial identity or motivation to manage risk (as included in the SAPROF-SO). This would perhaps provide a more holistic picture of offending aetiology and treatment change, improving our ability to generate meaningful and accurate causal theories to explain these phenomena.

Another important limitation of the current resesrch is the small sample size used for some of the studies, including Study Two and the analyses using the VRS-SO in Study Four. It is important that these analyses are replicated using larger sample sizes to ensure greater confidence in the findings. Relatedly, it is also important that the analyses are replicated using independent data sets. All studies in the current thesis used data from a sample of males, predominately White males, who had been convicted and imprisoned for sexual offending against children. It is important to improve the generalisability of results by replicating these findings with samples including female offenders, offenders of diverse ethnicities, and potentially individuals who have offended against adults. Although previous research has found that a large portion of the sample used in the current research (the Kia Marama sample) is lower risk than other international incarcerated samples (Olver, Beggs Christofferson, Grace, et al., 2014), it is also important that the results are replicated using a community sample of lower risk offenders.

Finally, it is important to note that much of the data used in the current study was collected via offender self-report. Although there is reason to question whether socially desirable responding in self-report has a significant negative impact on data quality (Stevens et al., 2016; Tan & Grace, 2008), using multiple different data collection methods is likely to improve the reliability and validity of findings. Particularly in the area of treatment and change, there are a number of external factors that might influence the honesty of responses to self-report measures, such as parole board decisions or impacts on security classification.

Future Research

The current research also highlights a number of promising avenues for future research. For example, the finding that cognitive transformation and motivational stage may be important factors in offender change raises interesting questions regarding whether motivational stage (or change group membership) may be more predictive of recidivism than overall levels or amounts of change made. The VRS-SO does include an assessment of motivational stage through assessment of stage of change, however this is confounded with how total change is measured using this tool and therefore was not an appropriate test of this question (and was perhaps available for too small a sample for sufficient power to detect an effect, in any case). Future research could investigate this possibility to identify the most appropriate variables or proxies capturing change to incorporate into risk assessment.

As mentioned above, it is also important for future research to further explore potential mechanisms of change, in light of current findings that change is best conceptualised as a categorical construct. Incorporating protective factors into this research may provide new avenues of information not already uncovered by existing research. Further, results from the current research suggest that assessing the validity of cognitive transformation and motivation as key components of offender change and desistance is an important focus for future research.

The categorical nature of change suggested by the current research also highlights the need for further development and refinement of existing measures of offender change. As discussed above, these measures will need to be able to effectively discriminate between change groups, particularly in areas where these groups may overlap in outward behaviour or characteristics. To support this work, further research is needed to enhance our understanding of the characteristics associated with each change group, and the mechanisms driving the patterns of change seen for these groups. Relatedly, the use of prospective research designs that measure change at multiple time points are needed to identify when or if individuals move between different groups over the course of treatment.

Concluding Statement

The research presented in the current thesis represents an important contribution to the literature in a number of ways. It represents an attempt to bridge the gap between data-driven research practices to a more abductive exploration of the aetiology of offending and nature of offender change. It also provides tentative validation of an influential theory regarding the causes of sexual offending against children, increasing our understanding of how we might prevent future offending and target causal factors in treatment (rather than just correlates of recidivism, or symptoms of offending; van den Berg et al., 2018). The research also provides the first study into the latent structure of treatment change amongst individuals who have sexually offended against children, exploring how change is best conceptualised and measured, and the possible mechanisms underlying this change.

Overall, the research adds to the growing body of literature highlighting the heterogeneity amongst individuals who sexually offend against children, and the need for individualised intervention that focusses on the promotion of prosocial identiy and skill acquisition.

References

Abel, G. G., Gore, D. K., Holland, C. L., Camp, N., Becker, J. V., & Rather, J. (1989). The measurement of the cognitive distortions in child molesters. *Annals of Sex Research*, *2*(2), 135–152. http://doi.org/10.1007/BF00851319

Ahmed, A. O., Green, B. A., McCloskey, M. S., & Berman, M. E. (2010). Latent structure of intermittent explosive disorder in an epidemiological sample. *Journal of Psychiatric Research*, *44*(10), 663–672.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *Automatic Control, IEEE Transactions On*, *19*(6), 716–723. http://doi.org/10.1109/TAC.1974.1100705

Allan, M., Grace, R. C., Rutherford, B., & Hudson, S. M. (2007). Psychometric Assessment of Dynamic Risk Factors for Child Molesters. *Sexual Abuse: A Journal of Research and Treatment*, *19*(4), 347–367. http://doi.org/10.1007/s11194-007-9052-5

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Boston: Anderson Publishing Ltd.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). *The Level of Service/Case Management Inventory*. Toronto, Canada: Multi-Health Systems.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The Recent Past and Near Future of Risk and/or Need Assessment. *Crime & Delinquency*, *52*(1), 7–27. http://doi.org/10.1177/0011128705281756

Andrews, D. A., Bonta, J., Wormith, J. S., Guzzo, L., Brews, A., Rettinger, J., & Rowe, R. (2011). Sources of variability in estimates of predictive validity: A specification with Level of Service general risk and need. *Criminal Justice and Behavior*, *38*(5), 413–432. http://doi.org/10.1177/0093854811401990

Barbaree, H. E., Seto, M. C., Langton, C. M., & Peacock, E. J. (2001). Evaluating the Predictive Accuracy of Six Risk Assessment Instruments for Adult Sex Offenders. *Criminal Justice and Behavior*, *28*(4), 490–521. http://doi.org/10.1177/009385480102800406

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*(1), 1–26. http://doi.org/10.1080/1068316X.2011.607820

Beauchaine, T. P. (2007). A Brief Taxometrics Primer. *Journal of Clinical Child & Adolescent Psychology*, *36*(4), 654–676. http://doi.org/10.1080/15374410701662840

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX.

Beech, A. R. (1998). A Psychometric Typology of Child Abusers. *International Journal of Offender Therapy and Comparative Criminology*, *42*(4), 319–339. http://doi.org/10.1177/0306624X9804200405

Beech, A. R., & Ford, H. (2006). The relationship between risk, deviance, treatment outcome and sexual reconviction in a sample of child sexual abusers completing residential treatment for their offending. *Psychology, Crime & Law*. Taylor & Francis Group. http://doi.org/10.1080/10683160600558493

Beech, A. R., Friendship, C., Erikson, M., & Hanson, R. K. (2002). The relationship between static and dynamic risk factors and reconviction in a sample of U.K. child abusers. *Sexual Abuse : A Journal of Research and Treatment*, *14*(2), 155-167; discussion 195-197. http://doi.org/10.1177/107906320201400206

Beggs, S. M. (2008). *Treatment Outcome, Risk Assessment, and Recidivism among Sexual Offenders against Children* (Unpublished doctoral thesis). University of Canterbury.

Beggs, S. M. (2010). Within-treatment outcome among sexual offenders: A review. *Aggression and Violent Behavior*, *15*(5), 369–379. http://doi.org/10.1016/j.avb.2010.06.005

Beggs, S. M., & Grace, R. C. (2010). Assessment of dynamic risk factors: an independent validation study of the Violence Risk Scale: Sexual Offender Version. *Sexual Abuse : A Journal of Research and Treatment*, *22*(2), 234–251. http://doi.org/10.1177/1079063210369014

Beggs, S. M., & Grace, R. C. (2011). Treatment gain for sexual offenders against children predicts reduced recidivism: a comparative validity study. *Journal of Consulting and Clinical Psychology*, *79*(2), 182–92. http://doi.org/10.1037/a0022900

Blanchard, R., & Barbaree, H. E. (2005). The strength of sexual arousal as a function of the age of the sex offender: Comparisons among pedophiles, hebephiles, and teleiophiles. *Sexual Abuse*, *17*(4), 441–456.

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk-20: Professional guidelines for assessing risk of sexual violence*. Vancouver, BC, Canada: Mental Health, Law and Policy Institute.

Bonta, J., & Andrews, D. A. (2016). *The Psychology of Criminal Conduct: 6th edition*. London.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity.

Psychological Review, *111*(4), 1061–1071. http://doi.org/10.1037/0033-295X.111.4.1061

Broman-Fulks, J. J., Hill, R. W., & Green, B. A. (2008). Is perfectionism categorical or dimensional? A taxometric analysis. *Journal of Personality Assessment*, *90*(5), 481–490. http://doi.org/10.1080/00223890802248802

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 and the Law*, *30*(4), 431–447.

Brown, S. L., Amand, M. D. S., & Zamble, E. (2009). The dynamic prediction of criminal recidivism: a three-wave prospective study. *Law and Human Behavior*, *33*(1), 25–45.

Burgess, A. W., Commons, M. L., Safarik, M. E., Looper, R. R., & Ross, S. N. (2007). Sex offenders of the elderly: Classification by motive, typology, and predictors of severity of crime. *Aggression and Violent Behavior*, *12*(5), 582–597. http://doi.org/10.1016/j.avb.2007.02.006

Burt, M. R. (1980). Cultural myths and supports for rape. *Journal of Personality and Social Psychology*, *38*(2), 217–230. http://doi.org/10.1.1.474.5745

Byrne, J. M., & Roberts, A. R. (2007). New directions in offender typology design, development, and implementation: Can we balance risk, treatment and control? *Aggression and Violent Behavior*, *12*(5), 483–492. http://doi.org/10.1016/j.avb.2007.02.009

Check, J. V. P. (1985). *The hostility towards women scale*. University of Manitoba.

Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, (74), 829–836.

Cohen, J. (1994). The Earth Is Round (p < .05). *The American Psychologist*, *49*(12), 997–1003. http://doi.org/10.1037/0003-066X.49.12.997

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral and health sciences*. Hoboken, NJ: John Wiley.

Colliver, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity... and back? *Medical Education*, *46*(4), 366–371. http://doi.org/10.1111/j.1365-2923.2011.04194.x

Connolly, M. (2004). Developmental Trajectories and Sexual Offending: An Analysis of the Pathways Model. *Qualitative Social Work*, *3*(1), 39–59. http://doi.org/10.1177/1473325004041131

Cording, J. R., Beggs Christofferson, S. M., & Grace, R. C. (2016). Challenges for the theory and application of dynamic risk factors. *Psychology, Crime and Law*, *22*(1–2), 84–103. http://doi.org/10.1080/1068316X.2015.1111367

Cortoni, F., & Marshall, W. L. (2001). Sex As a Coping Strategy and Its Relationship to Juvenile Sexual History and Intimacy in Sexual Offenders. *Sexual Abuse*, *13*(1), 27–43. http://doi.org/10.1177/107906320101300104

Craig, L. A., Browne, K. D., Stringer, I., & Beech, A. R. (2004). Limitations in actuarial risk assessment of sexual offenders: A methodological note. *British Journal of Forensic Practice*, *6*, 16–32.

Craissati, J., & Beech, A. R. (2003). A review of dynamic variables and their relationship to risk prediction in sex offenders. *Journal of Sexual Aggression*, *9*(1), 41–55. http://doi.org/10.1080/355260031000137968

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. http://doi.org/10.1.1.491.5048

Crowne, D. P., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: Wiley.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Davis, C. G., Thake, J., & Weekes, J. R. (2012). Impression managers: Nice guys or serious criminals? *Journal of Research in Personality*, *46*(1), 26–31. http://doi.org/http://dx.doi.org/10.1016/j.jrp.2011.11.001

Dempster, R. J., & Hart, S. D. (2002). The relative utility of fixed and variable risk factors in discriminating sexual recidivists and nonrecidivists. *Sexual Abuse: A Journal of Research and Treatment*, *14*(2), 121–138. http://doi.org/10.1177/107906320201400204

Descutner, C. J., & Thelen, M. H. (1991). Development and validation of a Fear-of-Intimacy Scale. *Psychological Assessment*, *3*(2), 218–225. http://doi.org/10.1037/1040-3590.3.2.218

Edens, J. F., Marcus, D. K., Lilienfeld, S. O., & Poythress, N. G. (2006). Psychopathic, not psychopath: Taxometric evidence for the dimensional structure of psychopathy. *Journal of Abnormal Psychology*, *115*(1), 131–144.

Ennis, L., Buro, K., & Jung, S. (2016). Identifying Male Sexual Offender Subtypes Using Cluster Analysis and the Static-2002R. *Sexual Abuse: Journal of Research and Treatment*, *28*(5), 403–426. http://doi.org/10.1177/1079063214527481

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*(1), 75–98.

Fisher, D., Beech, A. R., & Browne, K. D. (1999). Comparison of sex offenders to nonoffenders on selected psychological measures. *International Journal of Offender Therapy and Comparative Criminology*, *43*(4), 473–491.

Fortune, C.-A., & Ward, T. (2017). Problems in protective factor research and practice. *Aggression and Violent Behavior*, *32*, 1–3. http://doi.org/https://doi.org/10.1016/j.avb.2016.12.008

Fox, B. H., & DeLisi, M. (2018). From Criminological Heterogeneity to Coherent Classes: Developing a Typology of Juvenile Sex Offenders. *Youth Violence and Juvenile Justice*, *16*(3), 299–318. http://doi.org/10.1177/1541204017699257

Fox, B. H., & Farrington, D. P. (2016). Creating Burglary Profiles Using Latent Class Analysis A New Approach to Offender Profiling, *39*(12), 1582–1611. http://doi.org/10.1177/0093854812457921

Francis, B., Bowater, R., & Soothill, K. (2004). *Using homicide data to assist murder investigations*. London, UK.

Gagnon, J. H. (1990). The Explicit and Implicit Use of the Scripting Perspective in Sex Research. *Annual Review of Sex Research*, *1*(1), 1–43. http://doi.org/10.1080/10532528.1990.10559854

Gambrill, E. D., & Richey, C. A. (1975). An assertion inventory for use in assessment and research. *Behavior Therapy*, *6*(4), 550–561. http://doi.org/10.1016/S0005-7894(75)80013-X

Gannon, T. A., & Polaschek, D. L. L. (2005). Do Child Molesters Deliberately Fake Good on Cognitive Distortion Questionnaires? An Information Processing-Based Investigation . *Sexual Abuse: A Journal of Research and Treatment* . New York : Kluwer Academic

Publishers-Plenum Publishers . http://doi.org/10.1007/s11194-005-4604-z

Gannon, T. A., Ward, T., & Collie, R. (2007). Cognitive distortions in child molesters: Theoretical and research developments over the past two decades. *Aggression and Violent Behavior*, *12*(4), 402–416. http://doi.org/10.1016/j.avb.2006.09.005

Gannon, T., Terriere, R., & Leader, T. (2012). Ward and Siegert's Pathways Model of child sexual offending: a cluster analysis evaluation. *Psychology, Crime & Law*, *18*(2), 129–153. http://doi.org/10.1080/10683160903535917

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works. *Criminology*, *34*(4), 575–608.

Giordano, P. C., Cernkovich, S. A., & Rudolph, L. (2002). Gender, Crime, and Desistance: Toward a Theory of Cognitive Transformation. *American Journal of Sociology*, *107*(4), 990–1064. http://doi.org/10.1086/343191

Green, B. A., Ahmed, A. O., Marcus, D. K., & Walters, G. D. (2011). The latent structure of alcohol use pathology in an epidemiological sample. *Journal of Psychiatric Research*, *45*(2), 225–233.

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*(5), 457–480. http://doi.org/10.1177/0093854814553222

Grove, W. M., & Meehl, P. E. (1993). Simple regression-based procedures for taxometric investigations. *Psychological Reports*, *73*, 707–737.

Hagenaars, J. A., & McCutcheon, A. L. (2009). *Applied Latent Class Analysis*. Cambridge: Cambridge University Press.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*(4), 371–388. http://doi.org/10.1037/1082-989X.10.4.371

Haig, B. D. (2009). Inference to the best explanation: A neglected approach to theory appraisal in psychology. *The American Journal of Psychology*, *122*(2), 219–234.

Haig, B. D. (2012). From construct validity to theory validation. *Measurement: Interdisciplinary Research and Perspectives*, *10*(1–2), 59–62.

Haig, B. D. (2013). Detecting psychological phenomena: Taking bottom-up research seriously. *The American Journal of Psychology*, *126*(2), 135–153.

Haig, B. D. (2014). *Investigating the Psychological World: Scientific Method in the Behavioral Sciences*. Cambridge, MA.: MIT Press.

Hanson, R. K. (1997). *The development of a brief actuarial risk scale for sexual offense recidivism. (User report 97-04).* Ottawa.

Hanson, R. K., Babchishin, K. M., Helmus, L. M., Thornton, D., & Phenix, A. (2017). Communicating the Results of Criterion Referenced Prediction Measures: Risk Categories for the Static-99R and Static-2002R Sexual Offender Risk Assessment Tools. *Psychological Assessment*. Arlington: American Psychological Association. http://doi.org/10.1037/pas0000371

Hanson, R. K., & Bussiere, M. T. (1998). Predicting Relapse : A Meta-Analysis of Sexual Offender Recidivism Studies. *Journal of Consulting and Clinical Psychology*, *66*(2), 348–362.

Hanson, R. K., & Harris, A. J. R. (2000). Where Should We Intervene?: Dynamic Predictors of Sexual Offense Recidivism. *Criminal Justice and Behavior*, *27*(1), 6–35. http://doi.org/10.1177/0093854800027001002

Hanson, R. K., & Harris, A. J. R. (2001). A structured approach to evaluating change among sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, *13*(2), 105–122.

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

Hanson, R. K., Helmus, L., & Thornton, D. (2010). Predicting Recidivism Amongst Sexual Offenders: A Multi-site Study of Static-2002. *Law and Human Behavior*. Boston: Springer. http://doi.org/10.1007/s10979-009-9180-1

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: a meta-analysis of 118 prediction studies. *Psychological Assessment*, *21*(1), 1–21. http://doi.org/10.1037/a0014421

Hanson, R. K., & Thornton, D. (1999). *Static 99: Improving actuarial risk assessments for sex offenders. (User report 1999-02).* Ottawa.

Harris, D. A. (2014). Desistance From Sexual Offending: Findings From 21 Life History Narratives. *Journal of Interpersonal Violence*, *29*(9), 1554–1578. http://doi.org/10.1177/0886260513511532

Hart, S. D., Kropp, P. R., Laws, D. R., Klaver, J., Logan, C., & Watt, K. A. (2003). *The Risk for Sexual Violence Protocol (RSVP).* Burnaby, Canada: Simon Fraser University, Mental Health, Law and Policy Institute.

Heffernan, R., & Ward, T. (2015). The conceptualization of dynamic risk factors in child sex offenders: An agency model. *Aggression and Violent Behavior*, *24*, 250–260. http://doi.org/10.1016/j.avb.2015.07.001

Helmus, L. M. (2018). Sex Offender Risk Assessment: Where Are We and Where Are We Going? *Current Psychiatry Reports*, *20*(6), 1–9. http://doi.org/10.1007/s11920-018-

0909-8

Helmus, L. M., & Babchishin, K. M. (2017). Primer on risk assessment and the statistics used to evaluate its accuracy. *Criminal Justice and Behavior, 44*(1), 8–25.

Helmus, L. M., 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*. Los Angeles, CA: SAGE Publications. http://doi.org/10.1177/0093854812443648

Hildebrand, M., & de Ruiter, C. (2012). Psychopathic traits and change on indicators of dynamic risk factors during inpatient forensic psychiatric treatment. *International Journal of Law and Psychiatry*, *35*, 276–288. http://doi.org/10.1016/j.ijlp.2012.04.001

Hirschi, T., & Gottfredson, M. (1983). Age and the explanation of crime. *American Journal of Sociology*, *89*(3), 552–584.

Hogue, T. E. (1994). Goal attainment scaling: A measure of clinical impact and risk assessment. *Issues in Criminological & Legal Psychology*, *21*, 96–102.

Hudson, S. M., Wales, D. S., Bakker, L., & Ward, T. (2002). Dynamic risk factors: the Kia Marama evaluation. *Sexual Abuse: A Journal of Research and Treatment*, *14*(2), 103–119. http://doi.org/10.1177/107906320201400203

Hudson, S. M., Wales, D. S., & Ward, T. (1998). Kia Marama: A treatment program for child molesters in New Zealand. In W. L. Marshall, Y. M. Fernandez, S. M. Hudson, & T. Ward (Eds.), *Sourcebook of treatment programs for sexual offenders* (pp. 17–28). New York, NY: Plenum Press.

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*(8), 860–882.

Judge, J., Quayle, E., O'Rourke, S., Russell, K., & Darjee, R. (2014). Referrers' views of structured professional judgement risk assessment of sexual offenders: A qualitative study. *Journal of Sexual Aggression*, *20*(1), 94–109. http://doi.org/http://doi.org/10.1080/13552600.2013.767948

Jung, S., & Gulayets, M. (2011). Using clinical variables to evaluate treatment effectiveness in programmes for sexual offenders. *Journal of Sexual Aggression*, *17*(2), 166–180. http://doi.org/10.1080/13552601003802238

Kafka, M. P. (2003). Sex offending and sexual appetite: the clinical and theoretical relevance of hypersexual desire. *International Journal of Offender Therapy and Comparative Criminology*, *47*(4), 439–451. http://doi.org/10.1177/0306624X03253845

Kaseweter, K., Woodworth, M., Logan, M., & Freimuth, T. (2016). High-risk sexual offenders: Towards a new typology. *Journal of Criminal Justice*, *47*, 123–132. http://doi.org/10.1016/j.jcrimjus.2016.08.002

Knight, R. A., & Prentky, R. A. (1990). Classifying sex offenders: The development and corroboration of taxonomic models. In W. L. Marshall & H. E. Barbaree (Eds.), *Handbook of Sexual Assault: Issues, Theories, and Treatment of the Offender* (pp. 23–52). New York, NY: Plenum Press.

Lanterman, J. L., Boyle, D. J., & Ragusa-Salerno, L. M. (2014). Sex offender risk assessment, sources of variation, and the implications of misuse. *Criminal Justice and Behavior*, *41*(7), 822–843. http://doi.org/10.1177/0093854813515237

Lasher, M. P., & McGrath, R. J. (2017). Desistance From sexual and other violent offending among child sexual abusers: Observations using the Sex Offender Treatment Intervention and Progress Scale. *Criminal Justice and Behavior*, *44*(3), 416–431. http://doi.org/10.1177/0093854816670194

Lasher, M. P., McGrath, R. J., Wilson, D., & Cumming, G. F. (2015). Collaborative treatment planning using the Sex Offender Treatment Intervention and Progress Scale (SOTIPS): Concordance of therapist evaluation and client self-evaluation. *International Journal of Forensic Mental Health*, *14*(1), 1–9. http://doi.org/10.1080/14999013.2014.974087

Lawson, J. S., Marshall, W. L., & McGrath, P. (1979). The Social Self-esteem Inventory. *Educational and Psychological Measurement*, *39*(4), 803–811. http://doi.org/10.1177/001316447903900413

Lo, Y., Mendell, N., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778. http://doi.org/10.1093/biomet/88.3.767

Looman, J., Gauthier, C., & Boer, D. P. (2001). Replication of the Massachusetts Treatment Center child molester typology in a Canadian sample. *Journal of Interpersonal Violence*, *16*(8), 753–767. http://doi.org/https://doi.org/10.1177/088626001016008002

Lowenkamp, C. T., Latessa, E. J., & Holsinger, A. M. (2004). Empirical evidence on the importance of training and experience in using the Level of Service Inventory-Revised. *Topics in Community Corrections*, 49–53.

Loza, W., Dhaliwal, G., Kroner, D. G., & Loza-Fanous, A. (2000). Reliability, construct, and concurrent validities of the Self-Appraisal Questionnaire: A tool for assessing violent and nonviolent recidivism. *Criminal Justice and Behavior*, *27*(3), 356–374. http://doi.org/10.1177/0093854800027003005

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*(2), 191–217. http://doi.org/10.1177/1079063210366039

Marques, J. K., Wiederanders, M., Day, D. M., Nelson, C., & Van Ommeren, A. (2005). Effects of a relapse prevention program on sexual recidivism: Final results from California's Sex Offender Treatment and Evaluation Project (SOTEP). *Sexual Abuse*, *17*(1), 79–107.

Marshall, W. L. (1997). Pedophilia: Psychopathology and theory. In D. R. Laws & W. T. O'Donahue (Eds.), *Sexual deviance: Theory, assessment and treatment* (pp. 152–174). New York, NY: Guilford Press.

Marshall, W. L., Hamilton, K., & Fernandez, Y. (2001). Empathy deficits and cognitive distortions in child molesters. *Sexual Abuse : A Journal of Research and Treatment*, *13*(2), 123–130. http://doi.org/10.1177/107906320101300205

Marshall, W. L., Serran, G. A., Fernandez, Y. M., Mulloy, R., Mann, R. E., & Thornton, D. (2003). Therapist characteristics in the treatment of sexual offenders: Tentative data on their relationship with indices of behaviour change . *Journal of Sexual Aggression* . Taylor & Francis Group . http://doi.org/10.1080/355260031000137940

Marshall, W. L., Serran, G., Moulden, H., Mulloy, R., Fernandez, Y. M., Mann, R., & Thornton, D. (2002). Therapist features in sexual offender treatment: their reliable identification and influence on behaviour change. *Clinical Psychology & Psychotherapy*, *9*(6), 395–405. Retrieved from http://10.0.3.234/cpp.335

Martínez-Catena, A., Redondo, S., Frerich, N., & Beech, A. R. (2017). A Dynamic Risk Factors-Based Typology of Sexual Offenders. *International Journal of Offender*

Therapy and Comparative Criminology, *61*(14), 1623–1647. http://doi.org/10.1177/0306624X16629399

Mathie, N. L., & Wakeling, H. (2010). Assessing socially desirable responding and its impact on self-report measures among sexual offenders. *Psychology, Crime & Law*, *17*(3), 215–237. http://doi.org/10.1080/10683160903113681

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450–470. http://doi.org/10.1037/a0019216

McGrath, R. J., Lasher, M. P., & Cumming, G. F. (2012). The Sex Offender Treatment Intervention and Progress Scale (SOTIPS): Psychometric properties and incremental predictive validity with Static-99R. *Sexual Abuse: Journal of Research and Treatment*, *24*(5), 431–458. http://doi.org/10.1177/1079063211432475

McLachlan, G., & Peel, P. (2000). *Finite mixture models*. New York: John Wiley.

Meehl, P. E. (1973). MAXCOV-HITMAX: A taxonomic search method for loose genetic syndromes. In P. E. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 200–224). Minneapolis: University of Minnesota Press.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality*, *60*(1), 117–174.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist*, *60*, 266–275.

Meehl, P. E. (2004). What's in a taxon? *Journal of Abnormal Psychology*, *113*, 39–43.

Meehl, P. E., & Yonce, L. J. (1994). Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure). *Psychological Reports*, *74*, 1059–1274.

Meehl, P. E., & Yonce, L. J. (1996). Taxometric analysis: II. Detecting taxonicity using covariance of two quantitative indicators in successive intervals of a third indicator (MAXCOV procedure). *Psychological Reports*, *78*, 1091–1227.

Middleton, D., Elliott, I. A., Mandeville-Norden, R., & Beech, A. R. (2006). An investigation into the applicability of the Ward and Siegert Pathways Model of child sexual abuse with Internet offenders. *Psychology, Crime & Law*, *12*(6), 589–603. http://doi.org/10.1080/10683160600558352

Mills, J. F., & Kroner, D. G. (2006). Impression Management and Self-Report Among Violent Offenders. *Journal of Interpersonal Violence*, *21*(2), 178–192. http://doi.org/10.1177/0886260505282288

National Institute of Justice. (2007). *Helping probation and parole officers cope with stress*. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/220107.pdf

Neidigh, L., & Krop, H. (1992). Cognitive distortions among child sexual offenders. *Journal of Sex Education and Therapy*, *18*(3), 208–215.

Nowicki, S., & Duke, M. P. (1983). The Nowicki-Strickland Life-Span Locus of Control Scales: Construct Validation. In H. M. Lefcourt (Ed.), *Research with the locus of control construct: Vol. 2. Developments and social problems* (pp. 9–49). New York: Academic.

Nunes, K. L., & Babchishin, K. M. (2012). Construct Validity of Stable-2000 and Stable-2007 Scores. *Sexual Abuse: A Journal of Research and Treatment*, *24*(1), 29–45. http://doi.org/10.1177/1079063211404921

Nunes, K. L., Babchishin, K. M., & Cortoni, F. (2011). Measuring Treatment Change in Sex Offenders: Clinical and Statistical Significance. *Criminal Justice and Behavior*, *38*(2), 157–173. http://doi.org/10.1177/0093854810391054

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling*, *14*(4), 535–569. http://doi.org/10.1080/10705510701575396

Ó Ciardha, C., & Gannon, T. A. (2011). Are the cognitive distortions of child molesters in need of treatment? *Journal of Sexual Aggression*, *17*(2), 130–141. http://doi.org/10.1080/13552600.2011.580573

Olver, M. E., Beggs Christofferson, S. M., Grace, R. C., & Wong, S. C. P. (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*(5), 472–499. http://doi.org/10.1177/1079063213502679

Olver, M. E., Beggs Christofferson, S. M., & Wong, S. C. P. (2015). Evaluation and applications of the clinically significant change method with the Violence Risk Scale-Sexual Offender Version: Implications for Risk-Change Communication. *Behavioral Sciences and the Law*, *33*(1), 92–110.

Olver, M. E., Kingston, D. A., Nicholaichuk, T. P., & Wong, S. C. P. (2014). A psychometric examination of treatment change in a multisite sample of treated Canadian federal sexual offenders. *Law and Human Behavior*, *38*(6), 544–559. http://doi.org/http://dx.doi.org/10.1037/lhb0000086

Olver, M. E., Mundt, J. C., Thornton, D., Beggs Christofferson, S. M., Kingston, D. A., Sowden, J. N., … Wong, S. C. P. (2018). Using the Violence Risk Scale-Sexual Offense version in sexual violence risk assessments: Updated risk categories and recidivism estimates from a multisite sample of treated sexual offenders. *Psychological Assessment*, in press. http://doi.org/10.1037/pas0000538

Olver, M. E., Nicholaichuk, T. P., Kingston, D. a, & Wong, S. C. P. (2014). A multisite examination of sexual violence risk and therapeutic change. *Journal of Consulting and Clinical Psychology*, *82*(2), 312–24. http://doi.org/10.1037/a0035340

Olver, M. E., Nicholaichuk, T. P., & Wong, S. C. P. (2014). The predictive and convergent validity of a psychometric battery used to assess sexual offenders in a treatment programme: An 18-year follow-up. *Journal of Sexual Aggression*, *20*(2), 216–239. http://doi.org/10.1080/13552600.2013.816791

Olver, M. E., Sowden, J. N., Kingston, D. A., Nicholaichuk, T. P., Gordon, A., Beggs Christofferson, S. M., & Wong, S. C. P. (2018). Predictive Accuracy of Violence Risk Scale–Sexual Offender Version Risk and Change Scores in Treated Canadian Aboriginal and Non-Aboriginal Sexual Offenders. *Sexual Abuse*, *30*(3), 254–275. http://doi.org/10.1177/1079063216649594

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* . United States : American Psychological Association . http://doi.org/10.1037/1040-3590.19.3.318

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The Role of Constructs in Psychological and Educational Measurement* (pp. 49–69). Mahwah, NJ: Lawrence Erlbaum.

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66*(6), 1025–1060.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge: Cambridge University Press.

Prentky, R. A., Lee, A. F. S., Knight, R. A., & Cerce, D. (1997). Recidivism Rates among Child Molesters and Rapists : A Methodological Analysis. *Law and Human Behavior*, *21*(6), 635–659.

Prochaska, J. O., Diclemente, C. C., & Norcross, J. C. (1992). In Search of How People Change: Applications to Addictive Behaviors. *American Psychologist*, *47*, 1102–1114. http://doi.org/doi:10.3109/10884609309149692

Public Safety Canada. (2008). *Outside the walls*. Ottawa, Canada. Retrieved from http://www.publicsafety.gc.ca/cnt/cntrng-crm/csc-scc-rvw-pnl/report-rapport/otsd-wll-eng.aspx

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association.

Quinsey, V. L., Khanna, A., & Malcolm, P. B. (1998). A retrospective evaluation of the regional treatment centre sex offender treatment program. *Journal of Interpersonal Violence*, *13*(5), 621–644. http://doi.org/10.1177/088626098013005005

Robertiello, G., & Terry, K. J. (2007). Can we profile sex offenders? A review of sex offender typologies. *Aggression and Violent Behavior*, *12*(5), 508–518. http://doi.org/10.1016/j.avb.2007.02.010

Rogers, R., & Dickey, R. (1991). Denial and minimization among sex offenders. *Annals of Sex Research*, *4*(1), 49–63.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. http://doi.org/10.1037/h0042040

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335–392). Mahwah, NJ: Lawrence Erlbaum.

Ruscio, A. M., & Ruscio, J. (2002). The latent structure of analogue depression: Should the BDI be used to classify groups? *Psychological Assessment*, *14*, 135–145.

Ruscio, A. M., Ruscio, J., & Keane, T. M. (2002). The latent structure of posttraumatic stress disorder: A taxometric investigation of reactions to extreme stress. *Journal of Abnormal Psychology*, *111*(2), 290–301. http://doi.org/10.1037//0021-843X.111.2.290

Ruscio, J. (2000). Taxometric analysis with dichotomous indicators: The modified MAXCOV procedure and a case removal consistency test. *Psychological Reports*, *87*, 929–939.

Ruscio, J. (2014). Taxometric Programs for the R Computing Environment [User's Manual and code]. Retrieved May 17, 2015, from http://ruscio.pages.tcnj.edu/quantitative-methods-program-code/

Ruscio, J., Haslam, N., & Ruscio, A. M. (2006). *Introduction to the taxometric method: A practical guide*. Mahwah, NJ: Lawrence Erlbaum.

Ruscio, J., & Kaczetow, W. (2009). Differentiating categories and dimensions: Evaluating the robustness of taxometric analyses. *Multivariate Behavioral Research*, *44*(2), 259–280. http://doi.org/10.1080/00273170902794248

Ruscio, J., & Marcus, D. K. (2007). Detecting small taxa using simulated comparison data: A reanalysis of Beach, Amir, and Bau's (2005) data. *Psychological Assessment*, *19*(2), 241–246. http://doi.org/http://dx.doi.org/10.1037/1040-3590.19.2.241

Ruscio, J., Ruscio, A. M., & Meron, M. (2007). Applying the bootstrap to taxometric analysis: Generating empirical sampling distributions to help interpret results. *Multivariate Behavioral Research*, *42*(2), 349–386.

Russell, D., Peplau, L. A., & Cutrona, C. A. (1980). The revised UCLA Loneliness scale: Concurrent and discriminant validity evidence. *Journal of Personality and Social Psychology*, *39*, 472–480.

Ryan, J., Wilson, N., Kilgour, G., & Reynolds, N. (2013). Balancing efficiency and accuracy: Guidelines for the psychological assessment of extended supervision orders for child sex offenders in New Zealand. *Psychiatry, Psychology and Law*, *21*(1), 139–157. http://doi.org/10.1080/13218719.2013.793154

Sampson, R. J., & Laub, J. H. (1995). *Crime in the making: Pathways and turning points through life*. Cambridge, MA.: Harvard University Press.

Saunders, D. G. (1991). Procedures for adjusting self-reports of violence for social desirability bias. *Journal of Interpersonal Violence*, *6*(3), 336–344. http://doi.org/10.1177/088626091006003006

Scalora, M. J., & Garbin, C. (2003). A multivariate analysis of sex offender recidivism. *International Journal of Offender Therapy and Comparative Criminology*, *47*(3), 309–323. http://doi.org/10.1177/0306624X03252396

Schlager, M. D. (2009). The organizational politics of implementing risk assessment instruments in community corrections. *Journal of Contemporary Criminal Justice*, *25*(4), 412–423. http://doi.org/10.1177/1043986209344555

Schneider, A. L., Ervin, L., & Snyder-Joy, Z. (1996). Further exploration of the flight from discretion: The role of risk/need instruments in probation supervision decisions. *Journal of Criminal Justice*, *24*(2), 109–121. http://doi.org/http://dx.doi.org/10.1016/0047-

2352(95)00059-3

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Sclove, S. L. (1987). Application of model selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333–343.

Seager, J. A., Jellicoe, D., & Dhaliwal, G. K. (2004). Refusers, dropouts, and completers: Measuring sex offender treatment efficacy. *International Journal of Offender Therapy and Comparative Criminology*, *48*(5), 600–612. http://doi.org/10.1177/0306624X04263885

Serin, R. C., Chadwick, N., & Lloyd, C. D. (2016). Dynamic risk and protective factors. *Psychology, Crime & Law*, *22*(1–2), 151–170. http://doi.org/10.1080/1068316X.2015.1112013

Serin, R. C., Lloyd, C. D., Helmus, L., Derkzen, D. M., & Luong, D. (2013). Does intra-individual change predict offender recidivism? Searching for the Holy Grail in assessing offender change. *Aggression and Violent Behavior*, *18*(1), 32–53.

Seto, M. C., & Fernandez, Y. M. (2011). Dynamic Risk Groups Among Male Sexual Offenders. *Sexual Abuse : A Journal of Research and Treatment*, *23*(4), 494–507. http://doi.org/10.1177/1079063211403162

Slade, T., Grove, R., & Teeson, M. (2009). A taxometric study of alcohol abuse and dependence in a general population sample: evidence of dimensional latent structure and implications for DSM-V. *Addiction*, *104*(5), 742–751.

Spielberger, C. D. (1983). *State Trait Anxiety Inventory*. Palo Alto, CA.: Consulting Psychologists.

Spielberger, C. D. (1988). *State-Trait Anger Expression Inventory (STAXI) Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Spielberger, C. D. (1999). *State-Trait Anger Expression Inventory-2 (STAXI-2). Professional Manual*. Tampa, FL: Psychological Assessment Resources.

Stevens, C. D. J., Tan, L., & Grace, R. C. (2016). Socially desirable responding and psychometric assessment of dynamic risk in sexual offenders against children. *Psychology, Crime and Law*, *22*(5), 420–434. http://doi.org/10.1080/1068316X.2015.1120868

Stirpe, T. S., Wilson, R. J., & Long, C. (2001). Goal attainment scaling with sexual offenders: A measure of clinical impact at posttreatment and at community follow-up. *Sexual Abuse: A Journal of Research and Treatment*, *13*(2), 65–77. http://doi.org/10.1177/107906320101300201

Tan, L., & Grace, R. C. (2008). Social desirability and sexual offenders: A review. *Sexual Abuse, 20*(1), 61–87.

Tein, J.-Y., Coxe, S., & Cham, H. (2013). Statistical Power to Detect the Correct Number of Classes in Latent Profile Analysis. *Structural Equation Modeling : A Multidisciplinary Journal*, *20*(4), 640–657. http://doi.org/10.1080/10705511.2013.824781

Thornton, D. (1997). Structured anchored clinical judgement risk assessment for sex offenders. In *Proceedings of the NOTA Conference*. Brighton.

U.S. Department of Justice. (2007). *The importance of assessment in sex offender management: An overview of key principles and practices*. Silver Spring, MD. Retrieved from http://www.csom.org/pubs/assessment_brief.pdf

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psyschological Science*, *5*(3), 243–262.

van den Berg, J. W., Smid, W., Schepers, K., Wever, E., van Beek, D., Berg, J. W. Van Den, … Gijs, L. (2018). The predictive properties of dynamic sex offender risk assessment instruments : a meta-analysis. *Psychological Assessment*, *30*(2), 179–191. http://doi.org/10.1037/pas0000454

Vess, J., & Skelton, A. (2010). Sexual and violent recidivism by offender type and actuarial risk: reoffending rates for rapists, child molesters and mixed-victim offenders. *Psychology, Crime & Law*, *16*(7), 541–554. http://doi.org/10.1080/10683160802612908

Wakeling, H., Beech, A. R., & Freemantle, N. (2013). Investigating treatment change and its relationship to recidivism in a sample of 3773 sex offenders in the UK. *Psychology, Crime & Law*, *19*(3), 233–252. http://doi.org/10.1080/1068316X.2011.626413

Waller, G. D., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA: Sage Publications.

Walters, G. D., Diamond, P. M., Magaletta, P. R., Geyer, M. D., & Duncan, S. a. (2007). Taxometric analysis of the antisocial features scale of the personality assessment inventory in federal prison inmates. *Assessment*, *14*(4), 351–360. http://doi.org/10.1177/1073191107304353

Walters, G. D., Duncan, S. A., & Mitchell-Perez, K. (2007). The latent structure of psychopathy: A taxometric investigation of the Psychopathy Checklist–Revised in a heterogeneous sample of male prison inmates. *Assessment*, *14*(3), 270–278.

Walters, G. D., Knight, R. A., & Thornton, D. (2009). The Latent Structure of Sexual Violence Risk: A Taxometric Analysis of Widely Used Sex Offender Actuarial Risk

Measures. *Criminal Justice and Behavior*, *36*(3), 290–306. http://doi.org/10.1177/0093854808330341

Walters, G. D., & Ruscio, J. (2009). To sum or not to sum: Taxometric analysis with ordered categorical assessment items. *Psychological Assessment*, *21*(1), 99–111.

Walters, G. D., & Ruscio, J. (2010). Where do we draw the line? Assigning cases to subsamples for MAMBAC, MAXCOV, and MAXEIG taxometric analyses. *Assessment*, *17*, 321–333.

Ward, T. (2002). Good lives and the rehabilitation of offenders: promises and problems. *Aggression and Violent Behavior*, *7*(5), 513–528. http://doi.org/10.1016/S1359-1789(01)00076-3

Ward, T. (2014). The explanation of sexual offending: from single factor theories to integrative pluralism. *Journal of Sexual Aggression*, *20*(2), 130–141. http://doi.org/10.1080/13552600.2013.870242

Ward, T. (2016). Dynamic risk factors: scientific kinds or predictive constructs. *Psychology, Crime & Law*, *22*(1–2), 2–16. http://doi.org/10.1080/1068316X.2015.1109094

Ward, T., & Beech, A. R. (2014). Dynamic risk factors: a theoretical dead-end? *Psychology, Crime & Law*, *21*(2), 100–113. http://doi.org/10.1080/1068316X.2014.917854

Ward, T., & Casey, A. (2010). Extending the mind into the world: A new theory of cognitive distortions in sex offenders. *Aggression and Violent Behavior*, *15*(1), 49–58.

Ward, T., & Fortune, C.-A. (2016). The role of dynamic risk factors in the explanation of offending. *Aggression and Violent Behavior*, *29*, 79–88. http://doi.org/10.1016/j.avb.2016.06.007

Ward, T., Hudson, S. M., Johnston, L., & Marshall, W. L. (1997). Cognitive distortions in sex offenders: an integrative review. *Clinical Psychology Review*, *17*(5), 479–507.

Ward, T., Polaschek, D. L. L., & Beech, A. R. (2006). *Theories of sexual offending*. Chichester, England, England: Wiley.

Ward, T., & Siegert, R. J. (2002). Toward a comprehensive theory of child sexual abuse: A theory knitting perspective. *Psychology, Crime & Law*, *8*(4), 319–351. http://doi.org/10.1080/10683160208401823

Willis, G., Thornton, D., Kelley, S., & de Vries Robbé, M. (2018). *SAPROF-SO: Structured Assessment of Protective Factors for violence risk- Sexual Offence version: Pilot version*.

Wilson, G. (1978). *The secrets of sexual fantasy*. London: Dent.

Woessner, G. (2010). Classifying Sexual Offenders. *International Journal of Offender Therapy and Comparative Criminology*, *54*(3), 327–345. http://doi.org/10.1177/0306624X08331213

Wong, S. C. P., Olver, M. E., Nicholaichuk, T. P., & Gordon, A. (2003). *The Violence Risk Scale - Sexual Offender version (VRS-SO)*. Saskatoon: Regional Psychiatric Centre and University of Saskatchewan.

Wormith, J. S., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior*, *39*(12), 1511–1538. http://doi.org/10.1177/0093854812455741